UNIVERSIDAD DE MÁLAGA

E.T.S. INGENIERÍA INFORMÁTICA
UNIVERSIDAD DE MÁLAGA

Grado en Ingeniería Informática

Análisis y Modelado Temporal de Patrones de Consumo Energético en Vivienda Sostenible

Temporal Analysis and Modeling of Energy Consumption Patterns in Sustainable Homes

Realizado por
Isidro Javier García Fernández

Tutorizado por
José del Campo Ávila
Llanos Mora López

Departamento
Lenguajes y Ciencias de la Computación
UNIVERSIDAD DE MÁLAGA

MÁLAGA, junio de 2024

UNIVERSIDAD DE MÁLAGA

E.T.S. INGENIERÍA INFORMÁTICA
UNIVERSIDAD DE MÁLAGA

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA

GRADUADO EN INGENIERÍA INFORMÁTICA

Análisis y Modelado Temporal de Patrones de Consumo Energético en Vivienda Sostenible

Temporal Analysis and Modeling of Energy Consumption Patterns in Sustainable Homes

Realizado por
Isidro Javier García Fernández

Tutorizado por
José del Campo Ávila
Llanos Mora López

Departamento
Lenguajes y Ciencias de la Computación

UNIVERSIDAD DE MÁLAGA
MÁLAGA, JUNIO DE 2024

Fecha defensa: julio de 2024

Resumen

El objetivo de este estudio es aplicar técnicas de minería de datos en un conjunto de datos relacionados con el consumo energético en hogares, con el propósito de desarrollar modelos predictivos y caracterizar patrones de consumo. Estos patrones reflejan el consumo diario de energía eléctrica por vivienda y su variación a lo largo del tiempo. Capturan las fluctuaciones y tendencias en el consumo, revelando comportamientos específicos en relación con los momentos de mayor y menor demanda, factores ambientales y hábitos de uso. Con la disponibilidad de grandes conjuntos de datos sobre consumo energético es posible aplicar técnicas de minería de datos para analizar y modelar estos patrones con el fin de identificar correlaciones, predecir futuros picos de demanda, optimizar la distribución de recursos y desarrollar estrategias eficientes de gestión energética. La obtención de estos resultados involucra un enfoque teórico-práctico donde se analizan diversas técnicas de minería de datos para determinar cuáles son las más apropiadas para extraer patrones significativos del conjunto de datos y evaluar la calidad de los resultados obtenidos. Se han conseguido modelos que alcanzan precisiones de alrededor del 80 %.

Palabras clave: consumo energético, minería de patrones secuenciales, modelos de predicción

Abstract

The aim of this study is to apply data mining techniques to a dataset related to household energy consumption, with the purpose of developing predictive models and characterizing consumption patterns. These patterns reflect the daily electricity consumption per household and its variation over time. They capture fluctuations and trends in consumption, revealing specific behaviors concerning peak and off-peak demand times, environmental factors, and usage habits. With the availability of large datasets on energy consumption it is possible to apply data mining techniques to analyze and model these patterns to identify correlations, predict future demand peaks, optimize resource distribution, and develop efficient energy management strategies. Obtaining these results involves a theoretical-practical approach where various data mining techniques are analyzed to determine which are most appropriate for extracting significant patterns from the dataset and evaluating the quality of the results obtained. Models that reach a precision of around 80 % have been achieved.

Keywords: energy consumption, sequential pattern mining, prediction models

Índice

1 Introducción

1.1. Motivación

El consumo energético en las viviendas desempeña un papel fundamental en la búsqueda de la eficiencia y sostenibilidad en el uso de los recursos. En un contexto global cada vez más consciente de la importancia de conservar la energía y reducir las emisiones de carbono, comprender y modelar los patrones de consumo eléctrico en los hogares con el objetivo de conseguir reducir el consumo y/o desplazarlo a las horas más económicas se convierte en una tarea imprescindible.
En el ámbito de las comercializadoras de electricidad, este análisis es de gran importancia ya que permite anticipar la demanda de energía, identificar tendencias y anomalías y, en última instancia, promover prácticas más sostenibles y eficientes.
Resulta entonces interesante describir los siguientes conceptos en el ámbito de este trabajo:

Patrón de consumo energético: describe el consumo de energía eléctrica por vivienda y su variación en función del tiempo. Captura las fluctuaciones y tendencias en el consumo, revelando comportamientos específicos en relación con los momentos de mayor y menor demanda, factores ambientales, hábitos de uso y variaciones temporales.
Debido al gran volumen de datos que se genera en este dominio, se requieren estrategias automatizadas para ser analizados, como la minería de datos.

Minería de datos: proceso mediante el cual se extrae información de grandes conjuntos de datos con el propósito de convertirla en datos comprensibles y manejables. Utiliza métodos de estadística, aprendizaje automático y aprendizaje profundo [1].

Aprendizaje automático: es una rama de la inteligencia artificial donde se emplean algoritmos y modelos matemáticos para analizar datos, modelizar sistemas y reconocer patrones automáticamente. A medida que se proporciona más información, este proceso se vuelve más

hábil para realizar predicciones y tomar decisiones sin instrucciones directas [2].

El empleo de algoritmos de minería de datos para capturar estos patrones de consumo en cada vivienda permite predecir el consumo en momentos específicos para la anticipación en la demanda energética, la posibilidad de comparar y categorizar a los consumidores según su patrón de consumo para detectar comportamientos inusuales en el consumo, además de identificar posibles desperdicios de energía en ciertas viviendas, contribuyendo así a la mejora de la eficiencia energética.

1.2. Metodología

Con el propósito de alcanzar estos objetivos, las tareas que se van a llevar a cabo se organizarán siguiendo la metodología CRISP-DM [4] (*Cross-Industry Standard Process for Data Mining*), que es una metodología ampliamente reconocida en el campo de la minería de datos. CRISP-DM representa el ciclo de vida de un proceso de minería de datos y se divide en seis fases principales:

- **FASE 1: Comprensión del problema.** Comprender al detalle qué es lo que se quiere conseguir realmente. El objetivo es descubrir desde el principio factores importantes que pueden influir en el resultado del proyecto. Se debe registrar la información que se conoce sobre la situación al comienzo del proyecto, así como los criterios de éxito y utilidad del resultado.

- **FASE 2: Comprensión de los datos.** En esta etapa se recopilan los datos para su posterior estudio, se analiza su calidad y se describen, se exploran para comprender su comportamiento y se verifica su calidad (presencia de datos faltantes, errores en datos o metadatos, inconsistencias y similares), documentando hallazgos que guiarán las decisiones en el proyecto de minería de datos.

- **FASE 3: Preparación de los datos.** Se basa en el tratamiento inicial o preprocesamiento de los datos recopilados. Se seleccionan y limpian y, si es necesario, se construyen nuevos datos a partir de ellos.

- **FASE 4: Modelado.** Se eligen las metodologías de extracción de modelos, lo que puede incluir el ajuste de parámetros para optimizar el rendimiento o la propuesta e imple-

mentación de algoritmos innovadores para la obtención de los mismos. Utilizando estas metodologías, se generan los modelos, se comparan los resultados y se identifican aquellos que satisfacen los estándares mínimos establecidos por los expertos.

- **FASE 5: Evaluación.** Se evalúa la factibilidad de los modelos obtenidos, llevando a cabo un proceso de análisis crítico y sugiriendo mejoras para futuras iteraciones.
- **FASE 6: Despliegue y difusión:** Implementación de los modelos en el entorno de producción, asegurando que funcionen correctamente y generen valor de negocio. También incluye la preparación de documentos que recojan los avances conseguidos y permita su difusión.

En la Figura 1 se visualizan las etapas de esta metodología.

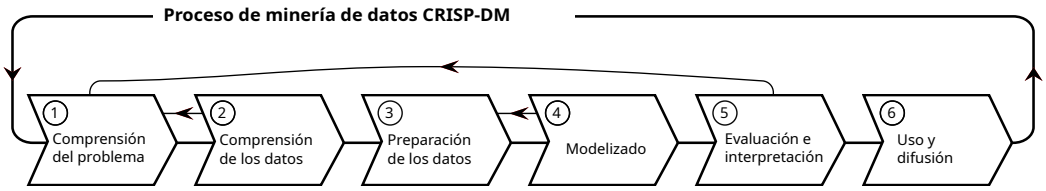


Figura 1: Esquema de la metodología CRISP-DM. Elaboración propia basada en el esquema original [4]

.

Las fases de esta metodología se realizarán en orden secuencial en este documento. El primer capítulo de esta memoria, la introducción, se corresponde con la fase inicial de la metodología CRISP-DM, donde se analiza el problema y se establecen los objetivos. La última etapa de esta metodología, la fase de despliegue, se corresponde con la redacción de esta memoria, donde se detalla todo el estudio realizado. Por último, se incluye la sección de conclusiones y líneas futuras, donde se proponen vías de desarrollo que quedan abiertas para un estudio más detallado.

1.3. Objetivos

Este trabajo de fin de grado se encuadra en el contexto de un proyecto de investigación que se encuentra en desarrollo paralelo, titulado “Desarrollo, implementación e integración de

modelos inteligentes para la gestión del autoconsumo compartido: asignación de coeficientes dinámicos de reparto e intercambio entre pares", de la convocatoria de Proyectos de Colaboración Público-Priva del Programa Estatal para impulsar la investigación científico-técnica y su transferencia, del Plan Estatal de Investigación Científica, técnica y de innovación. (Referencia: CPP2021-008403).

El objetivo primordial de este trabajo de fin de grado es aplicar procesos de minería de datos para descubrir patrones de consumo energético en viviendas. Esto se enmarca específicamente en el contexto del consumo de energía en hogares, con el propósito de lograr una clasificación y predicción en el consumo de energía o facilitar una adaptación gradual que permita un uso más eficiente de la energía. Esta investigación conlleva un proceso de selección, análisis y transformación de grandes conjuntos de datos para extraer patrones significativos. En este proceso será necesaria la aplicación de técnicas para minería de secuencias.

Podemos desglosar el objetivo principal en varias etapas con objetivos específicos:

1. Analizar detalladamente los datos de consumo energético disponibles, que incluyen información de múltiples viviendas y su consumo a lo largo del tiempo.

2. Identificar patrones de consumo energético existentes en los datos, considerando variaciones temporales y otros factores relevantes.

3. Desarrollar un modelo que permita asociar de manera efectiva los patrones de consumo con viviendas específicas, teniendo en cuenta variables temporales como el día de la semana y la hora del día.

4. Evaluar la eficacia del modelo propuesto en la asignación precisa de patrones de consumo a viviendas, utilizando métricas de rendimiento apropiadas.

5. Investigar y comprender las implicaciones de la asignación de patrones de consumo para la gestión de la eficiencia energética en entornos urbanos sostenibles.

1.4. Tecnologías usadas

Se han empleado diversas tecnologías y herramientas para llevar a cabo las diferentes etapas del proyecto. Principalmente, se ha hecho uso de los lenguajes de programación Python y R.

El lenguaje Python se ha utilizado para el preprocesado y análisis estadístico de los datos de consumo energético, con el uso de librerías como Pandas para la manipulación y análisis de datos, NumPy para operaciones numéricas, y Seaborn y Matplotlib para la visualización de datos con generación de gráficas. Se ha trabajado en el entorno de desarrollo Jupyter Notebook, pues permite una clara visualización del proceso.

R se ha empleado para la creación y evaluación de modelos de predicción de consumo de series temporales mediante técnicas de aprendizaje automático con librerías como caret, randomForest, HMM, markovchain o TraMineR y herramientas de visualización como igraph.

También se ha utilizado el lenguaje Java para evaluar modelos de análisis y predicción que utilizan algoritmos de minería de patrones secuenciales como CPT (*Compact Prediction Tree*), DG (*Dependency Graph*) y PrefixSpan. Estas herramientas y métodos se detallarán en la sección 2.5.

2

Preliminares

2.1. El problema de aprender secuencias de patrones

El problema de aprender secuencias de patrones es un área importante en el análisis de datos secuenciales. Consiste en la identificación de patrones de eventos o acciones que ocurren en una secuencia de datos a lo largo del tiempo. Existe una amplia variedad de aplicaciones, como el análisis de compras de clientes para descubrir la secuencia de productos que adquiere, el seguimiento de rutas de navegación en sitios web, la detección de patrones de comportamiento en sistemas de monitoreo, entre otros.

El desafío radica en encontrar estos patrones de manera eficiente y efectiva, especialmente cuando se trabaja con grandes volúmenes de datos. Se han desarrollado diversos algoritmos y técnicas para abordar este problema, desde enfoques basados en reglas hasta métodos más avanzados basados en modelos probabilísticos o de aprendizaje automático.
En la literatura, es común encontrar términos similares para abordar este problema, como análisis de secuencia de estados o de eventos. Es importante diferenciar este enfoque del análisis de series temporales, ya que no se manejan datos numéricos, sino categóricos. En lugar de puntos de datos continuos, se tratan secuencias de patrones que representan diferentes clases o categorías.

En este estudio, como se ha mencionado anteriormente, el objetivo es descubrir patrones de comportamiento de usuarios basándonos en una serie de datos secuenciales de consumo energético. En este contexto, se han utilizado diversos algoritmos basados en modelado y predicción de secuencias de patrones.

2.2. Métricas para la evaluación

Se han empleado diversas métricas durante el preprocesado de datos y además para evaluar el rendimiento de los modelos propuestos. Entre las métricas utilizadas se encuentran el error absoluto medio (MAE, del inglés Mean Absolute Error), el error cuadrático medio (MSE, del inglés Mean Square Error), la distancia euclidiana, la máxima verosimilitud y la tasa de acierto. Permiten una evaluación del rendimiento de un modelo en términos de la discrepancia entre los valores predichos y reales.

El error e_i para el punto i se define como la diferencia entre el valor real y_i y el valor predicho $\hat{y}_i$, es decir:

$$e_i = y_i - \hat{y}_i$$

Entonces, el error absoluto medio (MAE) se puede expresar como la media de los valores absolutos de los errores e_i:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|$$

donde n es el número total de valores.

El error cuadrático medio (MSE) se define como:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2$$

La distancia euclidiana entre dos puntos P y Q en un espacio n-dimensional se define como:

$$d_E(P,Q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

donde p_i y q_i son las coordenadas de los puntos P y Q en la dimensión i, respectivamente. La distancia euclidiana se puede utilizar para medir la similitud entre vectores o para comparar la distancia entre puntos en un espacio n-dimensional.

El estimador de máxima verosimilitud (MLE) es un método utilizado en la estimación de parámetros de un modelo estadístico. Dado un conjunto de observaciones $\{x_1, x_2, \ldots, x_n\}$ independientes e idénticamente distribuidas de una distribución de probabilidad con función de densidad $f(x;\theta)$, donde θ representa el vector de parámetros desconocidos del modelo, la función de verosimilitud se define como:

$$L(\theta|x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i;\theta)$$

El estimador de máxima verosimilitud $\hat{\theta}_{\text{MLE}}$ es aquel que maximiza la función de verosimilitud, es decir:

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\text{argmax}}\, L(\theta|x_1, x_2, \ldots, x_n)$$

En muchas aplicaciones, es más conveniente maximizar el logaritmo de la función de verosimilitud, ya que simplifica los cálculos y evita problemas de bajo flujo numérico.

$$\ell(\theta|x_1, x_2, \ldots, x_n) = \log L(\theta|x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} \log f(x_i; \theta)$$

El estimador de máxima verosimilitud $\hat{\theta}_{\text{log-MLE}}$ se obtiene maximizando la función:

$$\hat{\theta}_{\text{log-MLE}} = \underset{\theta}{\text{argmax}}\, \ell(\theta|x_1, x_2, \ldots, x_n)$$

La precisión (tasa de acierto) se define como:

$$\text{Precisión} = \frac{VP + VN}{VP + VN + FP + FN}$$

donde VP es el número de verdaderos positivos, VN es el número de verdaderos negativos, FP es el número de falsos positivos y FN es el número de falsos negativos que resultan al evaluar cada modelo.

2.3. Matriz de correlación

Durante el preprocesado de datos se ha empleado la matriz de correlación para ver similitudes entre vectores de datos y poder utilizarlas para posteriores aplicaciones en los modelos de predicción.

La **matriz de correlación** R es una matriz cuadrada $n \times n$ formada por los coeficientes de correlación de cada par de variables. Es simétrica, definida positiva y su determinante es no negativo.

$$R = \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1n} \\ r_{21} & 1 & r_{23} & \cdots & r_{2n} \\ r_{31} & r_{32} & 1 & \cdots & r_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & r_{n3} & \cdots & 1 \end{bmatrix}$$

El coeficiente de correlación muestral de Pearson puede definirse como una medida de dependencia lineal entre dos variables aleatorias cuantitativas. Es independiente de la escala de medida de las variables.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Donde:

- x_i y y_i son los valores individuales de las variables x e y, respectivamente.
- $\bar{x}$ y $\bar{y}$ son las medias de las variables x e y, respectivamente.

Es importante saber interpretar los resultados de la matriz de correlación:

- Si $r < 0$, hay correlación negativa: cuanto más próximo a -1 esté el coeficiente de correlación, más correlacionadas estarán inversamente. Si $r = -1$, se trata de correlación negativa perfecta.
- Si $r > 0$, hay correlación positiva: cuanto más próximo a $+1$ esté el coeficiente de correlación, más correlacionadas estarán directamente. Si $r = 1$, se trata de correlación positiva perfecta.
- Si $r = 0$, se dice que las variables están incorrelacionadas: no puede establecerse ningún sentido de covariación. Si dos variables son independientes estan incorrelacionadas, aunque el recíproco no es necesariamente cierto.

2.4. Validación cruzada

Además, se ha aplicado la validación cruzada como técnica para evaluar la capacidad de generalización de varios modelos.

La **validación cruzada** es una técnica para evaluar el rendimiento de modelos de aprendizaje automático que utiliza datos de entrenamiento y de prueba independientes para comprobar la calidad de generalización del modelo. Calcula la media de las ponderaciones de rendimiento obtenidas en cada una de las particiones. Destaca la **validación cruzada de K iteraciones** (*K-fold cross-validation*), donde se dividen los datos en k subconjuntos más pequeños. El modelo se entrena k veces utilizando en cada iteración k-1 subconjuntos como datos de

entrenamiento y el conjunto restante como datos de prueba. Esto proporciona independencia entre el conjunto de entrenamiento y de prueba, pues cada subconjunto es utilizado como datos de prueba una única vez.

Leave One Out Cross-Validation (LOOCV) es un caso específico de la validación cruzada de K iteraciones, donde K es el número de observaciones de los datos. En este enfoque, en cada iteración se utiliza una sola muestra como conjunto de prueba y el resto como conjunto de entrenamiento. Aunque proporciona una evaluación más precisa del modelo, puede ser computacionalmente costosa debido al elevado número de iteraciones requeridas, especialmente en conjuntos de datos grandes [5].

En el contexto de esta investigación se ha optado por utilizar la validación cruzada dejando uno fuera para algunos de los modelos empleados, ya que debido al tamaño moderado de la base de datos permite un mejor ajuste del modelo proporcionando una visión más cercana a la realidad.

2.5. Jerarquía de modelos

Se han utilizado gran variedad de modelos de minería de patrones secuenciales para datos categóricos. Estos se pueden dividir según su metodología. Podemos distinguir entre los basados en modelos de Markov, los basados en árboles de decisión, y los basados en grafos, entre otros, como se puede ver en la Figura 2.

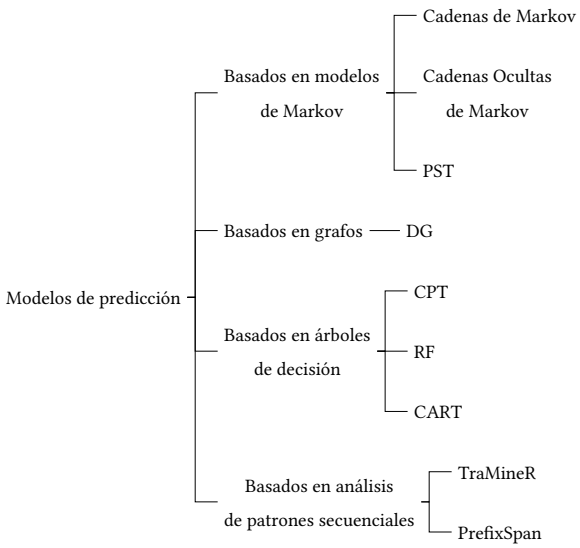


Figura 2: Tipos de los modelos de minería de patrones secuenciales para datos categóricos.

2.5.1. Modelos basados en cadenas de Markov

- **Cadenas de Markov**: se refiere a un proceso estocástico discreto donde la probabilidad de que ocurra un evento en una secuencia está condicionada únicamente por el evento inmediatamente anterior. Una Cadena de Markov $\{X_n : n = 0, 1, 2, \ldots\}$ con espacio de estados discreto S para cualquier entero $n \geq 0$ y para cualesquiera $x_0, x_1, \ldots, x_{n+1} \in S$ satisface la denominada propiedad de Markov:

$$P[X_{n+1} = x_{n+1} | X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n] = P[X_{n+1} = x_{n+1} | X_n = x_n]$$

  Dados $i, j \in S$ la probabilidad de ir del estado i en el tiempo n al estado j en el tiempo $n + 1$ viene dada por

$$p_{ij}(n, n + 1) = \mathrm{P}[X_{n+1} = j | X_n = i].$$

  Se dice que una Cadena de Markov es homogénea si la probabilidad de pasar del estado i al estado j no depende del tiempo n.

  Una vez definidas las probabilidades de transición para un único paso, representadas por p_{ij}, al variar los índices i y j a lo largo del conjunto de estados $S = 0, 1, 2, \ldots$, se genera una matriz P denominada matriz de transición de probabilidades en un paso.

$$P = \begin{bmatrix} p_{00} & p_{01} & p_{02} & \cdots \\ p_{10} & p_{11} & p_{12} & \cdots \\ p_{20} & p_{21} & p_{22} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

  Los subíndices (i, j) representan las probabilidades de pasar del estado i al estado j en un paso. La matriz P se dice que es una matriz estocástica [6].

- **Modelos Ocultos de Markov (MOM), del inglés *Hidden Markov Model* (HMM)**: En el ámbito de la probabilidad, los Modelos Ocultos de Markov son un tipo de modelo estadístico utilizado para modelar sistemas estocásticos donde se asume que cada observación de una secuencia fue generada por procesos no observables, denominados estados ocultos. Un Modelo Oculto de Markov se define por una 5-tupla (S, O, π, A, B), donde:

- $S = \{s_1, s_2, \ldots, s_N\}$ es el conjunto de estados ocultos, con N número total de estados.
- $O = \{o_1, o_2, \ldots, o_M\}$ es el conjunto de observaciones posibles, con M número total de observaciones.
- $\pi = \{\pi_i\}$ son las probabilidades iniciales. Probabilidad de iniciar en el estado s_i.

$$\pi_i = P(q_1 = s_i)$$

- $A = \{a_{ij}\}$ es la matriz de transición de estados, donde a_{ij} representa la probabilidad de transición del estado s_i al estado s_j.

$$a_{ij} = P(q_{t+1} = s_j | q_t = s_i)$$

- $B = \{b_j(o_t)\}$ es la matriz de observación, donde $b_j(o_t)$ representa la probabilidad observar o_t en el estado s_j.

$$b_j(o_t) = P(O = o_t | q_t = s_j)$$

El objetivo consiste en predecir el siguiente estado que sigue a la secuencia de observaciones $O = (O_1, O_2 ... O_T)$. Para ello se utiliza el Algoritmo de Baum-Welch [7] comprobando cuál es el conjunto de estados de transición y probabilidades que maximizan P$(O|Q, \lambda)$, donde $\lambda = (A, B, \pi)$ representa el modelo HMM.
El modelo HMM se inicializa (aleatoriamente o utilizando algún método heurístico) con ciertos parámetros que incluyen las probabilidades iniciales de estar en cada estado oculto, las probabilidades de transición entre estados ocultos y las probabilidades de emisión de cada símbolo en cada estado oculto.

El algoritmo de Baum-Welch es un algoritmo de optimización iterativo que se utiliza para estimar los parámetros óptimos del modelo HMM a partir de los datos observados. Durante cada iteración del algoritmo se realizan dos pasos: el paso de expectativa (*E-step*) y el paso de maximización (*M-step*). En [8] se profundiza detalladamente en el funcionamiento del modelo.

- En el paso de expectativa, se calculan las probabilidades posteriores de los estados ocultos dados los parámetros actuales del modelo y las observaciones. Esto se hace utilizando el algoritmo de avance hacia adelante (*forward algorithm*) y el algoritmo de retroceso (*backward algorithm*), que calculan la probabilidad de estar en cada estado oculto en cada paso de tiempo dado el modelo y las observaciones.

- En el paso de maximización, se utilizan las probabilidades posteriores calculadas en el paso anterior para actualizar los parámetros del modelo. Esto se hace maximizando la función de verosimilitud completa con respecto a los parámetros del modelo.

Los pasos de expectativa y maximización se repiten iterativamente hasta que los parámetros convergen a una solución óptima o hasta que se alcanza un criterio de convergencia. Para la aplicación en nuestro estudio, una vez que el modelo ha optimizado sus parámetros como la matriz de emisión de probabilidades, se puede utilizar para predecir el siguiente elemento de la secuencia, partiendo del último elemento de la misma.

2.5.2. Modelos basados en árboles de decisión

- **Árbol de clasificación (CART, *Classification and Regression Trees*)**: es un modelo de aprendizaje supervisado que se basa en la construcción de un árbol que divide el conjunto de datos recursivamente en subconjuntos más pequeños. Cada nodo hoja se corresponde con el siguiente estado que podría seguir al anterior. El objetivo es alcanzar nodos hojas que representen las predicciones de los próximos estados en la secuencia.

  El algoritmo CART utiliza una función de costo a minimizar utilizando el criterio de impureza de Gini [9] para determinar la probabilidad de clasificar incorrectamente una instancia y poder deducir así la mejor división de los datos:

$$I_G(D) = 1 - \sum_{i=1}^{c} p_i^2$$

  donde D es el conjunto de datos, c es el número de clases o estados y p_i es la proporción de elementos en D que pertenecen a la clase i. CART compara la diferencia entre la impureza antes y después de la división para determinar la mejor división de los datos. El proceso de división se detiene con un criterio de parada como puede ser alcanzar una profundidad máxima del árbol o un nivel de impureza cero.
  Una vez construido el árbol, la predicción para una secuencia se obtiene siguiendo el camino desde la raíz hasta un nodo hoja con la etiqueta de la clase predicha.

- **Random Forest (RF)**: es un método de aprendizaje supervisado utilizado para la clasificación y regresión. Se basa en la idea de construir múltiples árboles de decisión y combinar sus predicciones para obtener una predicción más precisa. Se trata de una mejora a la técnica conocida como *bagging* donde cada árbol se entrena con una muestra aleatoria de los datos y una selección aleatoria de características ayuda a reducir el sobreajuste.

- **Árbol Sufijo Probabilístico, del inglés *Probabilistic Suffix Tree* (PST)**: es un modelo que utiliza un árbol de sufijos para representar las probabilidades condicionadas de las letras en una secuencia. Aquí se detalla su funcionamiento interno: [10]. Implica la construcción de un árbol donde cada nodo representa un contexto de la secuencia. Sea $c = c_1, c_2, ..., c_k$ un contexto de longitud k y $x = x_1, x_2, ..., x_\ell$ una secuencia de longitud ℓ. Cada nodo almacena la probabilidad condicionada de la próxima letra dado el contexto representado por el nodo. La probabilidad condicional se calcula como:

$$P(\sigma|c) = \frac{N(c\sigma)}{\sum_{\omega \in A} N(c\omega)}$$

donde:

  - $N(c) = \sum_{i=1}^{\ell} \mathbb{1}[x_i, ..., x_{i+|c|-1} = c], \quad x = x_1, ..., x_\ell, \quad c = c_1, ..., c_k$
  - $N(c\sigma)$ es el número de veces que la subsecuencia $c\sigma$ aparece en la secuencia de entrenamiento x.
  - A es el conjunto de letras en el alfabeto.

El árbol se construye añadiendo nodos para cada contexto de la secuencia, comenzando desde el contexto más corto hasta el más largo. La etapa de crecimiento se detiene cuando cada subsecuencia distinta c de longitud $k \leq \ell - 1$ en los datos se ha añadido al árbol. El proceso puede controlarse mediante dos parámetros opcionales: L, la profundidad máxima del árbol, es decir, la longitud de contexto máxima permitida, y $nmin$, la frecuencia mínima requerida $N(c)$ de una subsecuencia c en los datos para agregarla en el árbol.
Dado un modelo PST entrenado S, se puede utilizar para predecir la probabilidad de una

secuencia específica $x \in A^{\ell}, \ell \in \mathbb{N}$.

$$P^S(x) = \prod_{i=1}^{\ell} P_S(x_i|x_1, x_2, ..., x_{i-1})$$

donde n es la longitud de la secuencia. Para evaluar la calidad de la predicción de $P^S(x)$, se puede utilizar como medida la pérdida logarítmica promedio. Puede definirse como:

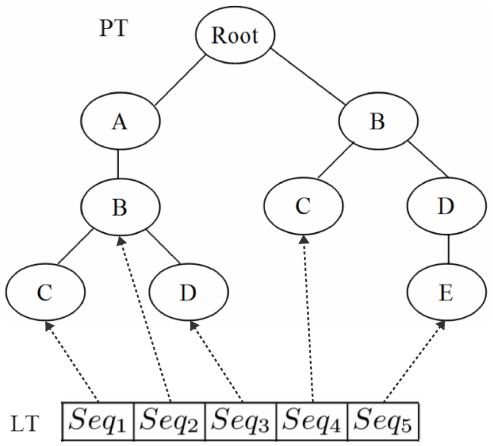
$$\text{logloss}(S, x) = -\frac{1}{\ell} \sum_{i=1}^{\ell} \log_2 P_S(x_i|x_1, x_2, ..., x_{i-1}) = -\frac{1}{\ell} \log_2 P^S(x)$$

Permite comparar las predicciones de secuencias de longitudes distintas midiendo la distancia entre la predicción de una secuencia y la predicción perfecta $P(x) = 1$, que produce una pérdida logarítmica de cero ($\text{logloss}(P^S, x) = 0$). Cuanto menor sea el valor de $\text{logloss}(P^S, x)$, mejor será la predicción de la secuencia [11].

- **Árbol de predicción compacto, del inglés *Compact Prediction Tree* (CPT)**: es un modelo de predicción de secuencias que utiliza una estructura de árbol compacta para almacenar y procesar secuencias de entrenamiento. Puede ser utilizado para predecir el siguiente elemento en una secuencia. Durante la fase de entrenamiento del modelo CPT, se construyen tres estructuras de datos principales: el Árbol de Predicción (PT), el Índice Invertido (II) y la Tabla de Búsqueda (LT).

  1. **Árbol de Predicción (PT):** cada nodo representa un elemento de la secuencia. Inserta secuencias una a una en el árbol desde el nodo raíz creando ramas y nodos hoja.
  2. **Índice Invertido (II):** se representa como una tabla donde cada clave es un elemento único del alfabeto y cada valor es un conjunto de secuencias que contienen ese elemento. Se utiliza para encontrar en qué secuencias aparece un elemento dado.
  3. **Tabla de Búsqueda (LT):** Relaciona el índice invertido y el árbol de predicción, permitiendo recuperar secuencias del árbol utilizando los identificadores de secuencia.

  A continuación se muestra en la Figura 3 un ejemplo para las secuencias: $\langle A, B, C\rangle$, $\langle A, B\rangle, \langle A, B, D\rangle, \langle B, C\rangle, \langle B, D, E\rangle$.



II

Item	Seq_1	Seq_2	Seq_3	Seq_4	Seq_5
A	1	1	1	0	0
B	1	1	1	1	1
C	1	0	0	1	0
D	0	0	1	0	1
E	0	0	0	0	1

Figura 3: Ejemplo de PT, II y LT [12]

Dada una secuencia $S = (s_1, s_2, ..., s_n)$ y una longitud x, se buscan todas las secuencias en el árbol de predicción (PT) que contienen los últimos x elementos de S en cualquier orden y en cualquier posición. Puede verse como la intersección de los subconjuntos de los últimos x elementos de S, utilizando la estructura de Índice Invertido (II).

$$\text{Similar}(S) = \bigcap_{i=1}^{x} \text{Set}(s_{n-i})$$

Para cada secuencia similar $Y \in \text{Similar}(S)$, se obtiene su consecuente con respecto a S, que es la subsecuencia de Y que comienza después del último elemento en común con S hasta el final de Y. Luego se construye una estructura llamada Tabla de Conteo (CT) que contiene los posibles elementos candidatos (z_i) y sus puntuaciones asociadas para la predicción.

$$\text{Support}(z_i) = \text{Número de veces que } z_i \text{ aparece en Similar}(S)$$

$$\text{Confidence}(z_i) = \frac{\text{Support}(z_i)}{\text{Número total de secuencias de entrenamiento que contienen } z_i}$$

Se tiene que $\text{Score}(z_i) = \text{Support}(z_i)$ y en caso de coincidir entre varios elementos se considera $\text{Score}(z_i) = \text{Confidence}(z_i)$.

El elemento predicho es aquel con la puntuación más alta dentro de la tabla de conteo.

$$\hat{z} = \arg\max_{z_i \in \text{CT}} \text{Score}(z_i)$$

Es fácil comprobar que la creación de las estructuras utilizadas y el algoritmo de predicción tienen complejidad $O(n)$.

- **Grafo de dependencias (DG, *Dependency Graph*)**: es un grafo dirigido que representa la dependencia de varios elementos entre sí. Cada nodo apunta al nodo del que depende, y en el contexto de este estudio se puede interpretar como la relación entre los elementos consecutivos de una secuencia. Se puede entender como una unión de árboles de dependencia, como se detalla en [13].

  Dado un conjunto de elementos S y una relación transitiva $R \subseteq S \times S$ con $(a, b) \in R$ que representa ¨a depende de b", el grafo de dependencias es un grafo $G = (S, T)$ con $T \subseteq R$ la reducción transitiva de R, es decir, con la menor cantidad de aristas posibles entre cada par de nodos [14].

  En un grafo de dependencias, la presencia de ciclos no permite definir un orden de evaluación válido, ya que ninguno de los elementos en el ciclo puede ser evaluado primero. Sin embargo, si el grafo es acíclico es posible encontrar un orden de evaluación mediante ordenamiento topológico.

  Un orden de evaluación correcto consiste en una numeración $n : S \to \mathbb{N}$ de los objetos que forman los nodos del grafo de manera que: $n(a) < n(b) \Rightarrow (a, b) \notin R$ con $a, b \in S$. Esto garantiza que los objetos se evalúen en el orden adecuado según sus dependencias. Puede haber más de un orden de evaluación correcto.

  Cuando se accede a un elemento $a \in S$, tiene sentido prever la necesidad de evaluar previamente otro elemento $b \in S$ si la arista de a a b tiene un peso alto (generalmente por encima de un umbral p), lo que sugiere una alta probabilidad de que b sea necesario poco después [15].
  Como los anteriores modelos, también se ha empleado para predecir el siguiente elemento de una secuencia de patrones.

2.5.3. Otros modelos

- **TraMineR**: es un conjunto de herramientas de análisis de secuencias categóricas. Entre las características principales de TraMineR se incluyen:

  - Funciones para describir y representar secuencias.
  - Cálculo de distancias entre secuencias utilizando diferentes métricas.

- Cálculo de tasas de transición entre estados en un conjunto de secuencias, es decir, la probabilidad de cambiar de un estado a otro.

La función *seqtrate* en TraMineR calcula las tasas de transición entre cada par de estados (s_i, s_j) en un conjunto de secuencias categóricas S. Para calcular estas tasas, se utiliza la siguiente función:

$$p(s_j|s_i) = \frac{\sum_{t=1}^{L-1} n_{t,t+1}(s_i, sj)}{\sum_{t=1}^{L-1} nt(s_i)}$$

Donde:

- $n_t(s_i)$ es el número de secuencias que no terminan en t con el estado si en la posición t.
- $n_{t,t+1}(s_i, s_j)$ es el número de secuencias con el estado s_i en la posición t y el estado s_j en la posición $t + 1$.
- L es la longitud máxima de secuencia observada.

La salida de la función es una matriz donde cada fila i proporciona una distribución de transición desde el estado de origen s_i en t a los estados en $t + 1$; es decir, la suma total de cada fila es igual a uno. Esta matriz proporciona información sobre los cambios de estado más frecuentes en las secuencias observadas y sobre la estabilidad de cada estado analizando la diagonal de la matriz [16].

- **PrefixSpan**: es un algoritmo de minería de patrones secuenciales cuyo objetivo principal en este estudio es la obtención de patrones de una secuencia. Vamos a definir algunos conceptos:

  Prefijo: es una subsecuencia inicial de una secuencia dada que permite encontrar patrones en los datos.

  Proyección: la proyección de una secuencia con respecto a un prefijo implica observar la secuencia resultante de eliminar los elementos que preceden a ese prefijo.

  Sufijo: es la parte restante de una secuencia después de eliminar el prefijo. Es la parte de la secuencia que aún no ha sido examinada en la búsqueda de patrones.

El algoritmo PrefixSpan [17] opera de la siguiente manera: comienza con un prefijo vacío y gradualmente lo expande para formar secuencias más largas. En cada paso, examina las secuencias que siguen al prefijo actual y busca patrones frecuentes. Si encuentra un patrón frecuente, lo registra y continúa buscando patrones adicionales.

3

Comprensión y Preparación de los datos

En esta sección se englobarán las tres primeras fases de la metodología CRISP-DM: comprensión del problema, comprensión de los datos y preparación de los datos.

3.1. Comprensión del problema

En el capítulo introductorio de este documento se encuentra detallada la etapa de comprensión del problema. En los apartados 1.1 y 1.3 se describe el contexto y lo que se pretende conseguir, así como los criterios de éxito y utilidad del resultado. Se plantea como objetivo descubrir un modelo capaz de predecir patrones y tipos de consumo que tienen los usuarios basándose en sus patrones de consumo energético. Se considerarán satisfechos los objetivos en el caso en el que se encuentren modelos para predecir consumos horarios a corto plazo.

Se añade el código utilizado durante este estudio para poder ser utilizado con otros datos de consumo y ser configurable en cada caso.

3.2. Comprensión de los datos

Este estudio se realiza paralelamente a un proyecto de investigación mencionado en el capítulo introductorio, por lo tanto, se cuenta con un conjunto de datos y otras investigaciones previas. Se dispone de:

- **Conjunto de datos original**: se dispone de un conjunto de datos público de Irlanda [18], proporcionado por el Proyecto de Medición Inteligente de la Comisión de Regula-

ción de Energía (CER). Este conjunto de datos contiene información sobre el consumo eléctrico de hogares. El formato de los datos es de 3 columnas que representan:

  - Identificador de usuario
  - Código de 5 dígitos
    - Los primeros tres dígitos representan el código del día. El día 1 corresponde al 1 de enero de 2009.
    - Los últimos dos dígitos representan el código de tiempo, con valores del 1 al 48, donde cada número representa un intervalo de 30 minutos desde las 0 horas de un día.
    - La cantidad de electricidad consumida en intervalos de 30 minutos (en kWh)

- **Conjunto de datos procesados**: se realiza un filtrado de datos del conjunto original para mejorar la calidad de las observaciones y posterior entrenamiento de modelos de predicción.

- **Modelos de clasificación**: Se cuenta con un conjunto de modelos de clasificación obtenidos a partir del algoritmo de K-medias. Estos grupos de consumo han sido obtenidos mediante el modelo de K-medias. Finalmente, se consigue clasificar en 21 grupos que distinguen tipos de consumo horario durante un día en los usuarios de la base de datos. Este método es detallado en el artículo publicado en [19].

3.3. Preparación y descripción de los datos

Como se ha descrito en la sección 3.2, el fichero de datos de consumo energético se presenta en intervalos de 30 minutos para cada día y para cada usuario. Como el objetivo es trabajar con consumos horarios para cada usuario y día, se ha hecho una transformación de los datos para obtener esos valores horarios. Se obtienen los siguientes resultados:

- Hay un número de 6 435 usuarios únicos con consumo energético.

- Hay datos de consumo para 536 días, comenzando desde el día 195 (15 de julio de 2009) y terminando en el día 730 (1 de enero de 2011).

- Hay un total de 3 291 258 observaciones diarias de datos de consumo energético en la base de datos.

El objetivo del estudio es analizar y predecir los patrones de consumo energético en hogares, luego es fundamental garantizar que los datos utilizados para el análisis de secuencias de consumo sean representativos del comportamiento típico de consumo residencial y estén libres de ruido que corresponda a consumos no domésticos.

Por lo tanto, se implementa un proceso de filtrado que elimina ciertos tipos de observaciones que podrían distorsionar los patrones de consumo deseados:

1. Se considera como consumidores no residenciales a aquellos con un consumo eléctrico superior a 15 kWh en cualquier hora de cualquier día. Estos usuarios han sido eliminados del conjunto de datos. Se identificaron 946 usuarios que en alguna hora tuvieron un consumo energético de más de 15 kWh. Estos usuarios han sido eliminados completamente.

2. Se eliminan las observaciones de días donde no hay datos de consumo de alguna hora. Es necesario disponer de los datos diarios completos.

3. Se eliminan las observaciones con un consumo diario total inferior a 100 Wh. Estos datos de consumo son demasiado bajos para representar el consumo típico de energía diario en un hogar. Se identificaron 220 usuarios que algún día tuvieron un consumo diario por debajo de 100 Wh. Se detectó un total de 15 729 ocasiones donde existió un consumo por debajo del mínimo establecido. Estos días de consumo han sido eliminados para esos usuarios.

Después de aplicar los filtros a los usuarios con datos de consumo diarios se obtienen los siguientes resultados:

- Hay un número de 5 483 usuarios únicos con consumo energético que cumplen los criterios especificados.

- Hay datos de consumo para 536 días, comenzando desde el día 195 (15 de julio de 2009) y terminando en el día 730 (1 de enero de 2011).

- Hay un total de 2 772 102 observaciones diarias de datos de consumo energético en la base de datos.

Una vez se dispone de una base de datos filtrada de consumo energético horario agrupado por días, el siguiente paso consiste en clasificar el tipo de consumo diario de cada usuario. Esta clasificación se realiza utilizando un modelo de K-medias con 21 grupos (clústeres) distintos que representan diferentes patrones de consumo. En la Figura 4 se examinan los diferentes tipos de consumo diario.

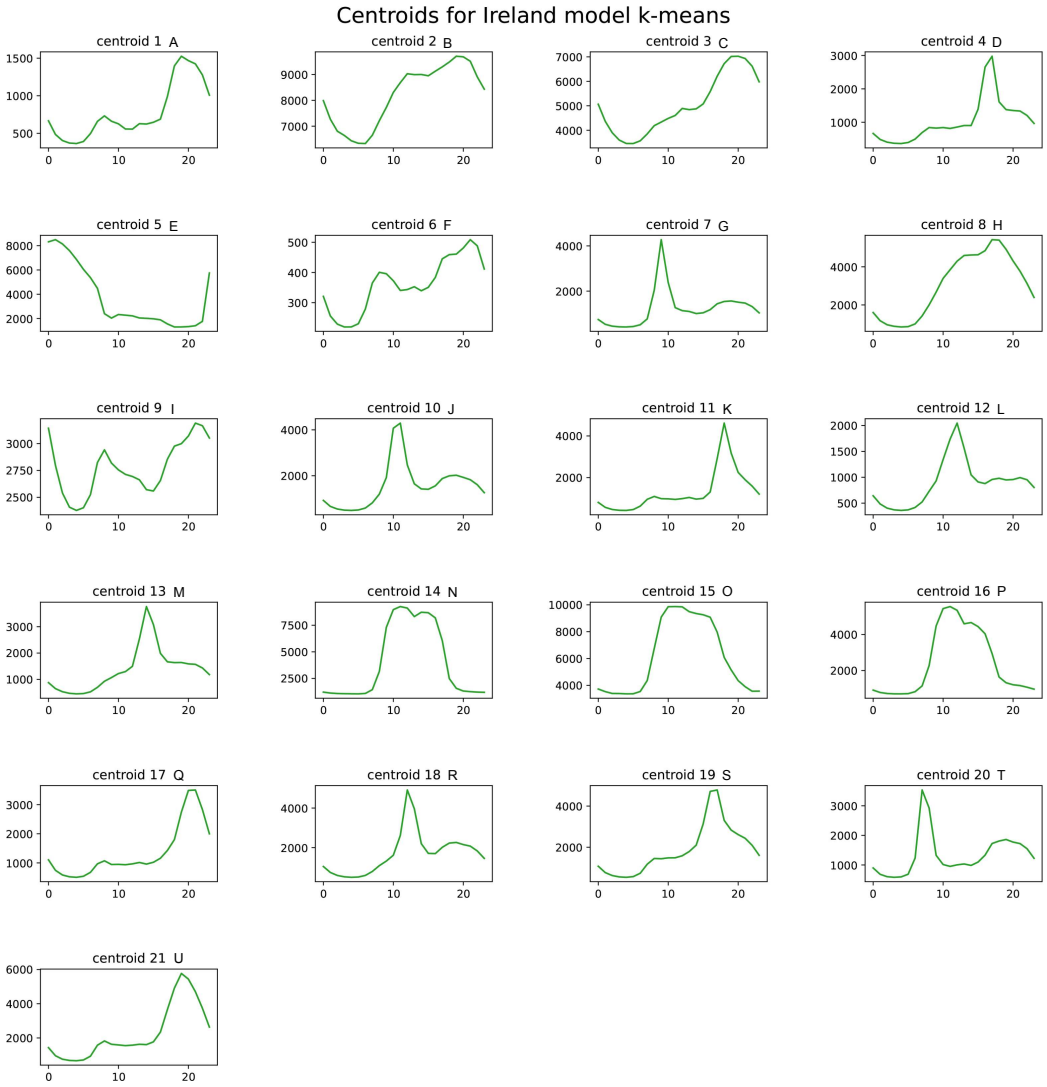


Figura 4: Clústeres de consumo energético obtenidos con K-medias [19].

En la Figura 5 se muestran superpuestas las gráficas de los diferentes clústeres, lo que permite observar con detalle qué grupos clasifican un mayor consumo y cuáles presentan un consumo más reducido según la franja horaria.

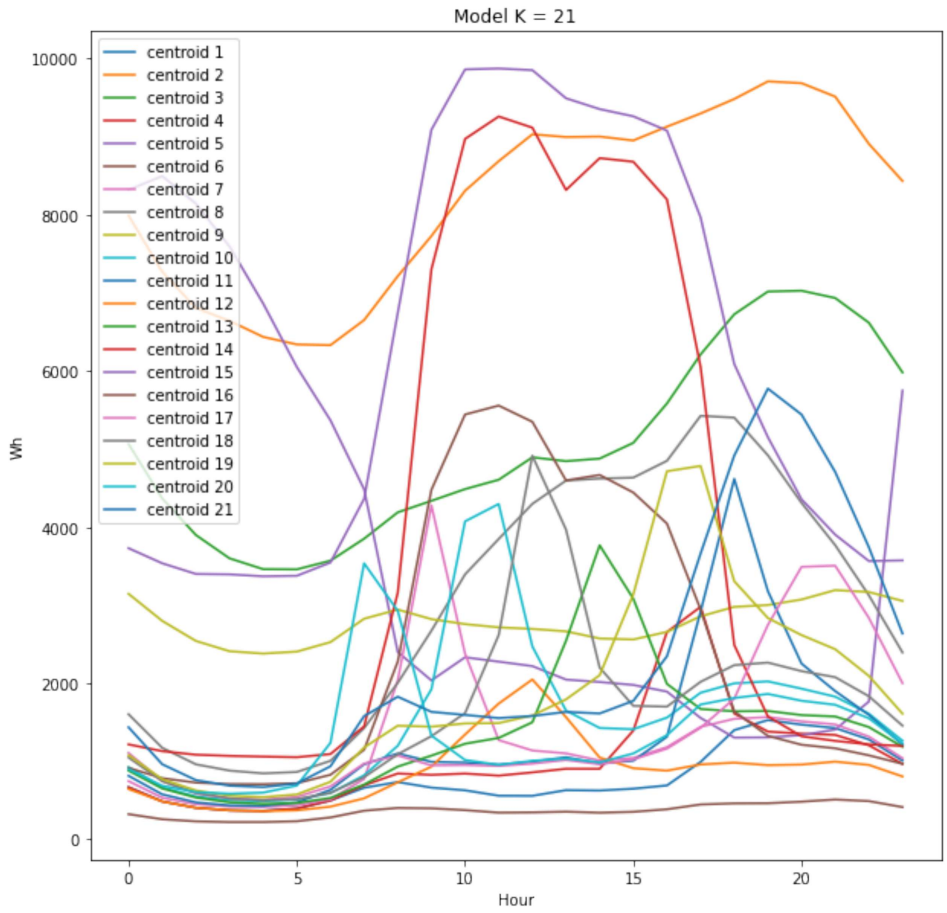


Figura 5: Clústeres para 21 tipos de perfiles de consumo [19].

Utilizando la distancia euclidiana entre el consumo diario de cada usuario y los 21 clústeres, se asigna la etiqueta del grupo más cercano, lo que permite determinar el comportamiento de consumo de cada día. De esta manera, la base de datos queda estructurada en tres columnas: identificador de usuario, día de consumo y clúster que representa el tipo de consumo de ese día.

Realizando un análisis estadístico básico, se puede examinar la frecuencia de cada clúster para usuarios individuales. Entre los grupos más asignados se encuentran los clústeres F y A con frecuencias de 748 779 (27 %) y 524 708 (18,9 %), respectivamente. Por otro lado, los grupos menos asignados fueron E y B con frecuencias de 2 818 (0,1 %) y 583 (0,02 %), respectivamente.

El objetivo previo a la sección de modelado es obtener una secuencia de patrones de consumo energético para cada usuario, es decir, se pretende conseguir una serie temporal categórica donde cada elemento de la secuencia represente los clústeres asignados a cada día de consumo. Adicionalmente, se diseñó una tabla de distribución de los clústeres donde los identificadores de los usuarios se encuentran en cada fila y en cada columna se muestra la frecuencia de aparición de ese clúster en los datos de consumo diarios del usuario correspondiente.

Para este estudio, se han establecido ciertos criterios en la selección de secuencias de patrones de consumo. Se ha fijado un número mínimo de 6 clústeres distintos y un máximo de 10 para asegurar un análisis general y realista. Con esto, se evitan las secuencias más repetitivas donde apenas varía el tipo de consumo, ya que es fácil conocer el tipo de consumo en esos casos y no es necesario modelarlo. También es importante establecer un número de días consecutivos con valores para las secuencias de clústeres. Para la fase de modelado, se requiere un conjunto de entrenamiento que siga ciertas pautas de consistencia, como mantener una longitud fija en las secuencias o asegurar un mismo contexto temporal, es decir, que todas las secuencias comiencen y terminen en el mismo día para todos los usuarios.

Por esta razón, se desarrolló un *script* para obtener la gráfica de la Figura 6, que muestra el número de usuarios en la base de datos que tienen datos de consumo diarios consecutivos durante secuencias de 90 hasta 537 días desde el día inicial.

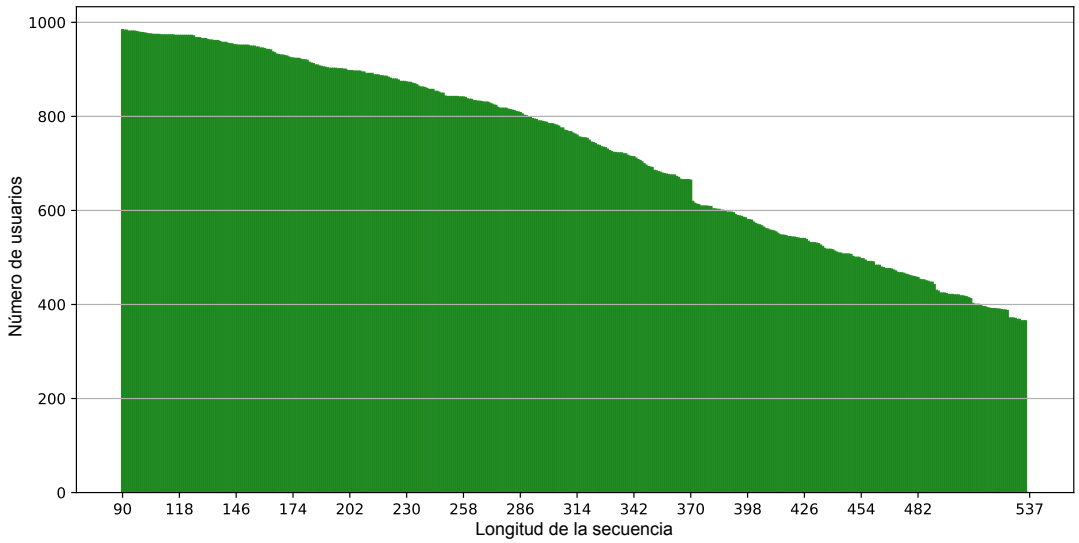


Figura 6: Histograma de usuarios con secuencias de consumo.

Se seleccionó una longitud de 400 días para las secuencias de patrones de consumo, ya que se consideró interesante estudiar un período de consumo superior a un año y entrenar los modelos de predicción en este contexto. Esto dio lugar a la creación de una base de datos con 493 usuarios, cada uno con su respectiva secuencia de patrones de consumo.

Finalmente, se generaron series temporales categóricas que se utilizarán como conjunto de entrenamiento para los modelos de predicción de secuencias. Graficar todas las secuencias de consumo a la vez resultaría poco práctico y claro. Sin embargo, en la Figura 7 se puede analizar la distribución de los clústeres más utilizados a lo largo de los días de la secuencia de 400 días.

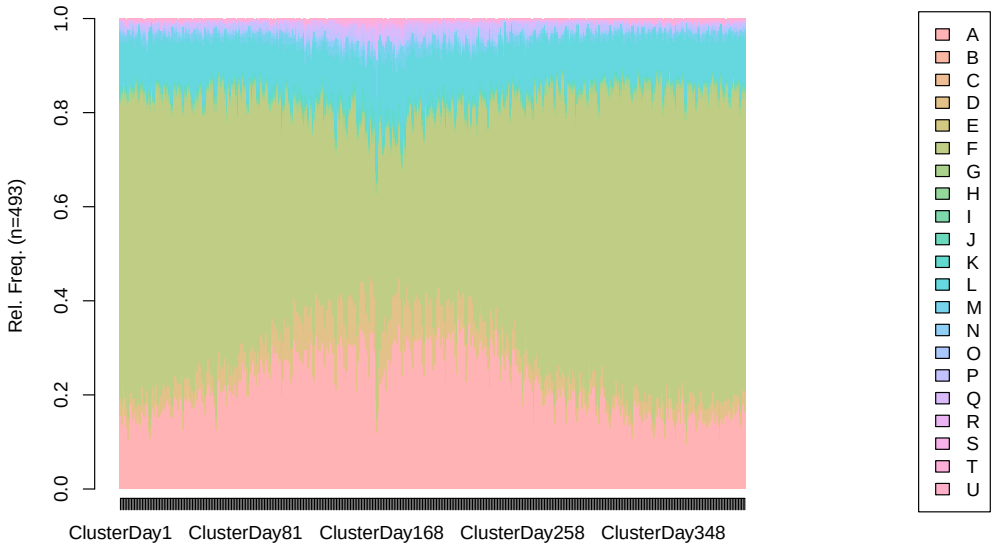


Figura 7: Distribución de los clústeres más presentes.

El grupo más utilizado, no sorprendentemente, es el tipo F, como se mencionó anteriormente. Sin embargo, cabe destacar la anomalía que se observa en los días donde hay una variación agresiva de los tipos de consumo. En la Figura 8, se puede observar cuál fue el grupo más frecuente en la secuencia de datos.

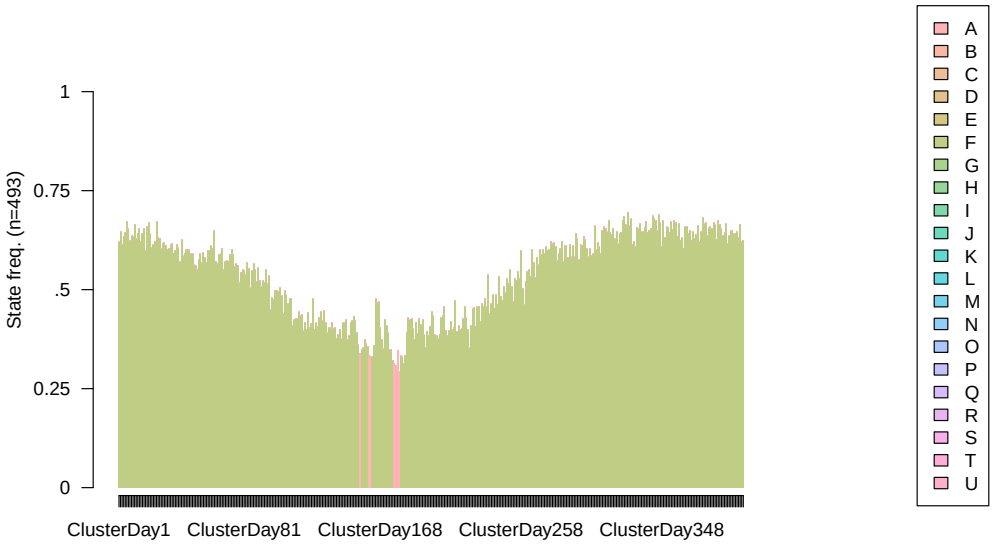


Figura 8: Clúster más frecuente en la secuencia completa.

Es posible que este caso se haya dado por una variación en el tiempo meteorológico durante esos días de consumo. De hecho, el clúster más frecuente pasó a ser A. En la Figura 4 se puede apreciar que existe una diferencia en el consumo en kWh entre el clúster A (número 1) y el clúster F (número 6).

Además, para otros modelos de predicción como RF (Random Forest) o CART, los datos se han preparado implementando un enfoque diferente que se basa en la aplicación de una **ventana deslizante** en la secuencia de consumo de un usuario. Este enfoque consiste en seleccionar una ventana de tamaño fijo que se desplaza a lo largo de la secuencia, avanzando una posición a la vez y seleccionando un subconjunto de elementos en cada paso. Es muy útil para analizar patrones locales dentro de una secuencia más amplia. Para este estudio en particular, se han utilizado ventanas de tamaño 7, 14, 21 y 28, correspondientes a intervalos semanales. Esto permite descubrir patrones de consumo más precisos o más generales según el tamaño de la ventana seleccionada. El proceso está diseñado para poder escoger un usuario de la base de datos y trabajar con su secuencia de consumo.

Entre la multitud de usuarios, se han seleccionado para un estudio particular aquellos que destacan por su diversidad en los hábitos de consumo. Esto implica evitar seleccionar secuencias de tipos de consumo repetitivas en las que predominen consistentemente un único tipo de consumo. Esta selección permite simular un escenario realista para el estudio de predicción de consumo, ya que refleja condiciones que son más representativas de la variabilidad natural del comportamiento del usuario. En concreto se han escogido los usuarios con identificadores 1022 y 5121 para este estudio.

La elección de estos usuario se ha realizado mediante un análisis estadístico donde se ha estudiado la distribución de patrones de consumo diarios a lo largo de sus secuencias de consumo. En las Figuras 9, 10 y 11, se muestran diferentes aspectos relevantes sobre el comportamiento del usuario con identificador 1022, como la distribución de distintos tipos de consumo que tiene el usuario a lo largo de la secuencia o la frecuencia de uso de los mismos.

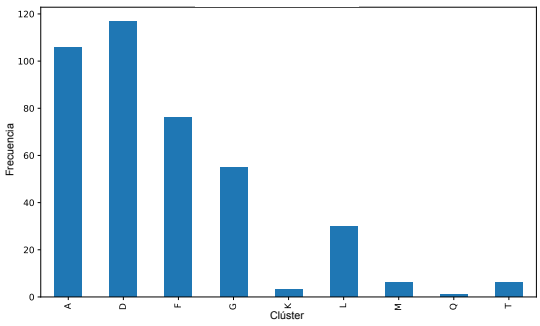


Figura 9: Histograma de clústeres para el usuario con identificador 1022.

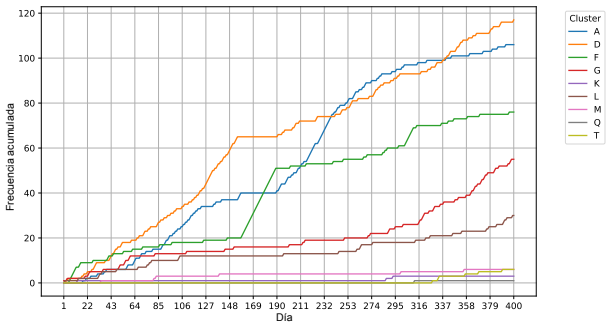


Figura 10: Variación de los clústeres a lo largo de la secuencia del usuario con identificador 1022.

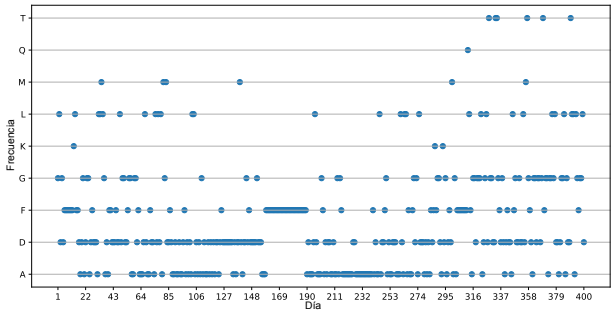


Figura 11: Valores de los clústeres a lo largo de la secuencia del usuario con identificador 1022.

Del mismo modo, se examinaron diferentes aspectos del comportamiento del usuario con identificador 5121, representados en las Figuras 12, 13 y 14. Estas figuras ofrecen información relevante sobre la distribución de varios tipos de consumo a lo largo de la secuencia temporal del usuario, así como la frecuencia de uso de cada uno de ellos.

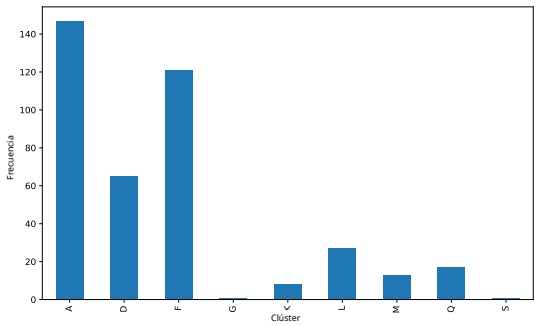


Figura 12: Histograma de clústeres para el usuario con identificador 5121.

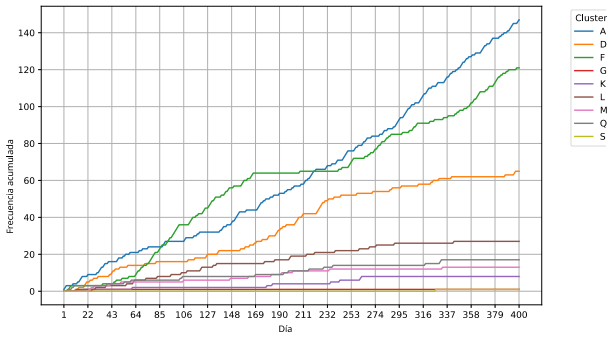


Figura 13: Variación de los clústeres a lo largo de la secuencia del usuario con identificador 5121.

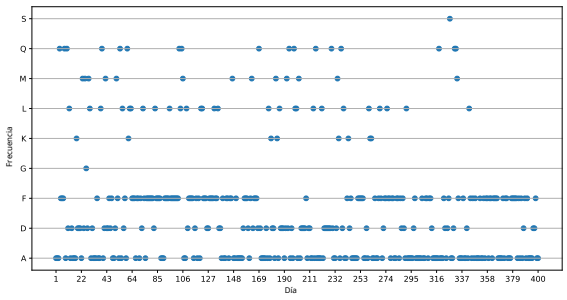


Figura 14: Valores de los clústeres a lo largo de la secuencia del usuario con identificador 5121.

Finalmente se ha ampliado este nuevo conjunto de datos basados en ventanas deslizantes con una columna adicional que indica el mes del año al que pertenece la secuencia estudiada. Durante los meses de invierno, es probable que se observen patrones de consumo energético distintos a los del verano, dado el uso común de sistemas de calefacción. Esto permite a los modelos basados en árboles de decisión mejorar la precisión en la fase de predicción, al permitirles clasificar con un parámetro adicional y relacionar secuencias de patrones que pertenecen a la misma época del año.

Para la asignación de meses, se sigue un esquema numérico que relaciona cada par de meses de la siguiente manera, como se puede ver en la Tabla 1.

Número	**Mes**	
1	Enero	Diciembre
2	Febrero	Noviembre
3	Marzo	Octubre
4	Abril	Septiembre
5	Mayo	Agosto
6	Junio	Julio

Cuadro 1: Codificación numérica de los meses del año.

De esta manera se relacionan intrínsecamente los meses cálidos y los fríos, lo que generalmente se traduce en una equivalencia en los patrones de consumo observados durante esos meses.

4

Modelado y Evaluación

Este capítulo abarca las dos siguientes etapas de la metodología CRISP-DM: Modelizado y Evaluación e interpretación de los resultados.
Se seleccionarán los modelos de minería de patrones secuenciales para realizar predicciones, seguidos de una comparación de los resultados según su metodología.

4.1. Modelado

En la sección 2 se han detallado los diversos modelos de predicción que se utilizarán en este estudio. Se dividirá el estudio de los modelos según su base de datos de entrenamiento y su estrategia de predicción.

- Modelos que utilizan toda la secuencia de datos. Parten de una secuencia de 400 días consecutivos de patrones de consumo y tienen como objetivo predecir el consumo del día siguiente basándose en patrones analizados durante la secuencia. Son modelos como Cadenas de Markov, Cadenas Ocultas de Markov, PST y TraMineR.

- Modelos que emplean ventana deslizante. Utilizan secuencias de 7, 14, 21 o 28 días utilizando la técnica de ventana deslizante y predicen el consumo del día siguiente en la secuencia. Se implementa con los modelos como Cadenas de Markov, DG, CPT, CART y RF.

- Modelos que identifican patrones en las secuencias de la base de datos. Los modelos utilizados ofrecen esta información, pero para disponer de un estudio analítico se ha utilizado el algoritmo PrefixSpan.

4.1.1. Modelos de Markov

Para la implementación del modelo de Cadena de Markov, se utilizó la biblioteca markovchain en R. El objetivo principal de este modelo es predecir el consumo del día siguiente basándose en una secuencia de 399 días previos. El modelo se ajustó utilizando el método de estimación de máxima verosimilitud (method = "mle").

Se presenta el modelo detallando primero un caso de uso para el usuario con identificador 1022. Se selecciona la secuencia de patrones de consumo de este usuario y se entrena el modelo de markov con la secuencia, excepto el último día de la secuencia. Se obtuvo la matriz de transición del modelo, que representa las probabilidades de transición entre los diferentes estados. En este caso el último elemento de la secuencia de entrenamiento (penúltimo de la secuencia completa) es el clúster L. Tomando este clúster como punto de partida se determinó la probabilidad de transición a cada posible estado siguiente. El resultado de la predicción consiste en seleccionar el estado siguiente con la probabilidad más alta. En la Figura 15 se resalta en rojo la fila del estado desde el cual parte la predicción y en verde el estado con mayor probabilidad de transición.

	A	D	F	G	K	L	M	Q	T
A	0.3773585	0.3490566	0.07547170	0.09433962	0.009433962	0.08490566	0.009433962	0.00000000	0.00000000
D	0.3362069	0.3189655	0.11206897	0.12931034	0.000000000	0.05172414	0.025862069	0.00000000	0.02586207
F	0.1052632	0.1710526	0.59210526	0.09210526	0.026315789	0.00000000	0.000000000	0.01315789	0.00000000
G	0.2000000	0.2727273	0.09090909	0.27272727	0.000000000	0.12727273	0.018181818	0.00000000	0.01818182
K	0.0000000	0.3333333	0.33333333	0.00000000	0.000000000	0.33333333	0.000000000	0.00000000	0.00000000
L	0.2413793	0.3793103	0.10344828	0.10344828	0.000000000	0.13793103	0.034482759	0.00000000	0.00000000
M	0.1666667	0.3333333	0.00000000	0.16666667	0.000000000	0.16666667	0.000000000	0.00000000	0.16666667
Q	0.0000000	0.0000000	0.00000000	0.00000000	0.000000000	1.00000000	0.000000000	0.00000000	0.00000000
T	0.0000000	0.0000000	0.16666667	0.50000000	0.000000000	0.16666667	0.000000000	0.00000000	0.16666667

Figura 15: Matriz de transición para el usuario 1022.

El siguiente estado según la predicción se corresponde con el clúster D. Sin embargo, es crucial no solo considerar la precisión, sino también el error estándar y el nivel de confianza que ofrece el modelo. El nivel de confianza se utiliza para calcular los intervalos de confianza alrededor de las estimaciones de las probabilidades de transición. En este caso, el modelo ofrece un nivel de confianza del 95 %. La matriz de errores estándar en la Figura 16 muestra los errores estándar asociados a las estimaciones de las probabilidades de transición entre estados en la cadena de Markov.

	A	D	F	G	K	L	M	Q	T
A	0.05966562	0.05738455	0.02668327	0.02983281	0.009433962	0.02830189	0.009433962	0.00000000	0.00000000
D	0.05383619	0.05243761	0.03108234	0.03338779	0.000000000	0.02111629	0.014931472	0.00000000	0.01493147
F	0.03721615	0.04744146	0.08826584	0.03481252	0.018608073	0.00000000	0.000000000	0.01315789	0.00000000
G	0.06030227	0.07041788	0.04065578	0.07041788	0.000000000	0.04810457	0.018181818	0.00000000	0.01818182
K	0.00000000	0.33333333	0.33333333	0.00000000	0.000000000	0.33333333	0.000000000	0.00000000	0.00000000
L	0.09123280	0.11436637	0.05972589	0.05972589	0.000000000	0.06896552	0.034482759	0.00000000	0.00000000
M	0.16666667	0.23570226	0.00000000	0.16666667	0.000000000	0.16666667	0.000000000	0.00000000	0.16666667
Q	0.00000000	0.00000000	0.00000000	0.00000000	0.000000000	1.00000000	0.000000000	0.00000000	0.00000000
T	0.00000000	0.00000000	0.16666667	0.28867513	0.000000000	0.16666667	0.000000000	0.00000000	0.16666667

Figura 16: Matriz de errores estándar para el usuario 1022.

El paquete igraph en R permite realizar un grafo representativo de la matriz de correlación donde cada estado se corresponde con cada clúster y cada arista con la probabilidad de transición entre estados, como se puede ver en la Figura 17.

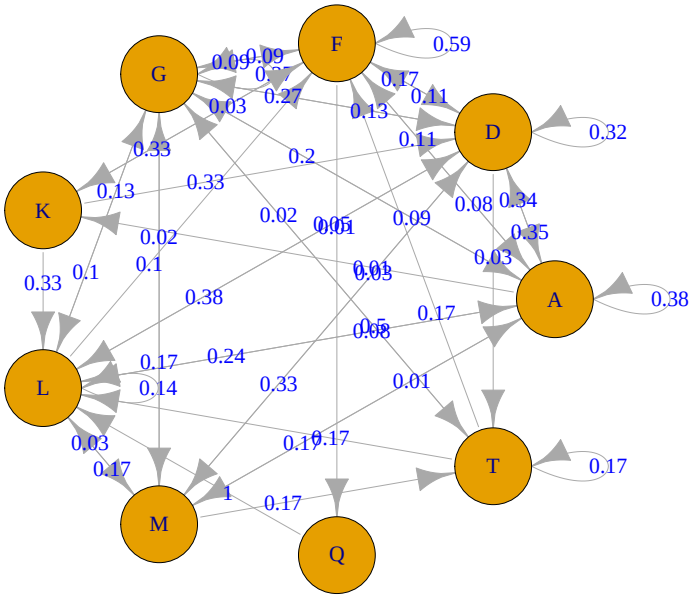


Figura 17: Grafo del modelo de Markov para el usuario 1022.

Aplicado este modelo a la base de datos descrita en la sección 3.3 se logra una precisión del 70,38 %.

El paquete markovchain en R permite aplicar el modelo entrenado para predecir secuencias de longitud determinada que continúen una secuencia dada utilizando la función *markovchainSequence*. Se aplicó este procedimiento para predecir una secuencia de 7 días que siguiera a los 393 días de datos de entrenamiento. Para el usuario con identificador 1022, se obtuvo la siguiente secuencia de patrones de consumo: (A, G, G, L, G, D, D). La cadena real para los clústeres de consumo asignados para los días 394 a 400 es: (L, G, F, G, G, L, D). Se consigue

una precisión del $42{,}85\,\%$. Sin embargo, es importante analizar los resultados y comparar las curvas de consumo de cada clúster en la predicción con la del clúster original. En la Figura 5 se puede apreciar una discrepancia entre las curvas de consumo de los clústeres D y L, evidenciando distintos patrones de consumo energético.

Generalizando este método para toda la base de datos, se puede utilizar no solo para predecir una secuencia de una semana, sino también para predecir un consumo de dos meses con una secuencia de 60 días a partir de los datos de consumo almacenados durante una larga secuencia de días. La precisión del modelo en este caso es del $50{,}41\,\%$. Aunque este resultado es mejorable, se puede explorar la precisión del modelo utilizando una matriz de confusión, una tabla que muestra la cantidad de errores que hay según la clase, lo que ayuda a identificar los errores de predicción en las secuencias. La matriz de confusión resultante de aplicar este modelo de predicción se muestra en la Figura 18.

	Reference																			
Prediction	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
A	1729	0	0	383	1	1590	52	0	0	5	72	389	62	1	0	3	57	7	3	11	0
B	0	1	0	0	0	0	0	0	1	0	0	0	0	2	4	0	0	0	0	2	0
C	0	0	37	0	0	0	0	12	21	0	0	0	0	0	11	4	0	0	0	0	0
D	275	0	0	187	0	326	12	1	0	4	21	100	26	1	0	1	15	1	1	6	0
E	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	4026	0	0	866	4	11799	150	1	2	65	202	1510	182	41	0	76	172	44	24	83	2
G	59	0	0	26	0	147	21	0	0	2	1	65	7	0	0	0	2	0	0	8	0
H	0	0	0	0	0	2	0	46	11	0	0	0	0	2	0	1	0	0	2	0	0
I	0	1	26	0	0	0	0	14	144	1	0	1	0	21	24	8	1	1	1	3	5
J	12	0	0	5	0	26	3	0	0	8	0	22	6	1	0	8	0	1	0	1	0
K	63	0	0	10	0	38	2	0	0	1	7	13	4	0	0	0	1	1	1	0	1
L	640	0	0	181	0	1102	45	0	0	55	27	625	66	5	0	70	15	12	0	14	0
M	57	0	0	29	0	98	2	2	0	6	4	28	75	9	0	36	2	5	3	1	0
N	0	0	0	0	0	1	0	4	11	2	0	0	0	17	11	10	0	0	0	2	0
O	0	2	9	0	0	0	0	3	22	0	0	0	0	21	51	2	0	0	0	9	0
P	6	0	0	0	18	34	0	2	2	2	0	2	9	24	0	123	0	2	3	0	0
Q	36	0	0	5	0	52	3	0	0	0	2	5	4	0	0	0	1	0	0	1	1
R	11	0	0	3	0	9	3	0	0	1	1	8	4	1	0	2	0	1	0	0	0
S	1	0	2	1	0	4	0	4	0	0	0	0	0	1	0	2	0	0	1	0	0
T	45	0	0	24	0	85	9	0	0	0	1	10	1	0	0	0	0	0	0	39	0
U	1	0	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figura 18: Matriz de confusión del modelo de Markov para 60 días.

Los elementos de la diagonal se corresponden con el número de aciertos al predecir ese tipo de consumo. Se observa un gran número de errores entre ciertos grupos, como por ejemplo los clústeres A y F.

Para abordar esta imprecisión en la predicción, se puede ampliar el método aplicando una matriz de correlación. Esta herramienta permite evaluar la gravedad de los errores de predicción al correlacionar los distintos tipos de consumo. Se calculan las distancias euclidianas entre las gráficas de consumo de los clústeres en la Figura 5, y se obtiene un coeficiente de correlación entre ellos, lo que ayuda a identificar la relación entre los distintos tipos de consumo.

En la Figura 19 se presenta la matriz de correlación calculada.

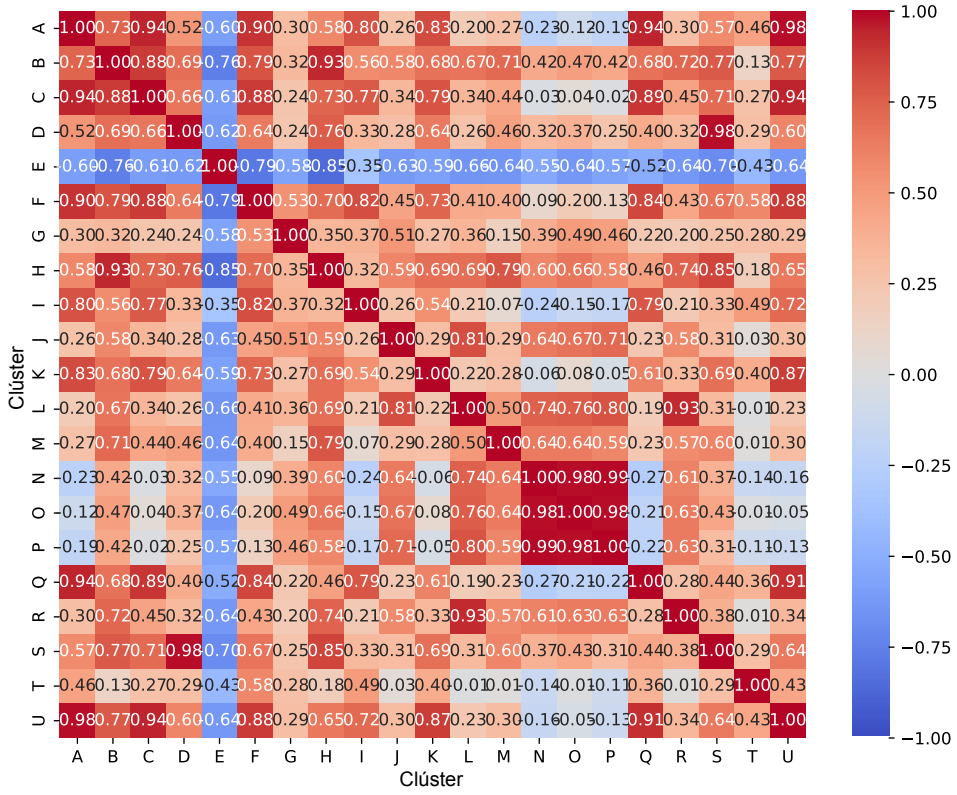


Figura 19: Matriz de correlación de los clústeres.

Existe una correlación inversa entre el tipo de consumo E y el resto de los clústeres, lo cual se evidencia claramente en la Figura 4. El clúster E representa un consumo muy reducido a lo largo del día, a diferencia del comportamiento de los otros grupos. Por otro lado, se observa una correlación muy alta entre otros clústeres de la lista.

Al aplicar esta matriz de correlación, con un valor mínimo de correlación de $0{,}9$, podemos identificar que las predicciones son válidas si los clústeres predichos tienen una alta correlación con el clúster real. Según los expertos en el dominio, es lógico dar por correctas aquellas predicciones que, aun no siendo exactas, implican un tipo de consumo similar. Esta forma de considerar los errores, conlleva un cambio a la hora de medir la la precisión del modelo, que para el caso de cadenas de Markov llega al $70\,\%$. La matriz de confusión resultante se muestra en la Figura 20

Prediction \ Reference	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
A	3376	0	0	383	1	0	52	0	0	5	72	389	62	1	0	3	0	7	3	11	0
B	0	1	0	0	0	0	0	0	1	0	0	0	0	2	4	0	0	0	0	2	0
C	0	0	37	0	0	0	0	12	21	0	0	0	0	0	11	4	0	0	0	0	0
D	275	0	0	188	0	326	12	1	0	4	21	100	26	1	0	1	15	1	0	6	0
E	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	866	4	15825	150	1	2	65	202	1510	182	41	0	76	172	44	24	83	2
G	59	0	0	26	0	147	21	0	0	2	1	65	7	0	0	0	2	0	0	8	0
H	0	0	0	0	0	2	0	46	11	0	0	0	0	2	0	1	0	0	2	0	0
I	0	1	26	0	0	0	0	14	144	1	0	1	0	21	24	8	1	1	1	3	5
J	12	0	0	5	0	26	3	0	0	8	0	22	6	1	0	8	0	1	0	1	0
K	63	0	0	10	0	38	2	0	0	1	7	13	4	0	0	0	1	1	1	0	1
L	640	0	0	181	0	1102	45	0	0	55	27	637	66	5	0	70	15	0	0	14	0
M	57	0	0	29	0	98	2	2	0	6	4	28	75	9	0	36	2	5	3	1	0
N	0	0	0	0	0	1	0	4	11	2	0	0	0	38	0	0	0	0	0	2	0
O	0	2	9	0	0	0	0	3	22	0	0	0	0	0	74	0	0	0	0	9	0
P	6	0	0	0	18	34	0	2	2	2	0	2	9	0	0	147	0	2	3	0	0
Q	0	0	0	5	0	52	3	0	0	0	2	5	4	0	0	0	38	0	0	1	0
R	11	0	0	3	0	9	3	0	0	1	1	0	4	1	0	2	0	9	0	0	0
S	1	0	2	0	0	4	0	4	0	0	0	0	0	1	0	2	0	0	2	0	0
T	45	0	0	24	0	85	9	0	0	0	1	10	1	0	0	0	0	0	0	39	0
U	0	0	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Figura 20: Matriz de confusión del modelo de Markov para 60 días.

En esta matriz, se observa una mejora en la asignación de los tipos de consumo, ya que se consideran como correctas las predicciones que resultaron en clústeres muy similares a los reales. Sin embargo, hay 23 patrones de tipo E, el cual nunca es predicho en ningún caso (ni por acierto, ni por fallo). Se confunde mucho con el patrón P.

De igual forma que para esta técnica de predicción de secuencias el uso de la matriz de correlación ha incrementado considerablemente la precisión del modelo, el resultado de aplicar este método en el primer modelo de cadenas de Markov descrito, donde se entrena el modelo con una cadena de 399 días y se predice el siguiente elemento en la secuencia, es de un $83{,}16\,\%$.

Paralelamente se ha realizado un estudio del Modelo de Markov donde se ha evaluado la precisión del modelo en función de la longitud de las secuencias de entrenamiento. Concretamente se han analizado todas las longitudes de secuencia desde 300 días consecutivos de consumo hasta 399 días utilizados en el modelo detallado anteriormente. En las Figuras 21 y 22 se muestra la evolución de la precisión de los modelos.

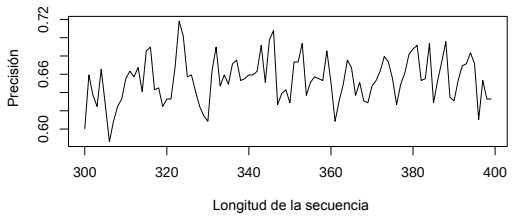


Figura 21: Variación de la precisión de modelos de Markov.

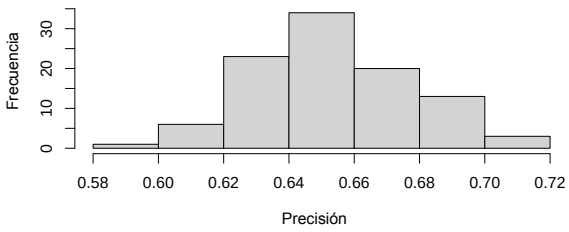


Figura 22: Histograma de la precisión de modelos de Markov.

La conclusión que se puede extraer de estas dos gráficas es que, con secuencias de días de consumo tan extensas, la longitud de la secuencia no resulta ser un factor determinante para la precisión del modelo creado.

4.1.2. Cadenas Ocultas de Markov

Para la implementación del modelo de Cadena Oculta de Markov, se utilizó la biblioteca HMM en R. Este modelo, al igual que el método anterior, tiene como objetivo predecir el consumo del día siguiente basándose en una secuencia de 399 días previos.

El algoritmo de Baum-Welch se emplea para ajustar el modelo HMM, con el fin de aprender las probabilidades de transición entre estados y las probabilidades de emisión asociadas a cada estado. Este proceso implica la iteración de pasos de estimación y maximización para encontrar los parámetros del modelo que mejor se ajusten a los datos observados.

Una vez entrenado el modelo, se puede utilizar para predecir el próximo estado en la secuencia. Por ejemplo, para el usuario con identificador 1022, dado el último estado observado en la secuencia de consumo (L) se calculan las probabilidades de transición a cada posible estado siguiente, como se puede observar en la Figura 23. El siguiente estado según la predicción se corresponde con el tipo de consumo D, pues tiene mayor probabilidad.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
L	0.2656642	0	0	0.2907268	0	0.1904762	0.1378446	0	0	0	0.007518797	0.07518797	0.01503759	0	0	0	0.002506266	0	0	0.01503759	0

Figura 23: Probabilidades de transición en HMM para usuario 1022

La generalización de este modelo a todas las secuencias de la base de datos proporciona una predicción global con una precisión de 52,94 %. Cabe destacar que es un proceso lento, costoso y que tampoco proporciona una precisión muy buena. Es posible tener una precisión más ajustada a las necesidades de los expertos considerando de nuevo el concepto de matriz de correlación de los patrones de consumo. De esta manera, como se ha descrito previamente, se darán por válidas las predicciones de tipos de consumo que sean parecidos entre sí.
La predicción global de las secuencias utilizando la matriz de correlación proporciona una precisión del 72,21 %.

4.1.3. Árboles Sufijo Probabilístico (PST)

Para la implementación del modelo de Árboles Sufijos Probabilísticos se utilizó la biblioteca PST en R. Este modelo tiene como objetivo predecir el consumo del día siguiente basándose en una secuencia de 399 días previos. Para construir un PST se analiza la secuencia de datos y se crea un árbol que representa las frecuencias y probabilidades de aparición de diferentes subcadenas (patrones) dentro de la secuencia. Primero se define un valor para el parámetro profundidad máxima del árbol (L). Se eligió una profundidad máxima del árbol de 14 niveles para capturar patrones de consumo que abarquen como máximo dos semanas. Esta decisión se basa en la suposición de que los patrones de consumo relevantes pueden extenderse hasta ese periodo de tiempo. El consumo de la vivienda se va a asemejar más al patrón de consumo reciente que a uno más distante en el tiempo.

Una vez construido el PST con estos parámetros, se puede utilizar para predecir el siguiente estado en la secuencia. Se obtienen las probabilidades de los posibles estados siguientes y se selecciona el estado más probable como la predicción para el siguiente elemento en la secuencia. Se aplica el modelo para el usuario con identificador 1022. En la Figura 24 se presenta la lista de probabilidades de transición entre estados desde el último día de la secuencia de entrenamiento. En este caso, el modelo de PST ha considerado que el siguiente estado debe ser el patrón de consumo A.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
0.6666667	0	0	0	0	0	0.3333333	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figura 24: Probabilidades de transición en PST para usuario 1022

Aplicando este modelo utilizando toda la base de datos se consigue una precisión del 67,34 %. La matriz de confusión correspondiente se puede visualizar en la Figura 25.

	Reference																
Prediction	A	C	D	F	G	H	I	J	K	L	M	N	O	P	Q	R	T
A	44	0	7	32	0	0	0	0	1	8	0	0	0	0	0	0	0
C	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	8	0	1	7	0	0	0	1	1	3	1	0	0	0	0	0	0
F	32	0	5	256	3	0	0	0	1	18	1	0	0	0	1	0	0
G	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	1	0
J	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
K	0	0	0	0	0	0	0	0	0	2	0	0	0	0	1	0	0
L	2	0	2	9	2	0	0	1	0	15	0	0	0	0	0	0	0
M	1	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0
N	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
O	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0
P	0	0	0	1	0	0	0	0	0	1	0	1	0	3	0	0	0
Q	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
R	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
T	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2

Figura 25: Matriz de confusión para modelo PST.

Aplicando el método de la matriz de correlación a este modelo permite que se den por válidos los tipos de consumo que más se asemejan. La nueva matriz de confusión en la Figura 26 relativa a este modelo corrige las asignaciones que previamente se habían asignado a otros clústeres.

	Reference																
Prediction	A	C	D	F	G	H	I	J	K	L	M	N	O	P	Q	R	T
A	76	0	7	0	0	0	0	0	1	8	0	0	0	0	0	0	0
C	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	8	0	1	7	0	0	0	1	1	3	1	0	0	0	0	0	0
F	0	0	5	288	3	0	0	0	1	18	1	0	0	0	1	0	0
G	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	1	0
J	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
K	0	0	0	0	0	0	0	0	0	2	0	0	0	0	1	0	0
L	2	0	2	9	2	0	0	1	0	16	0	0	0	0	0	0	0
M	1	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0
N	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
O	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0
P	0	0	0	1	0	0	0	0	0	1	0	0	0	3	0	0	0
Q	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2

Figura 26: Matriz de confusión para modelo PST con método matriz de correlación.

Si usamos este nuevo modelo y recalculamos la precisión, considerando como acierto el predecir una curva muy próxima a la real, se obtiene un valor más alto: 80,73 %.

4.1.4. TraMineR

Este algoritmo se implementa mediante el uso de la librería TraMineR, disponible en R. Esta herramienta permite generar cadenas que representan secuencias de tipos de consumo diario y crear sus respectivas matrices de transición entre los diferentes estados y se ha utilizado para predecir el tipo de consumo de los usuarios para el siguiente día que continúa a la secuencia. Veamos el funcionamiento de este modelo de nuevo para el usuario con identificador 1022 con longitud de secuencia como en los modelos anteriores. Al introducir la secuencia de consumo de este usuario se obtiene la siguiente matriz de transición de estados, ver Figura 27.

	[-> A]	[-> B]	[-> C]	[-> D]	[-> E]	[-> F]	[-> G]	[-> H]	[-> I]	[-> J]	[-> K]	[-> L]	[-> M]	[-> N]	[-> O]	[-> P]	[-> Q]	[-> R]	[-> S]	[-> T]	[-> U]
[A ->]	0.3773585	0	0	0.3490566	0	0.07547170	0.09433962	0	0	0	0.009433962	0.08490566	0.009433962	0	0	0	0.00000000	0	0	0.00000000	0
[B ->]	0.0000000	0	0	0.0000000	0	0.00000000	0.00000000	0	0	0	0.000000000	0.00000000	0.000000000	0	0	0	0.00000000	0	0	0.00000000	0
[C ->]	0.0000000	0	0	0.0000000	0	0.00000000	0.00000000	0	0	0	0.000000000	0.00000000	0.000000000	0	0	0	0.00000000	0	0	0.00000000	0
[D ->]	0.3362069	0	0	0.3189655	0	0.11206897	0.12931034	0	0	0	0.000000000	0.05172414	0.025862069	0	0	0	0.00000000	0	0	0.02586207	0
[E ->]	0.0000000	0	0	0.0000000	0	0.00000000	0.00000000	0	0	0	0.000000000	0.00000000	0.000000000	0	0	0	0.00000000	0	0	0.00000000	0
[F ->]	0.1052632	0	0	0.1710526	0	0.59210526	0.09210526	0	0	0	0.026315789	0.00000000	0.000000000	0	0	0	0.01315789	0	0	0.00000000	0
[G ->]	0.2037037	0	0	0.2777778	0	0.09259259	0.27777778	0	0	0	0.000000000	0.11111111	0.018518519	0	0	0	0.00000000	0	0	0.01851852	0
[H ->]	0.0000000	0	0	0.0000000	0	0.00000000	0.00000000	0	0	0	0.000000000	0.00000000	0.000000000	0	0	0	0.00000000	0	0	0.00000000	0
[I ->]	0.0000000	0	0	0.0000000	0	0.00000000	0.00000000	0	0	0	0.000000000	0.00000000	0.000000000	0	0	0	0.00000000	0	0	0.00000000	0
[J ->]	0.0000000	0	0	0.0000000	0	0.00000000	0.00000000	0	0	0	0.000000000	0.00000000	0.000000000	0	0	0	0.00000000	0	0	0.00000000	0
[K ->]	0.0000000	0	0	0.3333333	0	0.33333333	0.00000000	0	0	0	0.000000000	0.33333333	0.000000000	0	0	0	0.00000000	0	0	0.00000000	0
[L ->]	0.2413793	0	0	0.3793103	0	0.10344828	0.10344828	0	0	0	0.000000000	0.13793103	0.034482759	0	0	0	0.00000000	0	0	0.00000000	0
[M ->]	0.1666667	0	0	0.3333333	0	0.00000000	0.16666667	0	0	0	0.000000000	0.16666667	0.000000000	0	0	0	0.00000000	0	0	0.16666667	0
[N ->]	0.0000000	0	0	0.0000000	0	0.00000000	0.00000000	0	0	0	0.000000000	0.00000000	0.000000000	0	0	0	0.00000000	0	0	0.00000000	0
[O ->]	0.0000000	0	0	0.0000000	0	0.00000000	0.00000000	0	0	0	0.000000000	0.00000000	0.000000000	0	0	0	0.00000000	0	0	0.00000000	0
[P ->]	0.0000000	0	0	0.0000000	0	0.00000000	0.00000000	0	0	0	0.000000000	0.00000000	0.000000000	0	0	0	0.00000000	0	0	0.00000000	0
[Q ->]	0.0000000	0	0	0.0000000	0	0.00000000	0.00000000	0	0	0	0.000000000	1.00000000	0.000000000	0	0	0	0.00000000	0	0	0.00000000	0
[R ->]	0.0000000	0	0	0.0000000	0	0.00000000	0.00000000	0	0	0	0.000000000	0.00000000	0.000000000	0	0	0	0.00000000	0	0	0.00000000	0
[S ->]	0.0000000	0	0	0.0000000	0	0.00000000	0.00000000	0	0	0	0.000000000	0.00000000	0.000000000	0	0	0	0.00000000	0	0	0.00000000	0
[T ->]	0.0000000	0	0	0.0000000	0	0.16666667	0.50000000	0	0	0	0.000000000	0.16666667	0.000000000	0	0	0	0.00000000	0	0	0.16666667	0
[U ->]	0.0000000	0	0	0.0000000	0	0.00000000	0.00000000	0	0	0	0.000000000	0.00000000	0.000000000	0	0	0	0.00000000	0	0	0.00000000	0

Figura 27: Matriz de transición TraMineR.

El último estado de la secuencia de entrenamiento es el clúster L, luego el modelo escoge transitar al estado con mayor probabilidad, que es el tipo de consumo D.

El resultado de aplicar este modelo a todo el conjunto de datos proporciona una precisión del $60{,}49\,\%$. Se ha recurrido nuevamente al método de la matriz de correlación. Esta técnica permite analizar si el modelo tiende a equivocarse entre tipos de consumo similares. Al combinar el enfoque de TraMineR con el análisis de la matriz de correlación entre estados, se obtiene un valor significativamente más alto: $72{,}3\,\%$.

4.1.5. Modelos de Markov con ventana deslizante

A partir de ahora se trabajará con otra base de datos distinta que se detalló en el apartado 3.3, basada en el método de ventana deslizante y aplicado a los usuarios con identificadores 1022 y 5121. Inicialmente se han escogido tamaños de ventana de 14 días, ya que se considera que contienen información suficiente de manera local como para poder utilizarse para predecir un consumo próximo.

El modelo de Markov para secuencias de 14 días para el usuario con identificador 1022 ofrece una precisión del 40,56 %. En la Figura 28 se presenta la matriz de confusión de este modelo.

Prediction	Reference A	D	F	G	K	L	M	Q	T
A	70	47	19	24	0	13	3	0	1
D	23	44	12	18	2	9	2	0	3
F	6	4	35	1	0	2	0	1	0
G	3	17	2	8	0	5	0	0	2
K	0	1	0	0	0	0	0	0	0
L	2	2	1	1	0	0	1	0	0
M	1	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	0	0	0	0
T	1	0	0	1	0	0	0	0	0

Figura 28: Valores de los clústeres a lo largo de la secuencia del usuario con identificador 1022.

Aplicando de nuevo el método de matriz de correlación, se obtiene una precisión del 47,03 %. La Figura 29 muestra la nueva matriz de confusión del modelo, que corrige la asignación de algunos clústeres con respecto a la predicción anterior.

Prediction	Reference A	D	F	G	K	L	M	Q	T
A	76	47	0	24	0	13	3	0	1
D	23	44	12	18	2	9	2	0	3
F	0	4	54	1	0	2	0	1	0
G	3	17	2	8	0	5	0	0	2
K	0	1	0	0	0	0	0	0	0
L	2	2	1	1	0	0	1	0	0
M	1	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	0	0	0	0
T	1	0	0	1	0	0	0	0	0

Figura 29: Valores de los clústeres a lo largo de la secuencia del usuario con identificador 1022.

Par el usuario con identificador 5121, se obtienen los siguientes resultados al aplicar el modelo de Markov con ventana deslizante. La matriz de confusión correspondiente se muestra en la Figura 30

	Reference								
Prediction	A	D	F	G	K	L	M	Q	S
A	86	39	62	0	8	14	5	8	0
D	15	13	7	1	0	6	4	3	1
F	35	8	45	0	0	5	2	3	0
G	0	0	0	0	0	0	0	0	0
K	2	2	1	0	0	0	0	0	0
L	2	0	3	0	0	1	1	0	0
M	0	1	0	0	0	0	0	0	0
Q	2	1	0	0	0	0	1	0	0
S	0	0	0	0	0	0	0	0	0

Figura 30: Valores de los clústeres a lo largo de la secuencia del usuario con identificador 5121.

La precisión obtenida es del 37,47 %, lo que indica un rendimiento insatisfactorio del modelo. Al aplicar el método de matriz de correlación, se logra mejorar la precisión del modelo al 65,37 %. Esta forma de considerar los errores de predicción representa una mejora significativa, ya que los patrones de consumo de este usuario presentan curvas de consumo energético similares. El modelo da por acertadas las predicciones de clústeres que representan tipos de consumo similares. La nueva matriz de correlación se muestra en la Figura 31.

	Reference								
Prediction	A	D	F	G	K	L	M	Q	S
A	123	39	0	0	8	14	5	0	0
D	15	13	7	1	0	6	4	3	0
F	0	8	107	0	0	5	2	3	0
G	0	0	0	0	0	0	0	0	0
K	2	2	1	0	0	0	0	0	0
L	2	0	3	0	0	1	1	0	0
M	0	1	0	0	0	0	0	0	0
Q	0	1	0	0	0	0	1	8	0
S	0	0	0	0	0	0	0	0	1

Figura 31: Valores de los clústeres a lo largo de la secuencia del usuario con identificador 5121.

Analizando las dos matrices de confusión se puede observar que se realiza una mejor predicción aplicando el método con los clústeres de consumo correlados.

En el caso de la función *markovSequence* se utilizan como datos de entrenamiento secuencias de 49 días (7 semanas). Se predicen patrones de consumo para la próxima semana. Es fácil comprobar la precisión del modelo comparando con los 7 primeros elementos de la siguiente secuencia, pues utiliza el método de ventana deslizante. Se ha utilizado el usuario con identificador 5121. Se han predicho un total de 352 secuencias donde el modelo de Markov predice los patrones de consumo con una precisión de un $30{,}29$ %. En la Figura 32 se observa la matriz de confusión para todos los patrones de las secuencias predichas.

	Reference								
Prediction	A	D	F	G	K	L	M	Q	S
A	387	160	287	2	17	66	31	41	1
D	192	112	128	2	14	41	20	23	0
F	227	89	323	0	7	61	13	24	1
G	4	0	3	1	1	1	1	3	0
K	26	5	22	0	5	4	1	4	0
L	69	28	82	1	12	17	11	2	0
M	31	27	36	0	1	11	6	14	0
Q	48	31	41	1	3	8	6	13	0
S	5	1	0	0	0	0	0	0	1

Figura 32: Valores de los clústeres a lo largo de la secuencia del usuario con identificador 5121.

Aplicando la matriz de correlación al modelo de Markov se incrementa la precisión del modelo a un $54{,}09$ %. En la Figura 33 se observa la matriz de confusión para todos los patrones de las secuencias predichas por este nuevo modelo.

	Reference								
Prediction	A	D	F	G	K	L	M	Q	S
A	715	160	0	2	17	66	31	0	1
D	192	112	128	2	14	41	20	23	0
F	0	89	550	0	7	61	13	24	1
G	4	0	3	1	1	1	1	3	0
K	26	5	22	0	5	4	1	4	0
L	69	28	82	1	12	17	11	2	0
M	31	27	36	0	1	11	6	14	0
Q	0	31	41	1	3	8	6	61	0
S	5	0	0	0	0	0	0	0	2

Figura 33: Valores de los clústeres a lo largo de la secuencia del usuario con identificador 5121.

Se corrige considerablemente la predicción, pues los tipos de consumo A y F están altamente correlacionados. El modelo da por válida la predicción en estos casos y eso proporciona una mejora en la predicción de resultados.

4.1.6. Árboles Compactos Predictivos (CPT)

Este modelo de predicción de secuencias de patrones forma parte de un conjunto de algoritmos de minería de patrones conocido como SPMF [20], el cual está disponible en el lenguaje de programación Java.

Para ejecutar el algoritmo, es necesario configurar la base de datos de tal manera que las secuencias de letras sea de nuevo la numeración de clústeres, asignando a cada letra un número correspondiente, como se puede observar en la Figura 4. Por ejemplo, el tipo de consumo A se corresponde con el número 1 y el tipo U con el 21. Cada secuencia se formatea con el delimitador -1 entre los números y -1 -2 al final para indicar el final de la secuencia. El código implementado para el modelo CPT proporciona información sobre las secuencias leídas del conjunto de datos, como por ejemplo que la media de clústeres diferentes en las secuencias de 14 días es 4 y que la media de aparición de cada clúster en cada secuencia es de 3 veces.

Se han utilizado secuencias bisemanales de patrones de consumo del usuario con identificador 5121 utilizando el método de ventana deslizante. El modelo CPT predice correctamente el tipo de consumo del decimocuarto día con una precisión del $40{,}82\,\%$. Si bien este resultado puede mejorarse, se ha utilizado el método de la matriz de correlación para considerar válidos los tipos de consumo más correlacionados con el clúster real. Esto ha incrementado la precisión de la predicción a un $41{,}08\,\%$.

4.1.7. Grafos Degenerativos (DG)

Este modelo de predicción también forma parte del paquete SPMF y utiliza la misma base de datos formateada como se ha mencionado anteriormente.

Al ejecutar el algoritmo para secuencias de patrones de 14 días de consumo del usuario con identificador 5121, la tasa de acierto al predecir el tipo de consumo del decimocuarto día fue de un $40{,}57\,\%$. Este modelo ha resultado ser de los más rápidos en tiempos de ejecución, aunque no destaca por su alta precisión. Al introducir el método de la matriz de correlación se logró incrementar la tasa de acierto a un $40{,}83\,\%$.

En ambos modelos CPT y DG, la aplicación del método de la matriz de correlación apenas ha mejorado la precisión. Esto se debe a que los tipos de consumo predichos no guardaban correlación fuerte con los clústeres reales, lo que resultó en una imprecisión significativa.

4.1.8. Árbol de Clasificación y Regresión (CART) y Random Forest

En este estudio se comparan sus rendimientos en distintas configuraciones y con la adición de nuevas características. La base de datos consiste en secuencias de tipos de consumo durante 7, 14, 21 y 28 días consecutivos, utilizando el método de ventana deslizante. En el caso de Random Forest se ha utilizado un conjunto de 100 árboles para mejorar la robustez del modelo. El objetivo es comparar la eficacia de los algoritmos en función de la información de patrones de consumo que tiene de cada usuario. En ejemplo particular, se aplicarán los modelos a las secuencias de consumo del usuario con identificador 5121. En la Figura 34 se puede visualizar la evolución de la precisión de estos modelos.

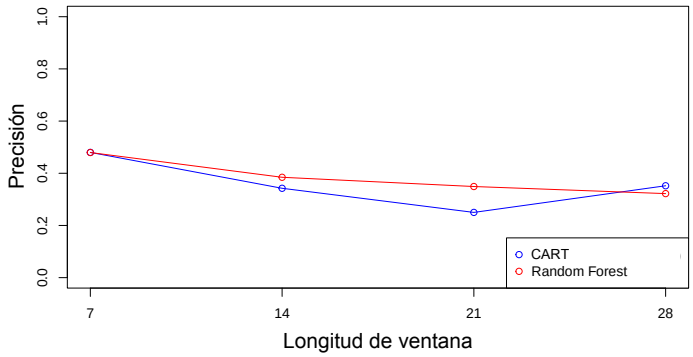


Figura 34: Evolución de la precisión de los modelos para el usuario 5121.

Modelo	**CART**	**RandomForest**
Secuencia 7 días	48 %	47,95 %
Secuencia 14 días	34,25 %	38,46 %
Secuencia 21 días	25 %	34,92 %
Secuencia 28 días	35,21 %	32,20 %

Cuadro 2: Resultados del modelo base para el usuario 5121.

Se añade una nueva característica a los datos, representada por el mes al que pertenece cada secuencia. Se ha ampliado el conjunto de datos basados en ventanas deslizantes con una columna adicional que indica el mes del año al que pertenece la secuencia estudiada, como se ha detallado en la sección 3.3.

En la Figura 35 se puede visualizar la evolución de la precisión de estos modelos utilizando el método de identificación mensual.

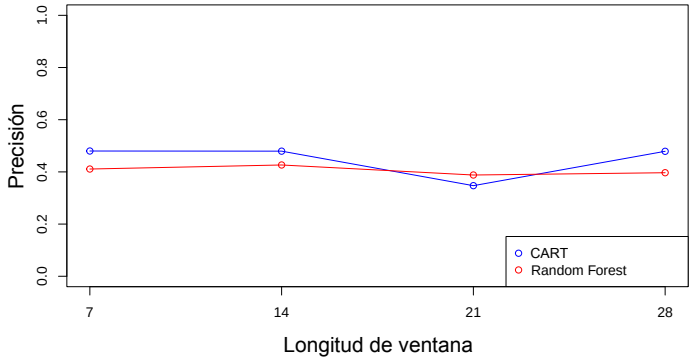


Figura 35: Valores de los clústeres a lo largo de la secuencia del usuario con identificador 5121.

Modelo	**CART**	**RandomForest**
Secuencia 7 días	48 %	0,4109589 %
Secuencia 14 días	47,95 %	42,65 %
Secuencia 21 días	34,72 %	38,81 %
Secuencia 28 días	47,89 %	39,68 %

Cuadro 3: Resultados con adición del mes para el usuario 5121.

Además de la adición del mes como característica, se utiliza el método aplicado en los modelos anteriores de la matriz de correlación para mejorar la precisión del modelo. En la Figura 36 se puede visualizar la evolución de la precisión de estos modelos aplicando los métodos de asignación mensual y matriz de correlación.

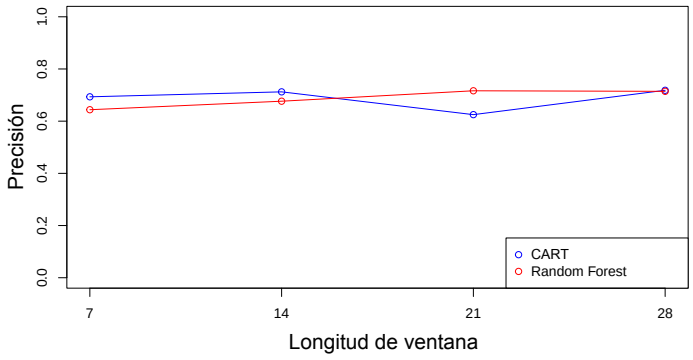


Figura 36: Valores de los clústeres a lo largo de la secuencia del usuario con identificador 5121.

Modelo	**CART**	**RandomForest**
Secuencia 7 días	69,33 %	64,38 %
Secuencia 14 días	71,23 %	67,65 %
Secuencia 21 días	62,50 %	71,64 %
Secuencia 28 días	71, 83 %	71, 43 %

Cuadro 4: Resultados con adición del mes y matriz de correlación para el usuario 5121.

Comparando los resultados de los tres experimentos, podemos observar que la inclusión del mes como característica mejora ligeramente la precisión de ambos modelos en la mayoría de las secuencias de tiempo. Esto sugiere que los patrones estacionales pueden influir en el consumo de los usuarios. Además, al aplicar la matriz de correlación, se produce una mejora significativa en la precisión de ambos modelos, lo que indica que las relaciones lineales entre las características pueden ser importantes para capturar patrones más complejos en los datos

4.1.9. PrefixSpan

Este modelo de predicción forma parte del paquete SPMF y se basa en la misma base de datos que se ha utilizado en los modelos CPT y DG, tal y como se mencionó anteriormente.
En este estudio, se empleó este modelo para descubrir patrones en las secuencias de consumo. Por ejemplo, para el usuario con identificador 5121, se aplicó a secuencias de 28 días de consumo utilizando el método de ventana deslizante para obtener patrones de consumo. Dado que la base de datos contiene la numeración de los clústeres de consumo, el modelo permite obtener secuencias de longitudes variables, incluyendo la frecuencia de aparición de cada patrón encontrado. En total, se identificaron 2,599 patrones de consumo. Los patrones descubiertos para el usuario 5121, con frecuencias próximas a la mitad de las secuencias analizadas, corresponden a comportamientos de consumo donde se mantiene una monotonía de consumo con el tipo de consumo A, el tipo F o alternando entre ellos.
Este modelo proporciona una amplia variedad de patrones, ya que puede obtener subsecuencias de longitud variable, desde longitud 1 hasta longitud 11 en este caso. Mientras que los patrones de longitud 1 ofrecen información sobre la frecuencia de cada clúster en cada secuencia, descubrir subsecuencias de patrones durante un número determinado de días resulta más interesante en este contexto.

4.2. Evaluación de los resultados

Esta sección se corresponde con la sexta fase de la metodología CRISP-DM.
Se ha comprobado que los modelos utilizados funcionan a alto nivel según su tipo y estructura. Para cada modelo, se han ajustado un conjunto de parámetros adaptados a la distribución de la base de datos y su relevancia para el análisis de patrones secuenciales de consumo energético. La comparación entre los distintos modelos se lleva a cabo considerando la metodología aplicada: predicción basada en secuencias prolongadas de consumo, predicción mediante el uso de ventana deslizante para secuencias de longitud específica, y análisis de patrones en secuencias de consumo energético. La aplicación de la matriz de correlación contribuye a mejorar la precisión de ciertos modelos según el tipo de consumo de cada usuario, al permitir la identificación de tipos de consumo altamente correlacionados. En la Figura 37 se presentan los resultados obtenidos de las predicciones entrenando los modelos con secuencias de 399 días y con el objetivo de predecir tipo de consumo del día subsecuente.

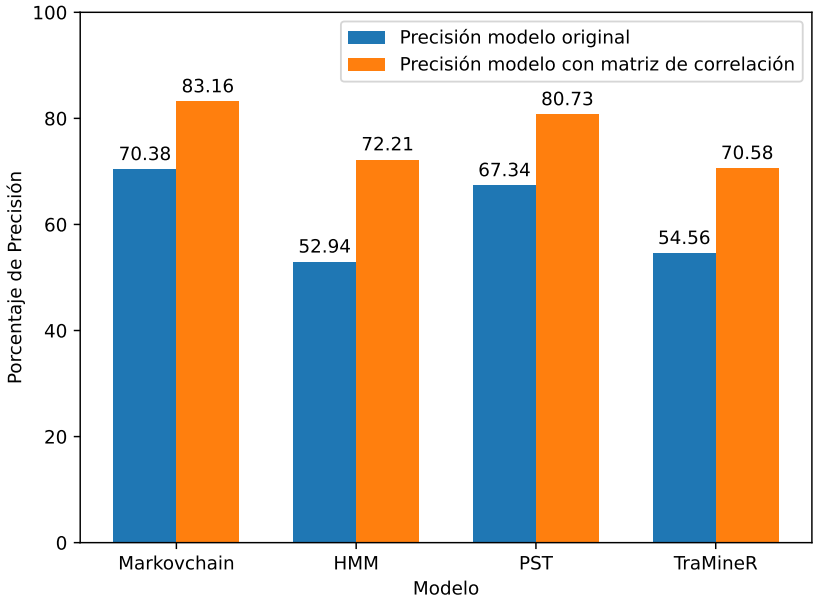


Figura 37: Precisión de los modelos para secuencias de 400 días.

Los modelos basados en cadenas de Markov, el modelo PST y la herramienta TraMineR realizan la predicción de consumo para el día siguiente a partir de una secuencia de 399 días para cada usuario. Sin embargo, debido a la complejidad espacial de la base de datos, estos modelos tienden a ser más lentos y algunos no alcanzan un rendimiento satisfactorio. Especi-

ficamente, los modelos de cadenas ocultas de Markov y PST son los más lentos, y el primero, en particular, no ha obtenido resultados de predicción notables. Los dos modelos con mayor precisión han sido aquellos basados en cadenas de Markov y árboles sufijos probabilísticos, aplicando la matriz de correlación, con precisiones del $83{,}36\,\%$ y $80{,}73\,\%$, respectivamente. Aunque la precisión no es el único criterio para designar a estos dos modelos como los mejores en este estudio, se han presentando las matrices de confusión para analizar detalladamente los errores de predicción de cada uno de ellos.

Evaluar los modelos que aplican el método de ventana deslizante puede puede requerir un esfuerzo computacional considerable, ya que su rendimiento está ligado al tipo de consumo del usuario seleccionado. En particular, el usuario con identificador 5121 utilizado en el estudio analítico de estos modelos, no presenta un patrón de consumo monótono como se puede ver en las Figuras 12, 13 y 14, lo que lo convierte en una elección adecuada para poner a prueba los modelos de predicción. En la Figura 38 se presentan los resultados obtenidos de las predicciones entrenando los modelos con ventanas de 14 días para el usuario 5121 y con el objetivo de predecir tipo de consumo del día siguiente en la secuencia.

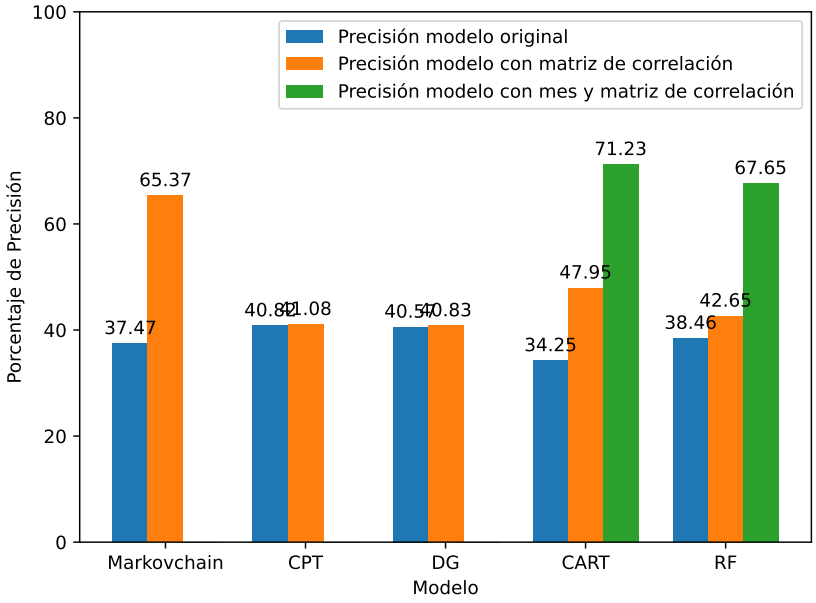


Figura 38: Precisión de los modelos para ventanas de 14 días para el usuario 5121.

En este contexto, algunos modelos muestran un rendimiento que puede ser mejorado, y de hecho, su precisión aumenta considerablemente con el uso de la matriz de correlación. Por ejemplo, se puede comparar el modelo basado en cadenas de Markov, que inicialmente tiene una precisión del $37{,}47\,\%$, y al aplicar la matriz de correlación, mejora la precisión a un $65{,}37\,\%$. Esto indica que el uso de la matriz de correlación aumenta la coherencia entre las clasificaciones del modelo y los valores reales.

Por otro lado, otros modelos como CPT y DG no han logrado mejorar significativamente su precisión al aplicar la matriz de correlación, lo que sugiere que sus errores de predicción fueron notables, ya que predicen tipos de consumo que no están altamente correlacionados con los clústeres reales.

Los modelos basados en árboles de predicción han conseguido mejorar sus resultados de predicción a medida que se incluye nueva información, como la inclusión del mes de cada secuencia o la matriz de correlación. Su rendimiento varía dependiendo de la secuencia de datos utilizada como datos de entrenamiento. En la Figura 36, se puede observar la evolución de estos modelos mejorados. En la Figura 39 se puede analizar la evolución de las precisiones de los distintos modelos entrenados con diversos tamaños de ventana para el usuario 5121 y con el objetivo de predecir tipo de consumo del día siguiente en la secuencia.

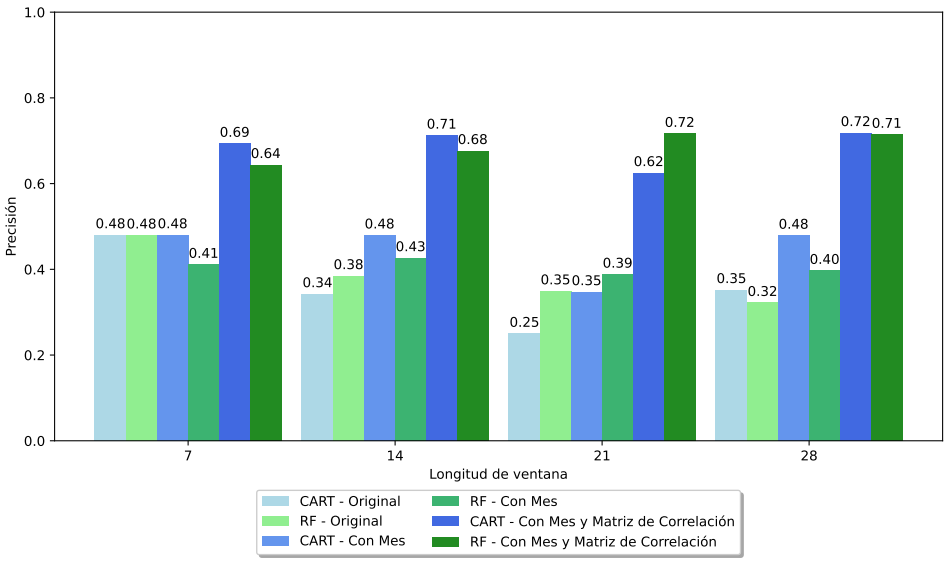


Figura 39: Precisión de los modelos para ventanas variables para el usuario 5121.

Por otro lado, el modelo PrefixSpan se presenta como una herramienta muy útil para encontrar patrones en las secuencias de consumo. Esta información resulta fundamental para el análisis de secuencias y permite llevar a cabo estudios mensuales de consumo para cada usuario, entre otros análisis de tipos de consumo.

5

Conclusiones y Líneas futuras

5.1. Conclusiones

La minería de datos ha demostrado ser una herramienta extremadamente útil para el análisis y la predicción de patrones de consumo energético en viviendas, además de proporcionar una base sólida para optimizar la distribución de recursos y desarrollar estrategias de gestión energética más eficientes. Los resultados obtenidos confirman la eficacia de estas técnicas de minería de datos para capturar tendencias en el consumo energético. Cada uno de los modelos predictivos utilizados ha mostrado sus ventajas y desventajas, dependiendo de la complejidad y naturaleza de los datos. Específicamente, los modelos basados en cadenas de Markov y árboles sufijos probabilísticos han obtenido precisiones del 83.36 % y 80.73 %, respectivamente, posicionándose como los más precisos en este estudio. Sin embargo, la precisión no ha sido el único criterio para evaluar el rendimiento de los modelos; la capacidad de cada modelo para adaptarse a la variabilidad de los datos y la eficiencia computacional también han sido factores determinantes.

Desde una perspectiva personal, este trabajo ha sido una oportunidad para profundizar en el campo de la minería de datos y su aplicación práctica en problemas reales. El proceso de investigar y experimentar con diferentes algoritmos permite desarrollar habilidades en la resolución de problemas y la toma de decisiones basada en datos. La experiencia obtenida a través de este proyecto no solo ha sido enriquecedora a nivel académico, sino también a nivel personal y profesional, pues aplicar conocimientos teóricos a situaciones prácticas refuerza el interés y confianza en el campo de la ciencia de datos.

5.2. Líneas futuras

Para futuras iteraciones de la metodología CRISP-DM empleada en este estudio, se podrían aplicar y evaluar los modelos basados en árboles de predicción para todos los usuarios de la base de datos con el objetivo de comprobar la precisión de estos modelos en cada caso. También parece razonable considerar la inclusión de factores ambientales que influyen en el tipo de consumo, además de la base de datos con tipos de consumo diarios. En este sentido, se ha optado por añadir el mes al que pertenece cada secuencia en la ventana deslizante, con el propósito de introducir un factor que pueda relacionar meses calurosos con meses fríos. Sin embargo, se pueden considerar otros factores como la temperatura, la humedad, las horas de sol o la precipitación, los cuales podrían mejorar significativamente la precisión de los modelos de predicción.

Otra alternativa sería la unificación de usuarios según sus patrones de consumo. Esto implica agrupar a los usuarios con patrones similares en sus secuencias de datos en un único grupo. En este sentido, el algoritmo PrefixSpan empleado en este estudio podría resultar útil. Por último, sería interesante explorar el paquete SPMF mencionado anteriormente, que ofrece una variedad de algoritmos para predicción y el análisis de patrones secuenciales, lo que podría complementar este estudio.

Referencias

[1] IBM,"¿Qué es la minería de datos? | IBM", IBM.com. https://www.ibm.com/es-es/topics/data-mining (consultado el 20 de Abril de 2024).

[2] Azure,"¿Qué es el aprendizaje automático? | Microsoft Azure", azure.microsoft.com. https://azure.microsoft.com/es-es/resources/cloud-computing-dictionary/what-is-machine-learning-platform (consultado el 24 de Abril de 2024)

[3] Rasch Measurement Transactions. "Data Mining and Rasch Measurement". Rasch Measurement Transactions. https://www.rasch.org/rmt/rmt152f.htm (consultado el 20 de abril de 2024).

[4] IBM, "Conceptos básicos de ayuda de CRISP-DM - Documentación de IBM", IBM.com. https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview (consultado el 20 de Abril de 2024).

[5] Payam Refaeilzadeh, Lei Tang, Huan Lui,"k-fold Cross-Validation", *Arizona State University*, 6 de noviembre de 2008

[6] J. R. Norris. *Markov Chains*. Cambridge: Cambridge University Press, 1997. doi: doi:10.1017/CBO9780511810633.

[7] "Modelos Ocultos de Markov HHM", https://web.archive.org/web/20110719170533/http://supervisadaextraccionrecuperacioninformacion.iespana.es/Markov.html, (consultado el 26 de Abril de 2024).

[8] djp3. (9 de abril de 2020). *HIDDEN MARKOV MODELS 12: THE BAUM-WELCH ALGORITHM*. Acceso: 01 de mayo de 2024. [Vídeo en línea]. Disponible en: https://youtu.be/JRsdt05pMoI?si=paQS-OudrHwSaQMB.

[9] L. Breiman, J. H. Friedman, R. A. Olsen y C. J. Stone. *Classification and Regression Trees*, CRC Press, 1984, doi: 0412048418.

[10] A. Gabadinho y G. Ritschard, "Analyzing State Sequences with Probabilistic Suffix Trees: The PST R Package", *Journal of Statistical Software*, vol. 72, pp. 1-39, American Statistical Association, Agosto 2016, doi: 10.18637/JSS.V072.I03.

[11] Christine Largeron-Leténo. "Prediction suffix trees for supervised classification of sequences." *Pattern Recognition Letters*, vol. 24, no. 16, pp. 3153-3164, Dec. 2003. doi: 10.1016/J.PATREC.2003.08.002.

[12] T. Gueniche, P. Fournier-Viger, and V. S. Tseng,"Compact Prediction Tree: A Lossless Model for Accurate Sequence Prediction", *Pattern Recognition Letters*, vol. 8347 LNAI, pp. 177-188, 2013. doi: 10.1007/978-3-642-53917-6_16.

[13] Nigel Franciscus, Xuguang Ren y Bela Stantic, "Dependency graph for short text extraction and summarization", *Journal of Information and Telecommunication*, vol. 3, no. 4, pp. 413-429, 2019. doi: 10.1080/24751839.2019.1598771.

[14] Hanane Amirat, Nasreddine Lagraa, Philippe Fournier-Viger y Youcef Ouinten, "Myroute: A graph-dependency based model for real-time route prediction", *Journal of Communications*, vol. 12, no. 12, pp. 668-676, diciembre 2017. DOI: 10.12720/JCM.12.12.668-676.

[15] Venkata N. Padmanabhan y Jeffrey C. Mogul, "Using predictive prefetching to improve world wide web latency", *Computer Communication Review*, vol. 26, no. 3, pp. 22-36, 1996. doi: 10.1145/235160.235164.

[16] Alexis Gabadinho, Gilbert Ritschard, Nicolas S. Müller y Matthias Studer, "Analyzing and Visualizing State Sequences in R with TraMineR", *Journal of Statistical Software*, vol. 40, no. 4, pp. 1-37, abril 2011. doi: 10.18637/JSS.V040.I04.

[17] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal y M. C. Hsu, "PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth", *Proceedings - International Conference on Data Engineering*, pp. 215-224, 2001. doi: 10.1109/ICDE.2001.914830.

[18] Commission for Energy Regulation (CER), CER Smart Metering Project - Electricity Customer Behaviour Trial, 2009-2010. 1st Edition. Irish Social Science Data Archive, 2012. URL: https://www.ucd.ie/issda/data/commissionforenergyregulationcer/.

[19] Francisco Rodríguez-Gómez, José del Campo-Ávila, Llanos Mora-López, "A novel clustering based method for characterizing household electricity consumption profiles", *Engineering Applications of Artificial Intelligence*, vol. 129, pág. 107653, marzo. 2024. doi: 10.1016/J.ENGAPPAI.2023.107653.

[20] Philippe Fournier-Viger, "SPMF: A Java Open-Source Data Mining Library", Philippe Fournier-Viger https://www.philippe-fournier-viger.com/spmf/ (consultado el 27 de Marzo de 2024).

Apéndice A
Manual de Instalación

Requerimientos:

Se proporcionan los documentos con el código donde se implementan los modelos de predicción. Debido a las restricciones del proveedor de los datos, no se incluyen los ficheros de datos en los que se basa este estudio. Sin embargo, pueden solicitarse en el siguiente enlace: https://www.ucd.ie/issda/data/commissionforenergyregulationcer/.

Es necesario disponer de una herramienta de visualización de cuadernillos capaz de ejecutar archivos .ipynb. También es necesario disponer de una versión de Python 3.9.12, una versión de R 4.3.2 y una versión de Java SE Development Kit (JDK) 1.8.0_411 o superiores. También se recomienda mantener las siguientes librerías con su correspondiente versión:

- R
  - PST: 0.94.1
  - TraMineR: 2.2-9
  - HMM: 1.0.1
  - igraph: 2.0.3
  - markovchain: 0.9.5
  - randomForest: 4.7-1.1
  - rpart: 4.1.21
  - caret: 6.0-94
- Python
  - numpy: 1.21.5
  - pandas: 1.4.2
  - seaborn: 0.11.2
  - matplotlib: 3.5.1

Apéndice B
Documentación

Entregables:

El código se ha organizado en ficheros ordenados según el orden de ejecución seguido durante el estudio y análisis de secuencias.

- **01-data_preparation**: cuaderno donde se implementa el preprocesado de datos y la generación de ficheros de datos para los modelos de predicción.
- **02-corMatrix.R**: script que permite trasponer la matriz de correlación calculada en el cuaderno para su posterior uso en los modelos de predicción.
- **03-markovchain.R**: implementación de un modelo basado en cadenas de Markov para secuencias de longitud 400.
- **04-hiddenMarkovchain.R**: implementación de un modelo de cadenas ocultas de Markov para secuencias de longitud 400.
- **05-hiddenMarkovchain_CorMatrix.R**: aplicación de la matriz de correlación en un modelo de cadenas ocultas de Markov.
- **06-markovchain_CorMatrix.R**: aplicación de la matriz de correlación en un modelo de cadenas de Markov.
- **07-markovchain_Evolution.R**: análisis de la evolución de las predicciones usando cadenas de Markov con diferentes longitudes de secuencia.
- **08-PST.R**: implementación de un modelo de árboles sufijos probabilísticos (PST).
- **09-PST_CorMatrix.R**: aplicación de la matriz de correlación en un modelo de árboles sufijos probabilísticos.
- **10-TraMineR.R**: uso del paquete TraMineR para la predicción utilizando secuencias de longitud 400.

- **11-TraMineR_CorMatrix.R**: aplicación de la matriz de correlación en el análisis de secuencias con TraMineR.
- **12-markovchain_Window.R**: implementación de un modelo de cadenas de Markov con ventanas deslizantes.
- **13-markovchain_Window_CorMatrix.R**: aplicación de la matriz de correlación en un modelo de cadenas de Markov utilizando ventanas deslizantes.
- **14-cart_rf.R**: implementación de modelos basados en árboles de decisión como CART y Random Forest utilizando ventanas deslizantes.
- **15-cart_rf_Months.R**: implementación de modelos de árboles de decisión como CART y Random Forest, añadiendo el mes de la secuencia en cada ventana del conjunto de datos.
- **16-cart_rf_Months_CorMatrix.R**: aplicación de la matriz de correlación en los modelos basados en árboles de decisión, añadiendo el mes de la secuencia en el conjunto de datos.
- **17-plotGraphics.R**: breve código para la generación de gráficos y visualizaciones de los datos.
- **MainTestCPT.java**: implementación de un modelo basado en árboles compactos predictivos (CPT).
- **MainTestCPTCorMatrix.java**: aplicación de la matriz de correlación en un modelo basado en árboles compactos predictivos (CPT).
- **MainTestDG.java**: implementación de un modelo basado en grafos degenerados (DG).
- **MainTestDGCorMatrix.java**: aplicación de la matriz de correlación en un modelo basado en grafos degenerados (DG).
- **MainTestPrefixSpan_saveToMemory.java**: implementación del algoritmo PrefixSpan para el análisis de patrones secuenciales.

Apéndice C
Figuras

Gráficas:

En esta sección se adjuntan gráficas de especial relevancia que no se han incluido en la memoria principal del estudio. Estas imágenes proporcionan información adicional y detallada que complementa los resultados y análisis presentados.

La Figura 40 muestra la base de datos sobre la que se realiza este estudio. Este histograma permite visualizar los usuarios con datos de consumo horarios completos para cada día en el conjunto original. Su utilidad radica en ofrecer una perspectiva clara sobre la distribución diaria de los datos de consumo energético.

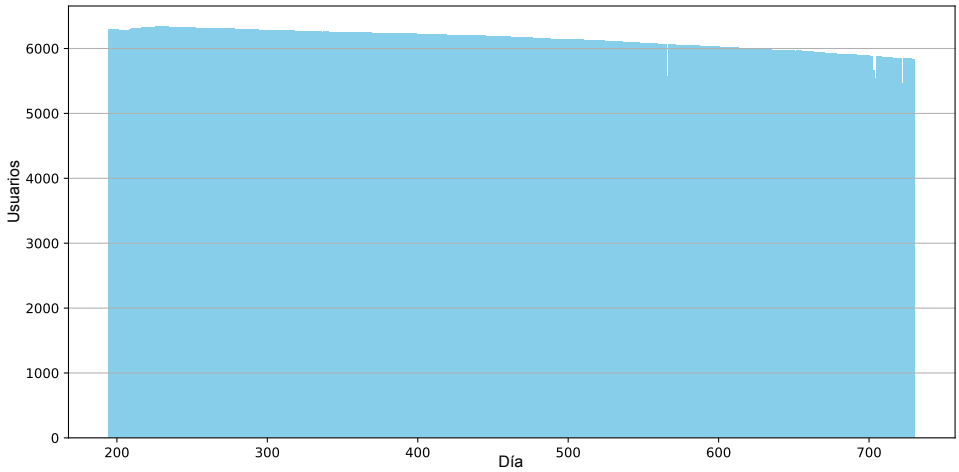


Figura 40: Histograma usuarios con datos de consumo horarios completos para un día en el conjunto original.

En la Figura 41 se muestra otro histograma, esta vez enfocado en los usuarios resultantes tras el filtrado que se ha detallado en la memoria. Comparando ambas figuras, se pueden obtener conclusiones sobre los usuarios que no cumplen los criterios establecidos, observándose que se han eliminado aproximadamente unos 1000 usuarios tras el filtrado.

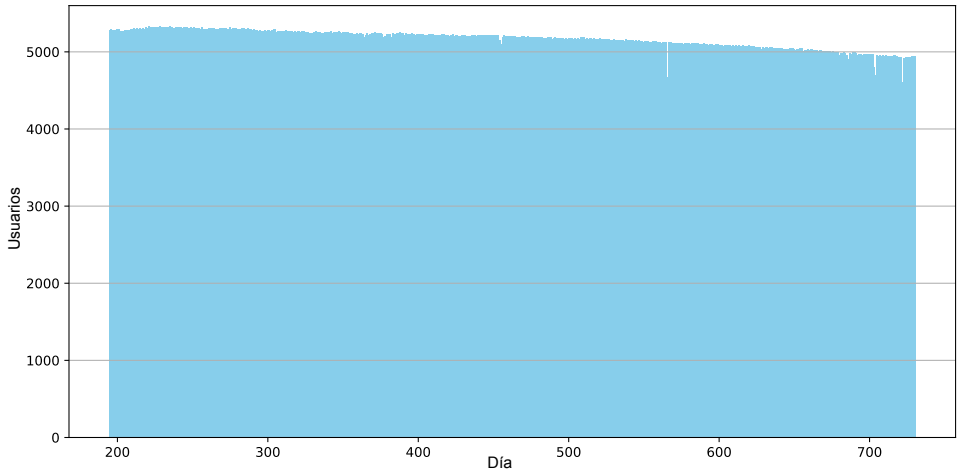


Figura 41: Histograma de usuarios con datos de consumo diario completos para un día en el conjunto original.

La Figura 42 presenta el índice de entropía para cada día de la secuencia completa. Este gráfico es fundamental para analizar la variabilidad y la incertidumbre en los datos de consumo a lo largo del tiempo. Un alto índice de entropía indica mayor desorden o diversidad en el consumo, mientras que un índice bajo sugiere patrones más predecibles y consistentes. Esta información es vital para comprender la dinámica temporal del consumo y para desarrollar modelos predictivos más precisos.

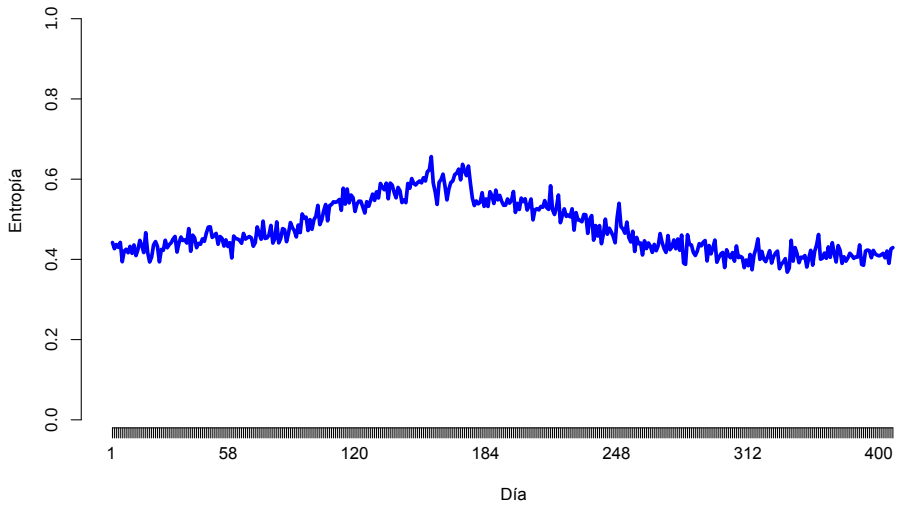


Figura 42: Índice de entropía para cada día de la secuencia completa.

La Figura 43 muestra la matriz de confusión del modelo TraMineR aplicada a toda la base de datos. Esta gráfica es esencial para evaluar el rendimiento del modelo, permitiendo identificar la precisión y los errores de clasificación.

	Reference																	
Prediction	A	C	D	F	G	H	I	J	K	L	M	N	O	P	Q	R	T	S
A	43	0	8	46	1	0	0	0	0	10	0	0	0	0	0	0	0	0
C	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
D	7	0	2	10	0	0	0	1	2	2	0	0	0	0	0	0	0	0
F	28	0	3	228	3	0	0	0	0	20	0	0	0	1	1	0	0	0
G	0	0	0	2	1	0	0	1	0	1	1	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	4	0	0	0	0	0	1	0	0	0	0	0
J	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K	2	0	0	4	0	0	0	0	0	3	1	0	0	0	1	0	0	0
L	8	0	2	13	1	0	0	0	0	11	0	0	0	0	0	0	1	0
M	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	1	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
O	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0	3	0	0	0	0
Q	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0
S	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

Figura 43: Matriz de confusión del modelo TraMineR para toda la base de datos.

En la Figura 44 se presenta la matriz de confusión del modelo TraMineR utilizando la matriz de correlación para toda la base de datos. Se pueden analizar las diferencias en la medición del error al clasificar como correctos los clústeres altamente correlacionados. Permite concluir que el modelo TraMineR predice frecuentemente clústeres que presentan curvas de consumo similares con el clúster original.

	Reference																
Prediction	A	C	D	F	G	H	I	J	K	L	M	N	O	P	Q	R	T
A	67	0	5	0	1	0	0	0	0	8	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	9	0	0	10	0	0	0	0	0	3	0	0	0	0	0	0	0
F	0	0	8	263	3	0	0	0	1	24	0	0	0	0	1	0	0
G	1	0	0	2	1	0	0	1	0	2	0	0	0	0	0	0	0
H	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
I	0	1	0	0	0	0	4	0	0	0	0	0	1	0	0	0	0
J	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0
K	1	0	0	2	0	0	0	0	0	0	0	0	0	0	1	0	0
L	9	0	0	23	0	0	0	1	1	8	0	0	0	0	0	0	1
M	0	0	1	2	0	0	0	0	0	1	2	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
O	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0
P	0	0	0	0	0	0	0	0	0	1	0	0	0	4	0	1	0
Q	0	0	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0
R	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
T	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1

Figura 44: Matriz de confusión del modelo TraMineR para toda la base de datos utilizando la matriz de correlación.

La Figura 45 muestra las estadísticas del modelo PST (*Probabilistic Suffix Tree*) para secuencias de 400 días. Este gráfico proporciona información sobre la distribución y sensibilidad de los patrones de consumo detectados por el modelo en un largo período. Permite comprender la evolución temporal de las secuencias y la efectividad del modelo en capturar patrones a largo plazo.

Statistics by Class:

	Class: A	Class: C	Class: D	Class: F	Class: G	Class: H	Class: I	Class: J	Class: K	Class: L	Class: M	Class: N	Class: O	Class: P	Class: Q	Class: R	Class: T
Sensitivity	0.50000	1.000000	0.066667	0.8366	0.166667	1.000000	0.750000	0.000000	0.000000	0.30612	0.250000	0.000000	1.000000	0.750000	0.000000	0.000000	1.000000
Specificity	0.88148	1.000000	0.956067	0.6738	1.000000	1.000000	0.997955	0.997963	0.993878	0.96396	0.993865	0.997967	1.000000	0.993865	0.997963	0.997967	0.997963
Pos Pred Value	0.47826	1.000000	0.045455	0.8076	1.000000	1.000000	0.750000	0.000000	0.000000	0.48387	0.250000	0.000000	1.000000	0.500000	0.000000	0.000000	0.666667
Neg Pred Value	0.89027	1.000000	0.970276	0.7159	0.989837	1.000000	0.997955	0.995935	0.993878	0.92641	0.993865	0.997967	1.000000	0.997947	0.995935	0.997967	1.000000
Prevalence	0.17850	0.002028	0.030426	0.6207	0.012170	0.002028	0.008114	0.004057	0.006085	0.09939	0.008114	0.002028	0.008114	0.008114	0.004057	0.002028	0.004057
Detection Rate	0.08925	0.002028	0.002028	0.5193	0.002028	0.002028	0.006085	0.000000	0.000000	0.03043	0.002028	0.000000	0.008114	0.006085	0.000000	0.000000	0.004057
Detection Prevalence	0.18661	0.002028	0.044625	0.6430	0.002028	0.002028	0.008114	0.002028	0.006085	0.06288	0.008114	0.002028	0.008114	0.012170	0.002028	0.002028	0.006085
Balanced Accuracy	0.69074	1.000000	0.511367	0.7552	0.583333	1.000000	0.873978	0.498982	0.496939	0.63504	0.621933	0.498984	1.000000	0.871933	0.498982	0.498984	0.998982

Figura 45: Estadísticas modelo PST para secuencias de 400 días.

La Figura 46 presenta las estadísticas del modelo PST utilizando el método de matriz de correlación para secuencias de 400 días. Comparando esta gráfica con la Figura 45, se puede evaluar el impacto de la matriz de correlación como parámetro de error en las estadísticas generadas por el modelo. Esto es útil para determinar si la incorporación de la correlación mejora la capacidad del modelo para identificar patrones y tendencias en los datos de consumo.

Statistics by Class:

	Class: A	Class: C	Class: D	Class: F	Class: G	Class: H	Class: I	Class: J	Class: K	Class: L	Class: M	Class: N	Class: O	Class: P	Class: Q	Class: R	Class: T
Sensitivity	0.8636	1.000000	0.066667	0.9412	0.166667	1.000000	0.750000	0.000000	0.000000	0.32653	0.250000	1.000000	1.000000	0.750000	0.000000	0.000000	1.000000
Specificity	0.9605	1.000000	0.956067	0.8449	1.000000	1.000000	0.997955	0.997963	0.993878	0.96396	0.993865	0.997967	1.000000	0.995910	0.997963	1.000000	0.997963
Pos Pred Value	0.8261	1.000000	0.045455	0.9085	1.000000	1.000000	0.750000	0.000000	0.000000	0.50000	0.250000	0.500000	1.000000	0.600000	0.000000	NaN	0.666667
Neg Pred Value	0.9701	1.000000	0.970276	0.8977	0.989837	1.000000	0.997955	0.995935	0.993878	0.92842	0.993865	1.000000	1.000000	0.997951	0.995935	0.997972	1.000000
Prevalence	0.1785	0.002028	0.030426	0.6207	0.012170	0.002028	0.008114	0.004057	0.006085	0.09939	0.008114	0.002028	0.008114	0.008114	0.004057	0.002028	0.004057
Detection Rate	0.1542	0.002028	0.002028	0.5842	0.002028	0.002028	0.006085	0.000000	0.000000	0.03245	0.002028	0.002028	0.008114	0.006085	0.000000	0.000000	0.004057
Detection Prevalence	0.1866	0.002028	0.044625	0.6430	0.002028	0.002028	0.008114	0.002028	0.006085	0.06491	0.008114	0.004057	0.008114	0.010142	0.002028	0.000000	0.006085
Balanced Accuracy	0.9121	1.000000	0.511367	0.8930	0.583333	1.000000	0.873978	0.498982	0.496939	0.64525	0.621933	0.998984	1.000000	0.872955	0.498982	0.500000	0.998982

Figura 46: Estadísticas modelo PST con método matriz de correlación para secuencias de 400 días.

UNIVERSIDAD DE MÁLAGA | uma.es

E.T.S de Ingeniería Informática
Bulevar Louis Pasteur, 35
Campus de Teatinos
29071 Málaga

E.T.S. DE INGENIERÍA INFORMÁTICA