Gobierno del Principado de Asturias
Consejería de Economía y Empleo

Unión Europea
Fondo Europeo
de Desarrollo Regional

RESULTADOS ENERCONFORT

Minimización del consumo energético de edificios optimizando el confort de sus usuarios

En Gijón , a 31 de Diciembre de 2017

1. MEMORIA TÉCNICA

El proyecto ENERCONFORT (Minimización del consumo energético de edificios optimizando el confort de sus usuarios) es ejecutado por El Centro Tecnológico CTIC entre el 1 de enero de 2016 y el 31 de diciembre de 2017.

El proyecto ENERCONFORT propone adoptar una solución en la cual se reduzcan los consumos energéticos de edificios asegurando o mejorando el confort térmico, acústico, lumínico y ambiental de los usuarios. Por tanto, el objetivo de este proyecto es desarrollar un sistema experto que permita alcanzar soluciones de compromiso que **minimicen el consumo energético y maximicen el confort** final de estos usuarios.

El proyecto se ha validado mediante un caso de uso concreto en un local de oficinas, coincidente con las instalaciones que CTIC dispone para desarrollar su actividad.

La ejecución del proyecto ENERCONFORT se dividió en cuatro paquetes de trabajo (Hitos) tal y como se muestra en el siguiente esquema:

Figura 1: Paquetes de trabajo

ANUALIDAD 2016

PT1: Problemática del consumo energético y del confort.

Tarea 1.1: Recopilación de legislación de eficiencia energética y confort

El objetivo de este hito era conocer la legislación correspondiente al consumo energético y de confort en el ámbito nacional y europeo, enfocado principalmente a la relacionada con edificios residenciales y de oficinas, para poder conocer la problemática que existe y analizar cuáles son los criterios en eficiencia energética y confort térmico, lumínico, acústico y de calidad del aire que se deben cumplir en los edificios de oficinas.

Eficiencia energética

En cuanto a la legislación en eficiencia energética para edificios en el ámbito europeo se recopilaron y analizaron las siguientes normativas de actuación:

- Directiva-2010-31-UE.
- Directiva 2012/27/UE.

Se recopilaron además el conjunto de normas relacionadas con estas directivas tal y como se recoge en el siguiente esquema:

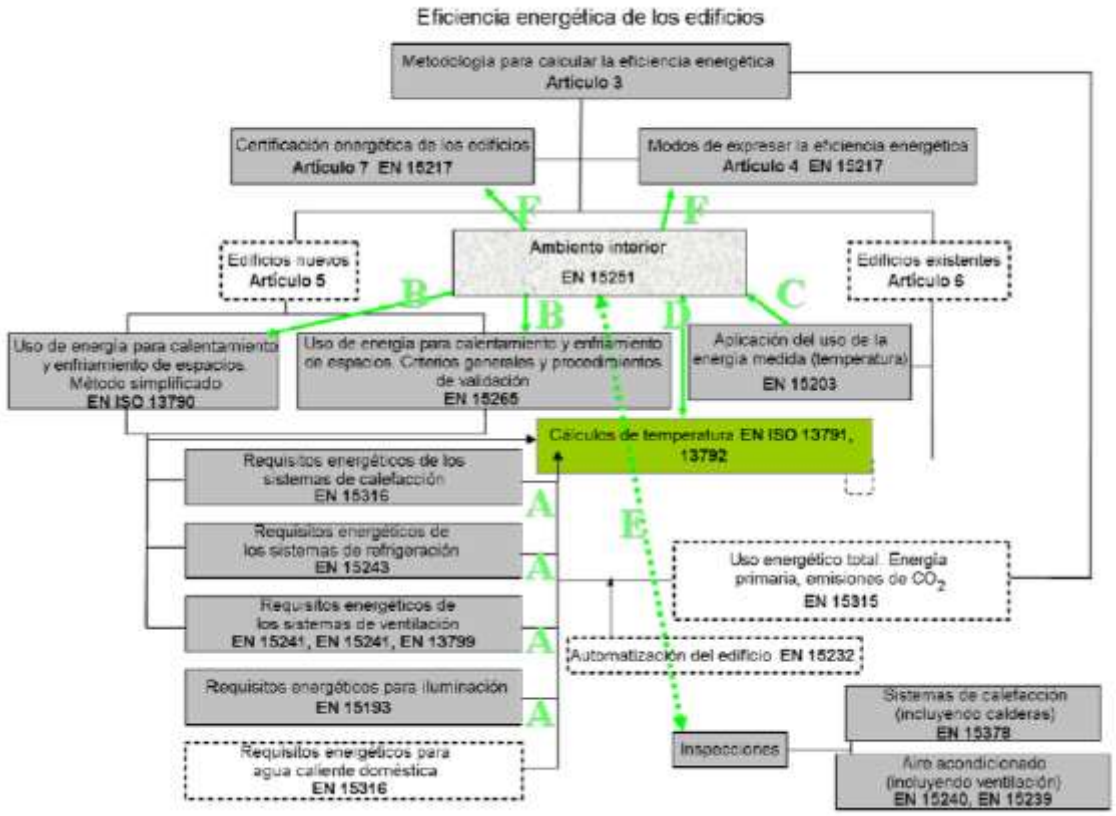


Figura 2: Normativa relacionada con la eficiencia energética en los edificios.
Fuente: Real Decreto RD 486/1997. Disposiciones mínimas de seguridad y salud en los lugares de trabajo

En lo que se refiere al ámbito nacional se analizó la siguiente legislación:

- Real Decreto 47/2007.
- Plan de acción de Ahorro y Eficiencia Energética 2011-2020-IDEA.
- Real Decreto 253/2013.
- Proyecto del Real Decreto relativo a la eficiencia energética.

En concreto, estos Reales Decretos están integrados en los siguientes instrumentos normativos:

- Código Técnico de la Edificación (CTE).
- Reglamento de Instalaciones Térmicas (RITE).
- Certificación de Eficiencia Energética de Edificios.

Confort

En lo relacionado con la legislación y recomendaciones correspondiente al confort se han recogido y analizado las siguientes normativas relacionadas con el confort térmico, lumínico, acústico y de calidad del aire:

Confort térmico:

- Artículo 7 y Anexo III del Real Decreto 486/1997.
- Ley 31/1995, de 8 de noviembre, de Prevención de Riesgos laborales. BOE nº269, de 10 de noviembre.
- Real Decreto 486/1997, de 14 de abril. BOE nº 97, de 23 de abril sobre lugares de trabajo.
- Guía Técnica para la evaluación y prevención de los riesgos relativos a la utilización de los lugares de trabajo. (Real Decreto 486/1997). INSHT.
- UNE EN 27243:95. Ambientes calurosos: Estimación del estrés térmico del hombre en el trabajo basado en el índice WBGT.
- UNE EN ISO 7726:02. Ergonomía de los ambientes térmicos: instrumentos de medida de las magnitudes físicas.
- UNE EN ISO 7933:05. Ergonomía del ambiente térmico. Determinación analítica e interpretación del estrés térmico mediante el cálculo de la sobrecarga estimada.
- UNE EN ISO 8996:05. Ergonomía del ambiente térmico: determinación de la tasa metabólica.
- UNE EN ISO 15265:05. Ergonomía del ambiente térmico. Estrategia de evaluación del riesgo para la prevención del estrés o incomodidad en condiciones de trabajo térmicas.
- UNE EN ISO 7730:06. Ergonomía del ambiente térmico. Determinación analítica e interpretación del bienestar térmico mediante el cálculo de los índices PMV y PPD y los criterios de bienestar térmico local.

Confort lumínico:

- UNE EN 15193. Eficiencia energética en los edificios. Requisitos energéticos de iluminación. UNE 12464.1. Iluminación de los lugares de trabajo en interior.

- RD 486/1997. Artículo VIII. Disposiciones mínimas de seguridad y salud en los lugares de trabajo. Iluminación.
- CTE HE3. Eficiencia energética de las instalaciones de iluminación.
- CTE SU-4. 12.4. Seguridad frente al riesgo causado por iluminación inadecuada.

Confort acústico:

- UNE-ISO 1996-1. magnitudes básicas que se deben utilizar para la descripción del ruido en ambientes comunitarios.
- UNE 74-022. Método para la medida del ruido en comunidades.
- UNE 74024-91. guía para la medida del ruido y su evaluación de sus efectos sobre el hombre.
- DB-HR Protección frente al ruido. Documento Básico para reglas y procedimientos que permiten cumplir las exigencias básicas de protección frente al ruido.

Calidad del aire:

- Norma UNE 171330-1:2008. Calidad Ambiental en Interiores. Parte I: Diagnóstico de calidad ambiental interior.
- Norma UNE 171330-2:2009. Calidad Ambiental en Interiores. Parte II: Procedimientos de inspección de calidad ambiental interior.
- Norma UNE-EN 13779:2005. Ventilación de edificios no residenciales.

A modo de resumen y necesario a objeto del desarrollo de este proyecto, la siguiente tabla recopila las condiciones ambientales de los lugares de trabajo:

Tabla 1: Condiciones ambientales en locales de trabajo

<table>
<tr><th colspan="4">LOCALES DE TRABAJO CERRADOS</th></tr>
<tr><td rowspan="2">Temperatura</td><td>Trabajos sedentarios</td><td>Trabajos ligeros</td><td>Locales riesgos eléctricos</td></tr>
<tr><td>Invierno 17< 27º C
Verano 23 < 27º C</td><td>Entre 14 y 25º C</td><td rowspan="2">< 50%</td></tr>
<tr><td>Humedad</td><td colspan="2">Entre el 30 y el 70 %</td></tr>
<tr><td rowspan="2">Velocidad del aire</td><td>Trabajos en ambientes no calurosos</td><td>Trabajos sedentarios en ambientes calurosos</td><td>Trabajos no sedentarios en ambientes calurosos</td></tr>
<tr><td>0,25 m/s</td><td>0,5 m/s</td><td>0,75 m/s</td></tr>
<tr><td rowspan="3">Excepción</td><td colspan="3">• Corrientes de aire para evitar estrés térmico
• Corrientes de aire acondicionado</td></tr>
<tr><td colspan="2">Trabajos sedentarios</td><td>Demás casos</td></tr>
<tr><td colspan="2">0,25 m/s</td><td>0,35 m/s</td></tr>
<tr><td rowspan="2">Renovación mínima del aire</td><td colspan="2">Trabajos sedentarios ambientes no calurosos, no contaminados</td><td>Casos restantes</td></tr>
<tr><td colspan="2">30m³/h/trabajador</td><td>50m³/h/trabajador</td></tr>
</table>

T1.2 – Estado del arte de las metodologías

En esta tarea se ha estudiado el estado del arte, así como las técnicas a tener en cuenta relativas a la eficiencia energética y el confort desde el punto de vista térmico, lumínico, acústico y de calidad de aire para edificios y más concretamente para locales de oficinas, con la perspectiva del IoT y el diseño emocional.

Tecnologías de monitorización en tiempo real enfocadas a la eficiencia energética.

En lo relacionado a las tecnologías de monitorización en tiempo real, se han estudiado aquellas tecnologías enfocadas a la eficiencia energética. En concreto se han estudiado aquellas dirigidas a la monitorización en edificios. Por un lado, se han analizado cuales son las tipologías de consumo, así como la distribución en cuanto al sector, actividad y uso de dichos consumos que se pueden medir en los edificios, diferenciando consumos térmicos y consumos eléctricos. Por otro lado, se han analizado cuales son los parámetros fundamentales para la evaluación de dichos consumos. De ello se concluye lo siguiente:

- Consumo eléctrico. Sus principales usos en la edificación son la iluminación, los consumos de los equipos eléctricos (electrodomésticos, ordenadores, televisiones, etc.), la climatización y la generación térmica para calefacción, y agua caliente sanitaria (equipos de calentamiento mediante Efecto Joule o bombas de calor).
- Consumo térmico. Fundamentalmente está destinado a la climatización de las estancias y a la generación de ACS. Su origen puede ser químico, a través de la reacción de oxidación de un combustible (gas, gasóleo, madera, etc.) en una caldera, provenir de un ciclo termodinámico (bomba de calor), de la captación de la energía solar térmica o del Efecto Joule, por el que un conductor (resistencia) se calienta debido al paso de la energía eléctrica a través de él.

Una vez analizada la tipología y la distribución del consumo energético en edificación, se estudiaron cuáles son las diferentes metodologías para la medida de los flujos energéticos, así como para analizar y transmitir los datos de dichos flujos.

Del estudio se han sacado las siguientes conclusiones:

- Los principales elementos para monitorizar la corriente eléctrica empleados en la medida del consumo eléctrico en instalaciones doméstica son los sensores no invasivos toroidales o de pinza amperimétrica.
- Los principales elementos para monitorizar el consumo térmico se realizan a través de la monitorización bien del gasto primario de energía (por ejemplo, combustible) o bien a través de la medida del caudal del medio de transmisión (agua o gas) y las temperaturas de impulsión y retorno, conociendo la capacidad calorífica del fluido portador. Esta última opción suele ser la

más habitual, utilizando para ello dispositivos llamados contadores de calorías, provistos de la sensórica necesaria para realizar el cálculo correspondiente y transmitiendo electrónicamente esa información a la plataforma de tratamiento de datos correspondiente.

- Los principales elementos para centralizar, operar y transmitir los datos son las llamadas placas electrónicas que permiten desarrollar soluciones "Ad-Hoc" como las Raspberry Pi, Arduino ó BeagleBone.

Tecnologías de monitorización en tiempo real enfocadas al confort.

Además de evaluar las tecnologías óptimas para la monitorización de consumos energéticos, se analizaron aquellas relacionadas con los diferentes parámetros de confort, que, en algunos casos, como el de las variables de confort térmico, estaban directamente relacionadas con la propia eficiencia energética. Se estudiaron los diferentes parámetros que se evalúan mediante los dispositivos de monitorización en el interior de los edificios. Éstos son:

- Temperatura. Los sensores más comunes para medir este tipo de parámetros son las termo resistencias y los termopares.
- Humedad ambiental. Los principales sensores de humedad ambiental son de tipo capacitivo. Los modelos utilizados habitualmente son los DHT11 y DJT22 en diferentes formas de encapsulados.
- Ruido. Existen múltiples tecnologías en cuanto a sensores se refiere (micrófonos piezoeléctricos, de condensador Electret, etc.). Los requisitos específicos de la aplicación (umbral de sensibilidad, direccionalidad, ganancia, etc.) definirán la elección de una u otra tecnología o modelo.
- Iluminación. Los principales elementos empleados como sensor en este campo son las fotorresistencias, los fotodiodos y los fototransistores.
- CO_2. Existen fundamentalmente dos tipos de sensores para este tipo de aplicaciones: de infrarrojos y de célula electroquímica.
- Compuestos orgánicos volátiles (COVs). Para la medida de este parámetro se puede optar fundamentalmente por dos opciones, una más económica basada en semiconductores de óxidos metálicos, u otra, más costosa y precisa, como es el PIV (Photo-Ionization Detector), que se fundamenta en la ionización de los gases mediante una fuente de luz ultravioleta.
- Velocidad del aire. Dadas las bajas velocidades que se necesita medir en el campo de la climatización se requieren sondas de hilo caliente para su monitorización.

Diseño emocional.

Una de las metodologías que se está aplicando en este proyecto es la ingeniería Kansei. Dicha metodología forma parte de una forma de integrar el diseño emocional, de forma que el diseño de un producto dependa no sólo de las variables objetivas propias del diseño convencional del producto sino también de variables subjetivas o percepciones que conforman la opinión de los usuarios de un producto.

Hay diferentes metodologías relacionadas con este diseño afectivo o emocional, que permiten orientar de forma fiable el diseño de un producto o de un servicio de acuerdo a las percepciones, nivel de satisfacción y necesidades del consumidor garantizando el éxito del producto y mejorando la calidad de vida y el confort de las personas. Para este proyecto se realizó una recopilación de las diferentes técnicas de diseño emocional, que se pueden resumir en las siguientes:

- Método diferencial semántico.
- Descripción semántica de ambientes.
- Análisis conjunto.
- Despliegue de la función de calidad.
- Modelo Kano.
- Herramienta PrEmo.

La ingeniería Kansei, contempla alguna de estas metodologías, y se fundamenta en facilitar la traducción de las expectativas emocionales de los clientes en especificaciones técnicas de diseño. Para el desarrollo de este proyecto, se estudió en profundidad el fundamento de esta metodología, así como los principales productos en los que la aplicación de la Ingeniería Kansei ha supuesto un éxito en ventas.

Además, se recopilaron proyectos y/o productos en los que la Ingeniería Kansei fue aplicada al confort, destacando los desarrollados en la Universidad de Alicante, que basan su estudio en la evaluación de la percepción del confort para el diseño de bibliotecas.

T1.3. Análisis de requisitos

Esta tarea fue necesaria para poder definir los requisitos que deberá cumplir el sistema para el correcto desarrollo del proyecto. Para ello, primero se definió el escenario y caso de uso y se identificaron aquellos requisitos, teniendo en cuenta la normativa específica para la eficiencia energética y confort y el estado del arte de la metodología, que permitió determinar las mejores técnicas, parámetros y dispositivos que se están utilizando en este proyecto.

El caso de uso seleccionado fue el Centro Tecnológico CTIC que coincide además con el lugar donde se está desarrollando el proyecto. El estudio se concentró en la evaluación del confort de los trabajadores de la planta -1 y +1 que han sido también sensorizadas para la monitorización del consumo energético y de las condiciones ambientales.

Una vez conocido el caso de uso y analizada la legislación y el estado de arte de las metodologías que ya se están implementando en el proyecto, se definieron los requisitos del sistema para la minimización del consumo energético en edificios optimizando el confort de sus usuarios. Se definieron dos tipos de requisitos: funcionales y no funcionales.

Entre los requisitos funcionales se han tenido en cuenta los relacionados con:

- La legislación en eficiencia energética y confort: valores mínimos y máximos.
- Encuestas: espacio semántico de percepciones, la capacidad del sistema para ser enviadas en el momento óptimo de forma automática, el desarrollo de los algoritmos apropiados de optimización.
- Plataforma: capacidad para la recopilación de datos de los sensores, del envío de los datos a la base de datos, capacidad de procesar dichos datos mediante algoritmos, capacidad de integración de los datos con la plataforma y capacidad para generar alarmas.
- Sistema de visualización: capacidad de consulta de la información y capacidad para enfocar el sistema para dos perfiles diferentes.

Los requisitos no funcionales se resumen en los siguientes:

- Los relacionados con el diseño de la plataforma: capacidad de incorporación de nuevas fuentes de información, escalable vertical y horizontalmente, capacidad para almacenar grandes volúmenes de datos, altos rendimientos de escritura y de lectura de datos, inclusión de documentación, etc.
- Los relacionados con la interoperabilidad y la seguridad de las comunicaciones: mecanismos de control de acceso, capacidad para realizar backups de información, etc.

PT2: Elaboración y procesado de encuestas de confort

T2.1. Elaboración de las encuestas de confort

Para la elaboración de las encuestas de confort fue necesario realizar un estado del arte en cuanto a las metodologías relacionadas con el diseño emocional (estudiadas en el PT1) así como de los diferentes tipos de Ingeniería Kansei existentes y métodos estadísticos para su evaluación.

En cuanto a los tipos de Ingenieria Kansei, se encontraron seis tipos diferentes:

- Tipo 1: Clasificación por categorías. En este método, las categorías del "Kansei" asociadas a un producto son descompuestas en más subconceptos para encontrar los rasgos físicos de un dominio de nuevo diseño. Por lo general, los rasgos físicos son transferidos a especificaciones más detalladas después de los experimentos llevados a cabo en la cuestión ergonómica.
- Tipo 2: También denominado sistema KES. Este método implica un sistema informático de apoyo para la elección de un bien de consumo y para el diseño y elaboración de un nuevo producto. El modelo de red neuronal, la teoría de conjuntos difusos, GA (Genetic Algorithm) se utilizan como un sistema inteligente con un motor de inferencia. Si introducimos el Kansei del consumidor, el sistema reconoce el Kansei y propone un candidato de nuevos diseños a través del sistema de inteligencia artificial. **Esta tipología fue la seleccionada para el desarrollo de este proyecto**.

- Tipo 3: El KES híbrido. Si el sistema de ingeniería Kansei va hacia delante, se parte de las palabras introducidas por el usuario para mostrar los elementos de diseño que satisfacen dichas sensaciones, y a continuación una vez visto el candidato que el sistema ha propuesto, se crea su nueva idea. Si el sistema de ingeniería Kansei va hacia atrás, se parte del boceto del diseñador para mostrar las palabras Kansei que el usuario asocia a dicho boceto, según lo define.
- Tipo 4: Modelo matemático. Este tipo se centra en modelos matemáticos predictivos que se utilizan en lugar de una base de reglas para obtener la salida óptima a partir de las palabras de entrada. Estos modelos son también validados como ocurre en los tipos II y III.
- Tipo 5: Ingeniería Kansei Virtual. Esta metodología integra técnicas de realidad virtual con sistemas de recolección de datos estándar. De esta manera, el cliente va a poder percibir de forma virtual, pero muy próxima a la realidad, como quedaría el producto en cuestión una vez acabado y listo para su utilización.
- Tipo 6: Sistema de diseño colaborativo Kansei. En este tipo, la base de datos Kansei es accesible vía internet. El diseño se apoya en el trabajo en grupo y la ingeniería concurrente, utilizando herramientas tipo QFD que se apoyan en las preferencias del consumidor.

Para elaborar las encuestas y poder tener en cuenta los parámetros de confort de una forma más precisa, se hizo necesario establecer una metodología de evaluación de las percepciones que de forma individual definen el confort de una persona. Las encuestas fueron elaboradas a partir del siguiente enfoque metodológico definido en el marco de la Ingeniería Kansei:

1) Se definió un dominio.

   El dominio incluye la definición del tipo de mercado y del público objetivo, el nicho de mercado, así como las especificaciones del nuevo producto. Para este caso, se podría decir que la definición del dominio vendrá a ser un estado del arte de las diferentes metodologías existentes para la mejora del confort. De esta forma:

   - Producto/Servicio: Mejorar el confort de los trabajadores de oficinas.
   - Problemática: El confort es un parámetro subjetivo y muy difícil de evaluar, por lo que las soluciones presentes no evalúan el confort para cada usuario desde un punto de vista subjetivo, ni tampoco el confort general.

     Una distribución inadecuada de luz puede provocar dolores de cabeza, incomodidad visual, errores, fatiga visual, confusiones, y sobre todo la pérdida de visión. Un mal confort acústico puede generar interferencias en la comunicación, así como una disminución de la concentración, del rendimiento, trastornos del sueño, etc. Un mal confort térmico puede ocasionar pérdida de concentración, o favorecer a la aparición de resfriados, gripes, etc.
   - Mercado: para este contexto, el mercado se atribuye en sí al público objetivo que serán los trabajadores de la oficina y los propietarios de la empresa.

2) Se generó el espacio semántico.

Esto implica la selección de atributos Kansei, es decir, la identificación de aquellas percepciones que se van a tratar de cuantificar. En este caso dichas percepciones se traducen en forma de adjetivos que permitirán evaluar los sentimientos de los usuarios con respecto al confort que sienten en su puesto de trabajo.

Para que el universo semántico sea lo más completo posible, es necesario tratar de conocer todas aquellas percepciones que, a priori, se considera que son representativas del dominio y que puedan tener relación directa con los elementos de diseño, que son, en este caso, el confort térmico, lumínico, acústico y de calidad del aire. Para ello, se hará uso de los diagramas de afinidad de forma que se organicen ideas de una forma que permitan ser discutidas, mejoradas e interaccionadas con todos los participantes.

La siguiente figura muestra una imagen del procedimiento realizado para la obtención del diagrama de afinidad:

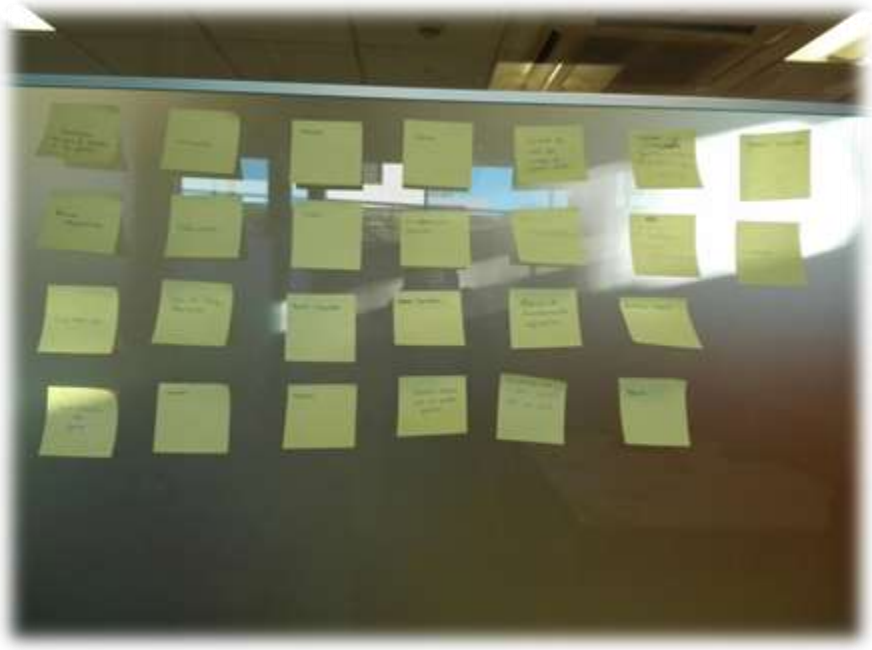


Figura 4: Percepciones obtenidas del diagrama de afinidad

En total se obtuvieron 28 percepciones que posteriormente se agruparon en 10 grupos de ejes semánticos (Figura 5).

Figura 5: Ejes semánticos

La tabla 2 muestra el listado con las diferentes percepciones agrupadas por grupos de ejes semánticos en las cuales aquellas las sombreadas en color azul son aquellas de nivel jerárquico superior que englobaría a las de nivel jerárquico inferior sin colorear.

Tabla 2: Percepciones

PERCEPCIONES
Tranquilo
Silencioso
Íntimo
Poco tránsito de gente
Bien iluminado
Buena iluminación natural
Buena iluminación artificial
Con buen tono de color de luz
Con ausencia de reflejos en la pantalla
Bien orientado
Vistas agradables
Espacioso
Amplio
Buena temperatura
Sin frío
Sin calor
Sin corrientes de aire

Bien ventilado
Con buen diseño
Con suficientes cajones
Con mobiliario adecuado
Ergonómico
Ambiente poco cargado
Limpio
Acogedor
Cómodo
Agradable
Humedad adecuada

3) Se definió el espacio de propiedades.

El espacio de propiedades para este proyecto se basa en la elección de las variables de diseño en las cuales se van a modificar de cara a la optimización del confort y la mejora de la eficiencia energética. Dichas variables, seleccionadas a partir de la bibliografía recopilada serán:

- Temperatura.
- Humedad.
- Luminosidad.
- Ruido.
- Ventilación.
- Ubicación del puesto de trabajo.

4) Elaboración de encuestas Kansei.

Para la elaboración de la encuesta Kansei, habiendo determinado ya el dominio, el espacio semántico y el de propiedades, se determina el tamaño de muestra. Éste ha sido el número de trabajadores que se encuentran en las salas comunes de las plantas -1 y +1. Han sido un total de 37 personas.

Las encuestas están formadas por dos partes:

- Una parte objetiva en la que se recopilaron datos objetivos del usuario: nombre, edad, sexo, categoría profesional, etc., que permiten además entender la muestra y relacionar algunos de esos datos con las percepciones que influyen en el confort de cada uno.
- Una parte subjetiva basada en cuestionarios relacionados con la influencia de las percepciones en el confort y con la influencia de dichas percepciones en las variables de diseño.

A continuación, se muestra el modelo de formulario que se les hizo llegar a los diferentes trabajadores de las plantas -1 y +1 del Centro Tecnológico CTIC.

Dicha encuesta formada por tres bloques tiene por objetivo evaluar el confort térmico, lumínico, acústico y de calidad de aire, de forma general y de forma individual. En ella se conocen las percepciones (primera parte de la encuesta), la influencia que tienen éstas en el confort de cada usuario (segunda parte de la encuesta) y cómo influyen los elementos de diseño (temperatura, humedad, ventilación, etc) en las percepciones (tercera parte de la encuesta).

Dado que la información que se obtiene de la segunda y tercera parte de la encuesta es algo constante, y que no cambia con las condiciones ambientales, sólo la primera parte de la encuesta (relacionada con las percepciones) se está haciendo pasar dos veces por cada estación, de forma que se pueda evaluar el confort en el momento en el que se contesta la encuesta y que estos resultados puedan ser contrastados con el valor real que los sensores de temperatura, iluminación, humedad, ruido y calidad del aire que se están tomando en tiempo real.

T2.2. Análisis de resultados de las encuestas de confort

Una vez contestadas las encuestas, se realiza un estudio de los resultados basado en métodos estadísticos y regresiones lineales. El objetivo de este estudio preliminar será evaluar el confort general

Como ya se explicó en la tarea 2.1, la encuesta está organizada en tres partes:

- La primera parte, que corresponde con la tabla 1 refleja el valor de las percepciones que cada trabajador tiene con respecto al confort en su puesto de trabajo.
- La segunda parte, que corresponde con la tabla 2 refleja la influencia de dichas percepciones en el grado de confort de cada trabajador
- La tercera parte, que corresponde con la tabla 3 refleja la influencia que tienen los elementos de diseño en las percepciones.

De esta forma se sacan los siguientes resultados:

- Evaluación de las percepciones

Del estudio realizado en el cual se ha analizado la mediana, media y desviación típica se concluye que, a nivel global, la percepción con peor valoración es la percepción “Íntimo”, con una calificación media de **2,61**.

Si se estudia la sensación de confort general de acuerdo a 5 tipos de confort (térmico, lumínico, acústico, calidad de aire, ergonómico) junto con la sensación de confort general, se obtienen los siguientes resultados:

Tabla 3: Confort general de los trabajadores

Confort	Mediana	Media	Desviación típica	%
Térmico	3	3,065	1,153	51,63
Lumínico	4	3,548	0,925	63,70
Acústico	4	3,645	0,839	66,13
Calidad del aire	4	3,806	0,910	70,15
Ergonómico	4	3,839	0,638	70,98
General	**4**	**3,774**	**0,497**	**69,35**

En general, la sensación de confort es buena **(69,35%),** siendo la más baja en términos medios la correspondiente al confort térmico **(51,63%).** Por su parte, el confort mejor valorado en términos medios es el confort ergonómico **(70,98%).**

Gracias al análisis de las respuestas relativas a la sensación de confort general y de cada uno de los otros 5 tipos de confort, se puede establecer la relación existente entre ellos. La ecuación resultante es la siguiente:

Confort general = 0.2134 * Confort térmico + 0.0247 * Confort lumínico + 0.0458 * Confort acústico – 0.1247 * Calidad del aire + 0.3395 * Ergonomía + 2.0253

Paralelamente se estudió la sensación de confort general por plantas (planta -1 y planta +1) extrayéndose que:

- La sensación general de confort de los usuarios que trabajan en la planta +1 es ligeramente superior a la sensación general de confort de los usuarios que trabajan en la planta -1.
- La sensación de confort térmico de la planta -1, que está fuertemente penalizada (40%), restando la media de confort general
- La sensación de confort de los otros 4 tipos es, en media, superior en la planta -1.
- El confort lumínico medio de la planta +1 es el peor valorado de todos.
- Influencia de las percepciones en el confort
- Relación existente entre las variables de entorno y las percepciones analizadas

De la relación entre la influencia que para los usuarios tienen las percepciones con las variables de entorno, se pueden extraer para cada elemento de diseño o variable de entorno, correspondiente con los resultados de la Tabla 3, se concluye que:

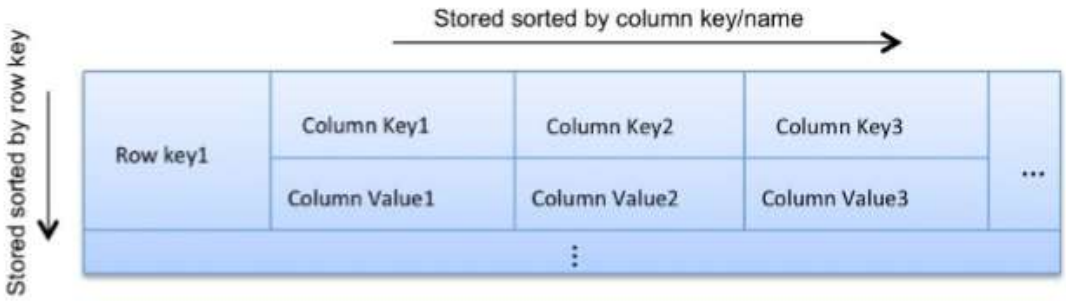
En general todas las percepciones están relacionadas en mayor o menor medida con alguna variable de entorno. Por otra parte, las percepciones “acogedor”, “cómodo” y “agradable” son transversales a todas las variables de entorno, es decir, no son discriminantes.

PT3: Sistema de adquisición y almacenamiento de datos

T3.1. Diseño, análisis y construcción de los sistemas de almacenamiento

Teniendo en cuenta los requisitos de rendimiento y gestión continuada de flujo de datos provenientes de la sensórica, gestión de grandes cantidades de datos y tiempos de respuestas rápidos, se han seleccionado las siguientes opciones para la adquisición y almacenamiento de datos:

- Respecto al almacenamiento de la información gestionada en el sistema ENERCONFORT, se ha de tener en cuenta la particular problemática que supone el almacenamiento de señales continuas de valores en el tiempo. Para dar respuesta a este problema, se ha decidido hacer uso del sistema Apache Cassandra5, que es una base de datos NoSQL concebida para gestionar grandes volúmenes de datos manteniendo una gran capacidad de escalabilidad que permite ofrecer un elevado rendimiento tanto en escritura como en lectura.
- Los datos se almacenarán en una misma "tabla" que permitirá guardar de forma independiente cada una de las variables de todos los elementos. Esta configuración proporciona una gran flexibilidad, ya que cualquier nueva variable proveniente de cualquier elemento puede ser registrada.



Con el objetivo de minimizar la redundancia de información en la base de datos Apache Cassandra en la que se almacena la información de carácter temporal, se necesita un sistema que permita contextualizar dicha información, habilitando con ello mecanismos de consulta dirigidos al dominio de explotación. Para ello se habilitará una herramienta en la que se modelará el dominio de la aplicación y que se asociará a las informaciones almacenadas en Apache Cassandra.

En este aspecto se contemplan dos alternativas.

Por una parte, si el modelado de los datos no fuese demasiado amplio y mantuviese una estructura homogénea, se podrían hacer uso de una herramienta relacional. El modelo relacional, para el modelado y la gestión de bases de datos, es un modelo de datos basado en la lógica de predicados y en la teoría de conjuntos. Su idea fundamental es el uso de relaciones. Estas relaciones podrían considerarse en forma lógica como conjuntos de datos llamados tuplas.

Por otra parte, en caso de que el modelado del dominio fuera demasiado complejo se optaría por la incorporación de un sistema de bases de datos NoSQL como podría ser MongoDB, que almacena la información a través de documentos. Cada uno de los documentos está formado por un conjunto de claves (campos) con valores asociados. Los documentos se organizan mediante el uso de colecciones que

permiten diferenciar tipos de documentos almacenando información diferente en cada una de las colecciones que se definan. Haciendo un símil con las tradicionales bases de datos relacionales (SQL) las colecciones serían como las tablas y los documentos como los registros. Una de las principales características de los documentos de MongoDB es la no existencia de esquemas predefinidos. Cada uno de los documentos que integran una colección puede seguir esquemas diferentes y son almacenados en formato BSON (versión modificada de JSON).

Respecto a la adquisición de datos se utiliza una capa de Internet de las Cosas, la arquitectura general del sistema contempla la incorporación de un mecanismo de mensajería general para que gestione la recepción y posterior tratamiento de los datos recibidos desde los dispositivos físicos. Apache Kafka7 es un sistema de mensajería que opera bajo el paradigma publicador/subscriptor y que se caracteriza por:

- *Rapidez*: Un bróker de Kafka puede manejar cientos de megabytes en lecturas y escrituras recibidas desde múltiples clientes.
- *Escalabilidad*: Un sistema Kafka puede ampliar sus capacidades de un modo elástico y transparente sin que se produzcan pérdidas de productividad. Los streams de datos pueden ser particionados y gestionados desde un clúster de máquinas, con la finalidad de que cargas de trabajo elevadas (superiores a la capacidad de cualquier nodo individual) pueden ser tratadas en coordinación de todas las máquinas que conformen el clúster.
- *Persistencia*: Los mensajes son almacenados en disco y replicados en el clúster, con la intención de evitar la pérdida de información. Cada bróker puede manejar magnitudes de mensajes del orden de Terabytes, sin que ello incurra en una penalización del rendimiento. El protocolo variará en función del dispositivo que se esté monitorizando, pudiendo ser, entre otros muchos, MODBUS (bien sobre TCP, bien sobre UDP, pudiendo ser la implementación RTU) o OPC.

PT4: Sistema de procesado, análisis y visualización de datos

T4.1. Sistema de procesado.

En cuanto al diseño de la arquitectura servidora Big Data, existen numerosas aproximaciones y enfoques para su estructuración y puesta en escena. Para el desarrollo del proyecto se valoraron tres opciones diferentes: Arquitectura kappa, Arquitectura lambda diferenciada y Arquitectura lambda unificada.

La primera, la arquitectura kappa, plantea un sistema en el cual toda la información es procesada y presentada en tiempo real. No existe como tal un almacenamiento de grandes volúmenes de datos que puedan ser analizados a posteriori. Dado el escenario del presente proyecto este tipo de diseño fue descartado, puesto que se pretende realizar un análisis inteligente de los datos recogidos de las diferentes fuentes que conforman el sistema.

La arquitectura lambda plantea una solución en la que conviven el procesamiento y presentación de datos en tiempo real con un análisis de los históricos de información recogidos. Teniendo en cuenta las necesidades del proyecto, resulta ser la aproximación más adecuada al diseño del sistema. En este tipo

de arquitecturas se plantean dos enfoques distintos: diferenciado y unificado. El primero considera el procesado y análisis de la información en tiempo real un módulo completamente independiente de aquel que utiliza los datos como un conjunto y realiza análisis offline. Los datos provenientes de ambos módulos son almacenados en sistemas independientes.

Por su parte, el enfoque unificado, no contempla la independencia de ambos módulos, estableciendo uno a continuación del otro en la cadena de procesado. Dadas las necesidades del proyecto, esta última opción, la arquitectura lambda unificada, es la que mejores prestaciones ofrece en su desarrollo. Una vez determinada la arquitectura de alto nivel, se procede a seleccionar las herramientas concretas con las que se cubrirán los requisitos de procesamiento en tiempo real y por lotes.

T4.2. Análisis de resultados

En esta tarea se están comenzando a diseñar y desarrollar algoritmos que predigan el confort de cada usuario.

El procedimiento consiste en recoger los resultados de las encuestas elaboradas por cada usuario y los datos recopilados por los sensores y aplicar técnicas de pre-procesamiento, que permitan, entre otros, filtrar los datos, detectar y eliminar posibles datos erróneos y calcular estadísticos básicos como son la media, la mediana o la desviación típica.En estos momentos se están desarrollando diferentes algoritmos de predicción. Algunos de estos algoritmos son la regresión lineal y logística, las redes neuronales, los árboles de regresión y el algoritmo conocido como Random Forest.

Los resultados devueltos por cada uno de estos algoritmos se evaluarán detenidamente, con el objetivo de identificar aquel que permita obtener el confort con el menor error posible.

ANUALIDAD 2017

A continuación, se explican detalladamente los trabajos y tareas desarrollados durante la anualidad 2017 con la que se ha concluido la ejecución del proyecto.

PT2: Elaboración y procesado de encuestas de confort

T2.2. Análisis de resultados de las encuestas de confort

Para la toma de datos con los que completar los análisis de resultados de las encuestas de confort, se han distribuido las encuestas en diferentes condiciones meteorológicas para una misma estación (un mal día de verano y uno bueno, un buen día en invierno y uno malo…), que han venido marcadas por los datos que registra el sistema. En un primer momento, se pensó utilizar algún repositorio de datos abiertos proveniente de la estación meteorológica más cercana de la que se pudiesen extraer dichos datos, como por ejemplo los datos proporcionados por la AEMET a través de su *AEMET OpenData API, que* permite el

acceso a los datos a diferentes usuarios con la posibilidad de que ese acceso ser periódico e incluso programado, desde cualquier lenguaje de programación, sin interfaces amigables, con posibilidad de autodescubrimiento y permite a los reutilizadores de información el incluir los datos de AEMET en sus propios sistemas de información. Pero a partir de un número concreto de peticiones diarias, se corta el servicio para la IP concreta y deja de tener acceso, porque determina que puede ser un ataque DDoS.

Sin embargo, finalmente se decidió optar por otra solución consistente en utilizar una pequeña estación meteorológica NETATMO ubicada en el edificio de CTIC.

Figura 3: Estación meteorológica Netatmo.

Se ha adoptado esta opción por un doble motivo. En primer lugar, para disponer de datos meteorológicos tomados “in situ” y, en segundo lugar, porque en los paquetes de trabajo posteriores, esta estación se convertirá en una “Thing” más que inyecte datos en la plataforma a implementar, quedando dichos datos almacenados y pudiendo ser procesados y visualizados de forma análoga a la de los datos que proporcionan el resto de sensores que integrarán el sistema.

Adelantamos en este punto una captura de pantalla de los datos que la estación meteorológica proporciona al sistema.

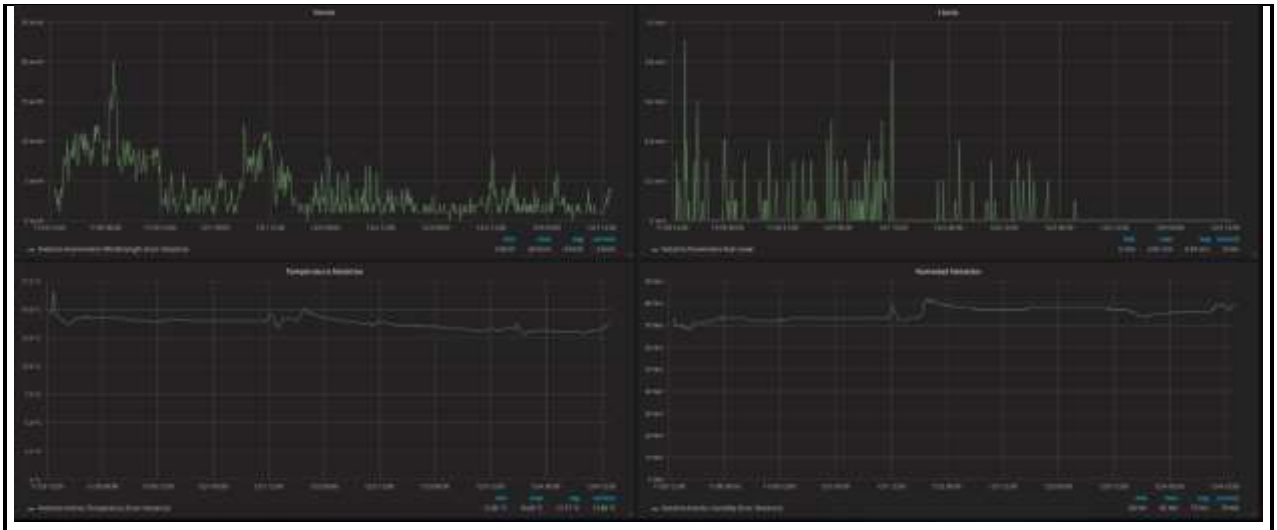


Figura 4: Datos capturados por la estación meteorológica Netatmo (de izquierda a derecha y de arriba abajo, velocidad del viento, lluvia, temperatura y humedad).

Así, en función de los datos proporcionados por la estación meteorológica Netatmo, y complementando a los trabajos de la anualidad anterior, se han ampliado los resultados de los análisis con los modelos desarrollados con el objetivo presentar los resultados de los modelos desarrollados, tanto de los modelos de cada tipo de confort, presentados en el apartado anterior, como del modelo de confort general presentado en el apartado "Evaluación de las percepciones".

La siguiente tabla muestra diferentes métricas de evaluación para cada uno de los 4 modelos entrenados (regresión lineal, M5P, Red Neuronal y Random Forest) con respecto al conjunto de datos de entrenamiento. En particular, se analizan el "*root mean squared error*", el "*mean absolute error*" y el coeficiente de correlación.

Tabla 3: Resultados obtenidos en la fase de entrenamiento para cada uno de los modelos entrenados

Confort	Regresión lineal			M5P			Red Neuronal			Random Forest		
	RMSE	MAE	Corr.	RMSE	MAE	Corr.	RMSE	MAE	Corr.	RMSE	MAE	Corr.
Térmico	0,440	0,194	0, 922	0,402	0,161	**0,937**	0,475	0,226	0,908	0,440	0,194	0,933
Lumínico	0,508	0,258	0,834	0,596	0,355	0,765	0,622	0,387	0,739	0,402	0,161	**0,901**
Acústico	0,508	0,258	0,802	0,568	0,323	0,796	0,475	0,226	0,823	0,359	0,129	**0,922**
Calidad del aire	0,539	0,290	0,799	0,596	0,355	0,754	0,539	0,290	0,799	0,539	0,290	**0,823**
Ergonomía	0,402	0,161	0,786	0,440	0,194	0,734	0,508	0,258	0,599	0,254	0,065	**0,920**
General	0,539	0,290	0,216	0,508	0,258	0,295	0,508	0,258	0,295	0,508	0,258	0,295

Como se puede apreciar, en el caso particular del cálculo del confort térmico, el algoritmo M5P es el que mejores resultados ofrece. Para el resto de tipos de confort, el algoritmo que mejor se ajusta es el conocido como Random Forest.

Nos quedaremos con el algoritmo M5P por ser más interpretable.

Con el objetivo de modelar un sistema que ofrezca mejores resultados que los mostrados anteriormente, se ha desarrollado un modelo experto que permite el cálculo del confort general. El procedimiento ha sido el siguiente:

1) Gracias al cuestionario ("Influencia de las percepciones en el confort") se conoce el grado de importancia de cada variable de diseño para los usuarios.
2) Cada tipo de confort está influido por una serie de variables de diseño, con un peso cada una. A partir de ellas podemos calcular el grado de importancia de cada confort para cada usuario.
3) Se contrasta el grado de importancia de cada confort con la puntuación recogida en la tabla 1 ("Evaluación de las percepciones"). Para ello se desarrolla una fórmula para el confort general que calcula la diferencia entre la puntuación y la importancia. Cuánto más negativa sea esa diferencia, peor (mayor será la importancia y menor la puntuación). Si la diferencia es positiva, es decir, tiene más puntuación que importancia, se fija a 0 (máxima puntuación).

La situación más desfavorable se da cuando la diferencia es igual a -4 (5 de importancia y 1 de puntuación). Por el contrario, la situación más favorable es que esa diferencia sea mayor o igual a 0, pero se fija siempre a 0.

A continuación, se realiza un cambio de escala, de manera que -4 (la peor situación) sea igual a 0 y 0 (la mejor situación) sea igual a 5.

Por último, se multiplica el valor resultante por la importancia de cada tipo de confort, para dar más peso según sea mayor la importancia.

Los resultados obtenidos con este modelo para el cálculo del confort general son los siguientes: RMSE = 0.516, MAE = 0.718 y correlación = 0.231. Estos resultados no mejoran los resultados proporcionados por el algoritmo M5P, por lo que se descarta su uso.

A continuación, se presentan los resultados obtenidos por los modelos ganadores para el conjunto de datos de test, es decir, un conjunto de datos nuevos, que no han sido utilizados en el entrenamiento de los algoritmos:

Tabla 4: Resultados obtenidos en la fase de validación para los mejores modelos obtenidos

Confort	Algoritmo	RMSE	MAE	Correlación
Térmico	M5P	0,943	0,667	0,612
Lumínico	Random Forest	0,882	0,556	0,457
Acústico	Random Forest	0,745	0,556	0,55
Calidad del aire	Random Forest	0,577	0,333	0,7
Ergonomía	Random Forest	0,667	0,444	0,655
General	M5P	0,667	0,444	0,438

Como conclusiones generales de la comparativa entre las plantas +1 y -1 se puede comprobar que la sensación general de confort de los usuarios que trabajan en la planta -1 es ligeramente superior a la sensación general de confort de los usuarios que trabajan en la planta +1. Esta diferencia sería aún más pronunciada de no ser por la sensación de confort térmico de la planta -1, que está fuertemente penalizada (44,05%), restando la media de confort general.

Por último, cabe destacar que, en media, el confort lumínico y térmico de la planta +1 son los peor valorados de esta planta.

Clústering jerárquico automatizado.

Para finalizar esta tarea, se ha realizado un clustering jerárquico que permita clasificar a los usuarios en base a similitudes en alguna de las características evaluadas. En este caso, se han realizado dos tipos de clusterings: el primero, basándose en las similitudes a la hora de evaluar las percepciones que tiene cada trabajador con respecto a su puesto de trabajo; el segundo, basándose en la sensación de confort que tiene que cada uno en relación a los 5 tipos de confort evaluados, además del confort general.

Clustering en base a las percepciones de cada usuario con respecto a su puesto de trabajo.

Como resumen, se puede concluir que los trabajadores de la planta +1 se quejan principalmente de reflejos en la pantalla y de la poca intimidad de sus sitios, debido en parte al elevado tránsito de gente por la planta. En cambio, los trabajadores de la planta -1 se quejan principalmente de que pasan frío y de la escasa iluminación natural.

Clustering en base a la sensación de confort de cada usuario.

En este caso se observa claramente que la sensación de confort de las dos islas superiores de la planta -1 es similar. Por otra parte, hay un grupo de personas en la planta +1 con sensaciones de confort similares.

Optimización.

Con el objetivo de incrementar el confort general de los trabajadores, se ha realizado un estudio que permita extraer qué variables de diseño hay que modificar, y a qué personas.

Según los modelos desarrollados, y antes de realizar ningún cambio en las condiciones de trabajo, los cinco tipos de confort, además del confort general, obtienen una valoración para cada uno de los trabajadores encuestados.

A partir de lo anterior, nos hemos fijado en aquellas percepciones que han sido valoradas con un 1 o con un 2 y que, por tanto, son susceptibles de ser mejoradas. Con este enfoque cumplimos dos objetivos: por un lado, que nadie valore negativamente (1-2) algún aspecto de su entorno de trabajo y, por otro lado, actuar sobre aquellas percepciones que tienen más margen de mejora.

Dicho lo anterior, se simula un incremento de las percepciones valoradas con un 1 o con un 2 hasta la valoración de 3.

Esta optimización pretende incrementar la sensación de confort general, por lo que nos fijaremos en la última columna. En este caso, vemos como hay 5 trabajadores que han aumentado su sensación de confort general, a pesar de que se han simulado cambios en las percepciones de 19 trabajadores. Estos 5 trabajadores sobre los que sí han surgido efecto los cambios, serán los elegidos para realizar realmente estas modificaciones. Sobre los 14 trabajadores restantes no se aplicará ningún cambio, ya que no tendrá efecto en su sensación de confort general.

Gracias a la actuación sobre las percepciones anteriores, conseguimos aumentar la sensación de confort general media de 3.47 a 3.61.

PT3: Sistema de adquisición y almacenamiento de datos

T3.1. Diseño, análisis y construcción de los sistemas de almacenamiento

Para la recepción de todos los datos entrantes del sistema, se ha optado por la utilización de Apache Kafka, un sistema de colas distribuido que garantiza la escalabilidad del sistema de adquisición de datos, independientemente del número de dispositivos que envíen datos.

La utilización de un sistema de colas distribuido para la recepción de los datos facilita su utilización por parte de rutinas de procesamiento en tiempo real, como las propias rutinas de almacenamiento o persistencia de datos. Los datos generados por otras rutinas de procesamiento que forman parte de la arquitectura también son publicados en Kafka, de manera que este componente actúa como punto de integración de todos los datos del sistema.

En el diseño desarrollado, las rutinas de almacenamiento leerán los datos publicados por los sensores en la cola Kafka en tiempo real, y los almacenarán de manera persistente en una base de datos.

Como base de datos a utilizar, se ha seleccionado Apache Cassandra, una base de datos noSQL especializada en el almacenamiento de series temporales. Su escalabilidad y gran rendimiento en la gestión de grandes volúmenes de datos la convierten en la opción óptima para el almacenamiento de los datos planteado como parte de ENERCONFORT. En particular, se ha seleccionado porque todas las fuentes de datos a monitorizar se presentan como series temporales, o un flujo de datos numéricos en tiempo real. En posteriores apartados se presenta un estudio de diferentes alternativas para el almacenamiento de datos y de la configuración específica de Cassandra para el almacenamiento de todos los datos de ENERCONFORT.

La utilización de un framework de procesamiento en tiempo real orientado al procesamiento Big Data para la implementación de las rutinas de almacenamiento garantizará a su vez la escalabilidad completa

del sistema, pues las tecnologías utilizadas para el sistema de recepción de datos (Apache Kafka) y almacenamiento (Apache Cassandra) han sido diseñadas con la escalabilidad como objetivo principal.

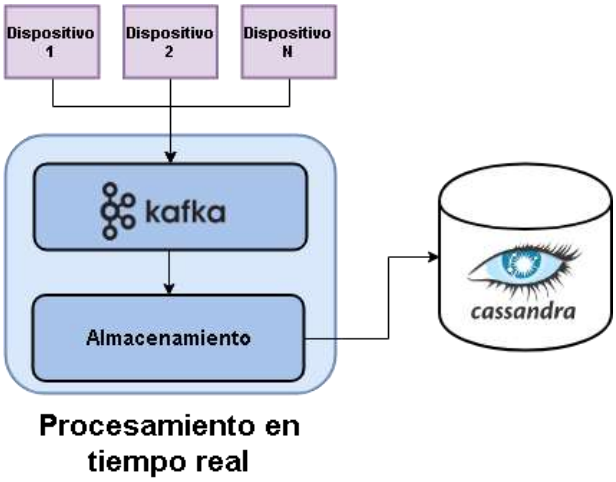


Figura 5: Diseño conceptual sistema almacenamiento

T3.2 Pruebas de concepto.

Para realizar las pruebas de concepto que validen los desarrollos de la tarea anterior, se definirá un entorno de pruebas y una serie de dispositivos a monitorizar. La evaluación del sistema de adquisición se realizará comprobando que un flujo de datos generada por los elementos sensores recorre todos los elementos de la arquitectura de procesamiento de datos definida. Esta está formada por un módulo de recepción de datos, una rutina de procesamiento de datos en tiempo real y una base de datos. Esta base de datos será utilizada por parte del dashboard de visualización para representar gráficamente la información almacenada.

Para la validación del sistema de adquisición y almacenamiento de datos implementado como parte de la tarea T3.1, se ha diseñado un banco de pruebas o prueba de concepto consistente en la monitorización de una serie de dispositivos desplegados en un edificio de varias plantas. Estos dispositivos actuarán como fuentes de datos, aportando valores en tiempo real de cada uno de los sensores que tienen instalados.

Estas pruebas consistirán en seguir la trazabilidad de los datos desde que estos son generados en uno de los dispositivos hasta que son mostrados al usuario en un dashboard web de visualización. La visualización de históricos asegura a su vez que todos los datos introducidos en la arquitectura definida han sido almacenados permanentemente.

Los siguientes apartados definen las distintas fuentes de datos utilizadas para la prueba de concepto, la metodología de evaluación y los resultados de la prueba de concepto.

Fuentes de datos

Como fuentes de datos para la evaluación del sistema de adquisición, se desarrollaron dos dispositivos que agrupan una serie de sensores. Estos dispositivos recogen los datos de la sensórica en tiempo real mediante unos módulos software de lectura de datos, y la introducen en la arquitectura definida mediante módulos de envío adicionales.

Los dispositivos fueron instalados en 7 distintos puestos de trabajo en varias plantas de un edificio de oficinas, obteniéndose, por tanto, un total de 14 fuentes diferentes. Su objetivo era el de medir las condiciones de confort a nivel de usuario, además de variables de consumo energético en cada planta.

Metodología.

La metodología de evaluación de la prueba de concepto consiste en corroborar que un subconjunto de los datos generados por los dispositivos recorre todos los elementos que forman parte del sistema de adquisición y almacenamiento de datos.

El flujo de un conjunto de datos desde que es generado en el dispositivo hasta que es visualizado en el dashboard web consta de los siguientes pasos:

1. Generación de los datos en el dispositivo
2. Publicación de datos en Kafka
3. Lectura de datos por parte de la rutina de almacenamiento
4. Almacenamiento de datos en Cassandra
5. Acceso a Cassandra mediante los servicios web de exposición de datos
6. Visualización de datos en el dashboard de visualización

Resultados.

Para la evaluación del sistema, se toman los valores más recientes de los sensores comentados con anterioridad.

El identificador único de cada sensor representa al sensor unívocamente en el interior de la arquitectura. Este será utilizado para realizar consultas sobre los datos provenientes de cada sensor en cada uno de los componentes del sistema.

De esta manera, si consultamos los últimos valores publicados en Kafka, podemos comprobar como las motas monitorizadas envían los datos relativos a los sensores a evaluar:

La utilización de un timestamp para la representación del instante en que se produce el valor de una variable permite identificar unívocamente la fecha, independientemente de la zona horaria en que esta se produzca.

PT4: Sistema de procesado, análisis y visualización de datos

T4.1. Sistema de procesado.

En esta tarea se evaluaron distintos sistemas de procesamiento de datos, teniendo en cuenta las características específicas de los datos de ENERCONFORT. En particular, se prestó atención a los frameworks de procesamiento Big Data, orientados al procesamiento de grandes volúmenes de datos.

A continuación, se mencionan las principales características de los sistemas de procesamiento de datos evaluados, agrupados por el tipo de procesamiento que realizan.

Sistemas de procesamiento Batch

El procesamiento por lotes (batch-processing) se realiza en conjuntos de datos estáticos de gran volumen, en los que el resultado completo se proporciona una vez se han procesado todos los datos.

Algunas características de los conjuntos de datos a los que se aplica el procesamiento por lotes son las siguientes:

- Finitud: El conjunto de datos tiene comienzo y fin, y su estado se mantiene a lo largo de todo el cálculo.
- Persistente: Los datos se encuentran almacenados en un sistema de almacenamiento.
- Gran volumen

El procesamiento por lotes es indicado para casos de uso en que se requiere analizar grandes volúmenes de datos estáticos, como por ejemplo el análisis de datos históricos. La desventaja de los sistemas de procesamiento por lotes es el elevado tiempo de computación, que no los hace apropiados para casos de uso en que la latencia o el tiempo de procesamiento resulten relevantes.

Hadoop

Apache Hadoop es un framework de procesamiento por lotes. Fue uno de los primeros frameworks de procesamiento Big Data open-source que posibilitaron el procesamiento de grandes volúmenes de datos.

Hadoop está formado por varios componentes:

- HDFS es un sistema de ficheros distribuido que se encarga del almacenamiento y replicación de datos en un clúster de servidores, ofreciendo alternativamente mecanismos de tolerancia a fallos. Es utilizado como fuente de datos, para el almacenamiento de resultados de cálculo intermedios y para la persistencia de los resultados finales.
- YARN es un componente encargado de la gestión de los recursos disponibles en un clúster de servidores. Es responsable de la coordinación de servicios y gestión de tareas a ejecutar en el clúster.
- MapReduce es el motor de ejecución por lotes de Hadoop.

El procesamiento de datos por lotes es realizado por MapReduce. Consiste principalmente en dividir un conjunto de datos extraído de HDFS en varios bloques que se distribuyen entre los distintos nodos de un clúster. Cada nodo aplica una serie de operaciones a cada bloque de datos independiente (Map) y los resultados intermedios de estas operaciones son combinados formando el resultado final (Reduce), siendo este almacenado en HDFS.

Ya que esta metodología depende del almacenamiento permanente proporcionado por HDFS, siendo necesaria la lectura y escritura a disco múltiples veces por tarea, suele resultar un rendimiento reducido. Por otra parte, ya que el espacio de disco es un recurso abundante, permite procesar cantidades de datos muy elevadas, siendo un framework recomendable en casos de uso en que el tiempo de procesamiento no es relevante.

Sistemas de procesamiento en tiempo real.

Los sistemas de procesamiento en tiempo real (*Stream processing*) procesan datos en el momento en que estos son introducidos en el sistema. Estos sistemas requieren un modelo de procesamiento diferente al del procesamiento por lotes. En lugar de definir operaciones que aplicar a un conjunto de datos en su totalidad, se definen operaciones que aplicar a cada dato que entre en el sistema.

Los conjuntos de datos para un sistema de procesamiento en tiempo real son considerados “infinitos”, ya que quedan determinados por la cantidad de datos que ha sido procesado por el sistema en un momento dado y es eternamente creciente.

El procesamiento está basado en eventos discretos y no termina a menos que se finalice explícitamente. Los resultados están disponibles inmediatamente y son actualizados en tiempo real a medida que llegan nuevos datos, al contrario del procesamiento por lotes que ofrece un único resultado al finalizar la computación del conjunto de datos completo.

Los sistemas de procesamiento en tiempo real pueden procesar grandes volúmenes de datos, pero sólo pueden procesar uno o un número reducido de elementos a la vez (micro-batching), siendo muy limitado el estado que es posible mantener entre datos procesados.

Las operaciones aplicadas por sistemas de procesamiento en tiempo real suelen estar compuestas de un conjunto de pasos discretos que no se ven afectados por el procesamiento de datos previos.

Este tipo de procesamiento es recomendable para determinadas cargas de datos, como la necesidad de procesamiento en tiempo real. La analítica de datos (predicciones en tiempo real) o procesamiento de datos de sensórica.

Storm.

Apache Storm es un framework de procesamiento de streams cuyo principal objetivo es proporcionar baja latencia, siendo uno de los más indicados para el procesamiento de grandes volúmenes de datos en tiempo real.

La idea principal tras Storm consiste en definir operaciones discretas interrelacionadas que aplicar a cada dato introducido en el sistema, formando un conjunto de operaciones conocido como topología.

Por defecto, Storm ofrece garantías de que cada dato introducido en el sistema será procesado al menos una vez, pudiendo darse el caso de que algún dato se procese varias veces en caso de error. Storm tampoco garantiza que los datos sean procesados en orden.

Storm es el framework recomendado para el procesamiento de streams puro con requisitos de latencias muy bajas. Al ser un framework específico de streaming, sería necesaria la utilización de software adicional si fuera necesario realizar procesamiento por lotes.

Sistemas de procesamiento híbridos.

Ciertos frameworks de procesamiento pueden procesar datos en lotes o en streams indistintamente. Estos frameworks estandarizan el procesamiento de ambos paradigmas de procesamiento mediante la definición de componentes y APIs comunes al procesamiento streaming y en lotes.

Mientras que los frameworks específicos para el procesamiento streaming o por lotes están más centrados en una tarea concreta, los frameworks híbridos plantean una solución general para el procesamiento de datos de todo tipo. Alternativamente, estos frameworks ofrecen librerías e integración con herramientas externas para la realización de tareas como análisis de grafos o machine learning.

Apache Spark.

Apache Spark es un framework de procesamiento por lotes con funcionalidad para el procesamiento de streams de datos. Basado en ciertos principios de frameworks de procesamiento anteriores como MapReduce, su principal ventaja es la mejora del rendimiento global al realizar el procesamiento en memoria RAM.

En el procesamiento por lotes, al contrario que MapReduce, Spark realiza todo el procesamiento en memoria, interactuando con la capa de almacenamiento al comienzo del proceso para cargar los datos en memoria o al final para almacenar los resultados. Todos los resultados intermedios son gestionados en memoria.

Spark es el framework más indicado para el procesamiento de cargas variadas, bien por lotes o en streams. El procesamiento por lotes ofrece un gran rendimiento aportado por el procesamiento de datos en memoria, y el procesamiento de streams, aunque no el óptimo para minimizar latencias, permite procesar datos en tiempo real con lotes de menos de un segundo.

Apache Flink.

Apache Flink es un framework de procesamiento de streams con funcionalidad adicional para el procesamiento por lotes. Esta funcionalidad es conseguida considerando un lote como un stream de datos finito, siendo posible tratar el procesamiento por lotes como un sub conjunto del procesamiento de streams.

Los streams son considerados como conjuntos de datos inmutables e infinitos, que entran continuamente en el sistema. El modelo de procesamiento de streams planteado por Flink se aplica a cada dato que llega al sistema.

El procesamiento de datos por lotes es una extensión del procesamiento de streams. En lugar de leer datos de un stream continuo, estos son leídos de un sistema de almacenamiento como un conjunto de datos finito. De esta manera Flink puede utilizar el mismo entorno de ejecución para ambos paradigmas de procesamiento.

Rutina procesamiento datos

De entre todos los frameworks de procesamiento evaluados, se optó por la utilización de Spark, debido a la versatilidad que ofrece a la hora de procesar cargas de datos variables, tanto por lotes como en tiempo real.

En particular, mediante este framework se implementó la rutina de adquisición de datos mencionada en la tarea T3.1. *Diseño, análisis y construcción de los sistemas de almacenamiento* del paquete de trabajo *PT3 – Sistema de adquisición y almacenamiento de datos.* A continuación, se presentan detalles de la implementación de la rutina, así como capturas que demuestran su funcionamiento.

Como se mencionó anteriormente, el framework de procesamiento en tiempo real Spark Streaming utiliza el paradigma de micro-batching, en el que los datos en tiempo real se procesan en pequeños lotes limitados por un intervalo de tiempo predefinido. En el caso de la rutina de adquisición de datos, este intervalo se fija en 2s, por lo que cada micro-batch contendrá todos los datos llegados al sistema en los 2s previos.

La rutina consta de los siguientes pasos:

1. Lectura de los mensajes presentes en un tópico de Kafka
2. Filtrado de mensajes válidos
3. Mapeado de mensajes a una estructura de datos compatible con Cassandra
4. Almacenado de resultados en una tabla de Cassandra.

Tanto la inicialización del contexto streaming como los pasos mencionados se pueden observar en la siguiente figura.

Figura 6: Rutina procesamiento

El proceso de filtrado de datos y el mapeo de datos mencionados aplican funciones específicas al flujo de datos entrante. La Figura 8 muestra la función de filtrado de datos que el framework de Spark Streaming aplicará a los datos entrante, de manera que la información inválida se elimine antes de ser procesada por el resto de funciones de la rutina.

Figura 7: Función filtrado datos

En la Figura 9 se pueden apreciar diferentes estadísticas sobre el procesamiento de los diferentes micro-batches o conjuntos de datos. En particular se mencionan el número de mensajes que llegaron al sistema en el intervalo de tiempo definido (Input Size) y el tiempo que se tardó en procesarlo (Processing Time). En ella se puede apreciar que **en 1 minuto se procesan aproximadamente 153 mensajes o valores independientes generados por sensores**, lo que supone el cumplimiento con creces del indicador de seguimiento para esta tarea, que se había establecido en un mínimo de 5 variables procesadas por minuto, siendo el sistema capaz de procesar más de 100.

La figura 10 muestra información global sobre todos los lotes de datos procesados por la rutina en tiempo real. En ella se puede apreciar la media de mensajes procesados por segundo (8.46 records/sec) o el tiempo de procesamiento medio de un lote (13 ms).

Análisis de datos.

A partir de las plataformas desarrolladas, se recogen de forma continua datos del estado de cada una de las áreas en estudio.

Como se desarrolla a lo largo de la distinta documentación del paquete de trabajo *PT2 - Elaboración y procesado de encuestas de confort*, se han realizado unas encuestas en las oficinas a los distintos trabajadores, disponiendo de, entre otros, la percepción de dicho usuario del confort general.

Todos los datos recogidos mediante dichas encuestas han sido obtenidos mediante escalas Likert (https://es.wikipedia.org/wiki/Escala_Likert) de, por lo general, cinco posibles valores. Dichas escalas Likert quedan determinadas por un número finito de posibles valores, los cuales están ordenados mediante un criterio previo, siendo habitual que el número de posibles valores oscile entre tres y siete.

Por último, también han sido proporcionados los consumos eléctricos registrados en el edificio de forma periódica, de modo que sea posible utilizar dicha información para la generación del modelo pertinente para la predicción del consumo eléctrico futuro.

Pre-procesado.

En cualquier proceso de analítica de datos es indispensable realizar una fase de pre-procesado suficientemente exhaustiva, pudiendo así analizar la tipología de los datos, obteniendo la adecuación de dicha información para el uso de los algoritmos. Debido a la dependencia de los modelos hacia los datos utilizados, una preparación inadecuada podrá resultar en un modelado erróneo.

En una fase de pre-procesado se pueden englobar tareas tales como el filtrado de los datos o la selección de atributos. En los siguientes subapartados, pasamos a desarrollar brevemente cada una de ellas.

Filtrado.

La primera parte ha consistido en analizar detalladamente los datos recibidos. En numerosas ocasiones, los dispositivos de recogida de información pueden tener un corte en la recepción de los datos, presentando casos vacíos que han de subsanarse. Para ello, se podrá proceder mediante técnicas de interpolación que permitan, a partir del resto de información disponible, generar los datos no recogidos estimándolos a partir de los valores cercanos. Existen distintas variedades de interpolación, dependiendo de la naturaleza de los datos.

También es necesario localizar valores anómalos que puedan haberse generado por errores en la recogida, en los procesos de comunicación o en el almacenamiento. Técnicas de detección de valores anómalos, comúnmente conocidos como *outliers*, resultan de especial importancia para evitar que los resultados se vean influenciados negativamente por parte de dicha información errónea.

Aparte de las técnicas generales de filtrado, ha de estudiarse cada uno de los tipos de datos recibidos por separado, analizando las características propias de cada conjunto, de modo que se puedan detectar aquellas situaciones anómalas que pueden no ser detectables por métodos de detección de *outliers* habituales. Véase casos donde la temperatura tome valores extremadamente bajos o altos, fuera de los valores usuales en una oficina, o si la humedad comienza a tomar valores fuera del intervalo 0-100, al tratarse de una variable con valores porcentuales.

Selección de atributos

En esta segunda etapa del pre-procesado se realiza un análisis conjunto de todos los datos disponibles para cada uno de los dos enfoques aquí presentados: clasificación multinomial y regresión.

Partiendo de todas las variables a analizar, el objetivo de esta etapa de pre-procesamiento será detectar qué variables son de interés de cara al análisis posterior para el modelado. Dicha fase se denomina *selección de atributos*, y existen diferentes herramientas que permiten desarrollarlo a partir de los datos previamente filtrados.

En concreto, es de especial interés la herramienta *Weka*, desarrollada por la Universidad de Waikato. Dicha herramienta, tras procesar los datos al formato correcto para su lectura, permite analizar cada una de las variables cargadas.

Además, proporciona, para cada variable seleccionada, un análisis descriptivo de ésta. En la siguiente figura, nos encontramos con el caso del grado de confort de los usuarios, utilizado en el conjunto de entrenamiento para proporcionar al sistema datos con los resultados conocidos, de modo que sea capaz de generar un modelo predictivo correcto.

Tras esto, es posible realizar diversos análisis de selección de atributos en la pestaña *Select Attributes* de *Weka*. Dentro del apartado *Attribute Evaluator*, aparece el desplegable mostrado abajo, el cual muestra las metodologías aplicables a los datos disponibles.

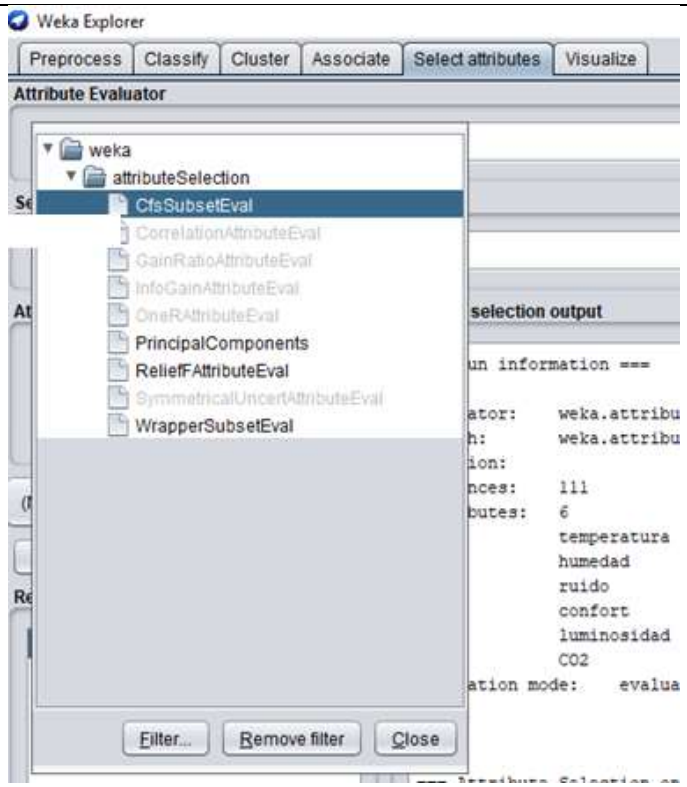


Figura 8. Opciones de selección de atributos para los datos cargados mediante la herramienta Weka

A modo de ejemplo, se muestra en la figura siguiente un análisis de componentes principales para dichos datos, teniendo en cuenta que el confort es la variable objetivo. Como se puede observar, proporciona la información indispensable dentro de un análisis de dicha tipología: la matriz de correlaciones, los valores propios, los vectores propios, y los nuevos atributos (componentes principales) generados, ordenados en función del porcentaje de varianza explicada.

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (unsupervised):
        Principal Components Attribute Transformer

Correlation matrix
  1     -0.35   0.01   0.03  -0.18
 -0.35   1      0.53  -0.2    0.53
  0.01   0.53   1     -0.11   0.24
  0.03  -0.2   -0.11   1      0.64
 -0.18   0.53   0.24   0.64   1


eigenvalue      proportion      cumulative
  2.0136          0.40272         0.40272       -0.591humedad-0.585CO2-0.424ruido+0.297temperatura-0.202luminosidad
  1.4877          0.29754         0.70026       0.759luminosidad+0.408CO2-0.344humedad-0.342ruido+0.148temperatura
  1.00418         0.20084         0.9011        0.843temperatura+0.534ruido+0.049luminosidad+0.039CO2-0.014humedad
  0.43274         0.08655         0.98765       0.641ruido-0.488humedad-0.418temperatura+0.304luminosidad-0.289CO2

Eigenvectors
 V1      V2      V3      V4
 0.297   0.1476  0.8432 -0.4185 temperatura
-0.5905 -0.3443 -0.0142 -0.4879 humedad
-0.4237 -0.3421  0.5338  0.6409 ruido
-0.2021  0.7592  0.0493  0.3041 luminosidad
-0.5853  0.4077  0.0385 -0.2891 CO2

Ranked attributes:
 0.5973  1 -0.591humedad-0.585CO2-0.424ruido+0.297temperatura-0.202luminosidad
 0.2997  2 0.759luminosidad+0.408CO2-0.344humedad-0.342ruido+0.148temperatura
 0.0969  3 0.843temperatura+0.534ruido+0.049luminosidad+0.039CO2-0.014humedad
 0.0124  4 0.641ruido-0.488humedad-0.418temperatura+0.304luminosidad-0.289CO2

Selected attributes: 1,2,3,4 : 4
```

Figura 9. Resultado de un análisis de componentes principales en Weka

Como es normal, la variable confort no está incluida dentro de este análisis al haberla seleccionado como variable objetivo. Al tratarse de una variable para modelar y contrastar, no deberá ser modificada ni eliminada por ningún proceso de selección de atributos.

T4.2. Análisis de resultados.

El objetivo general de esta tarea ha sido la generación de modelos que permitan:

- Predecir el grado de confort de los usuarios.
- Predecir el consumo eléctrico del edificio.

Para la obtención de dichos algoritmos de optimización, tanto para minimizar el consumo energético, como para la maximización del confort, se han realizado dos análisis independientes, donde **se ha desarrollado un algoritmo de regresión para la predicción del consumo energético, y un algoritmo de clasificación para la maximización del confort del usuario.**

Si bien dichos algoritmos de regresión y clasificación no pueden ser completamente equiparables a un modelo de optimización, permiten estimar los valores generados tanto para el consumo energético como para el confort del usuario. A partir de dichos modelos, es posible generar los valores estimados del

consumo energético y del confort de los usuarios, de modo que la modificación de los parámetros que interactúan en el desarrollo de los modelos, permitirá obtener aquellas configuraciones óptimas.

Como se ha destacado en la tarea anterior, el grado de confort de los usuarios se ha medido a través de escalas Likert en las encuestas Kansei desarrolladas. Es por ello que, el modelo de predicción del confort tendrá una salida de estructura análoga, es decir, con cinco posibles valores. Así, se tratará de un problema de clasificación multinomial, es decir, donde la salida del modelo generará un valor entre un conjunto finito de éstos.

Por otro lado, el consumo eléctrico del edificio está medido mediante una escala continua, es decir, dados dos valores cualesquiera, siempre será posible encontrar un valor intermedio para dicha escala. Por tanto, nos encontramos ante un problema de regresión.

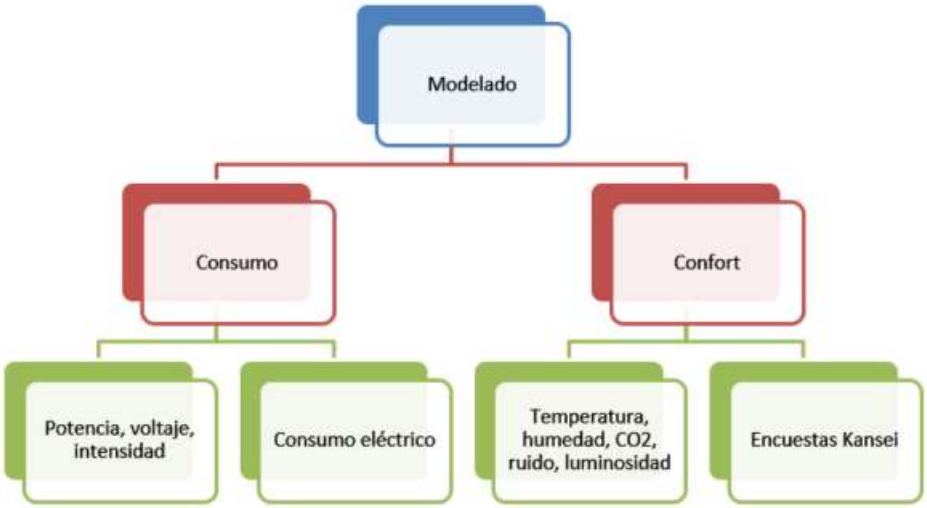


Figura 10. Esquema del modelado. Se muestran los dos enfoques: predicción del consumo eléctrico (regresión) y predicción del grado de confort (clasificación multinomial), junto con los datos utilizados por cada uno

A lo largo de los dos siguientes apartados, se contemplan los distintos algoritmos desarrollados para ambos enfoques: Predicción del consumo eléctrico (regresión) y predicción del grado de confort (clasificación multinomial), respectivamente.

1.- Algoritmo de predicción del consumo energético desarrollado.

En la Figura 11 se presenta de forma esquemática el procedimiento llevado a cabo para el desarrollo del módulo de predicción del consumo energético.

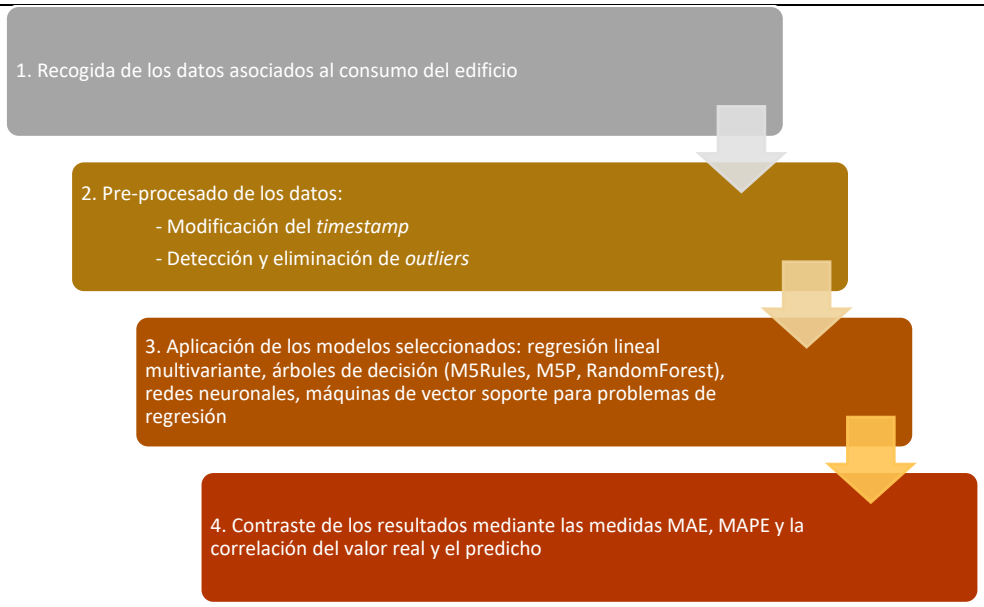


Figura 11. Esquema del proceso desarrollado para la generación del modelo de predicción del consumo energético

Las dos primeras tareas del algoritmo, recogida de datos y pre-procesado de los datos, se han explicado en los apartados anteriores de este informe técnico. A partir de los datos ya pre-procesados, se han aplicado los siguientes modelos de regresión, que tienen como característica distintiva que la variable dependiente u objetivo puede tomar valores en un espacio continuo, como ocurre en este caso con el consumo eléctrico (para cualesquiera dos valores de consumo eléctrico, es posible encontrar un valor intermedio posible).

A lo largo de los siguientes subapartados se desarrollan cada uno de los modelos de regresión considerados a lo largo del proyecto.

Regresión lineal.

Se trata de una de las técnicas estadísticas más utilizadas para el estudio de la relación entre variables, gracias a su adaptabilidad a diversas situaciones. Tanto en el caso con solo dos variables (regresión simple), como en las situaciones con varias variables independientes (regresión multivariable) que nos encontramos, es posible aplicar la regresión lineal con el objetivo de explorar y evaluar la relación entre la variable objetivo o dependiente y el resto de variables, comúnmente denominadas variables independientes. Dichos modelos vienen definidos mediante las siguientes expresiones:

- Modelo de regresión simple:

$$y = b_0 + b_1 \cdot x + u$$

- Modelo de regresión multivariable:

$$y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 + \ldots + b_k \cdot x_k + u$$

Como se indica, nos encontramos ante un problema multivariable, al disponer de más de una variable para la predicción del consumo eléctrico (potencia, intensidad, voltaje). En estos casos, a diferencia de la regresión simple, se genera un hiperplano en un espacio multidimensional.

Árboles de regresión.

Un árbol de regresión permite predecir el valor de la variable objetivo a partir de las diferentes variables de entrada disponibles. A partir de los datos de entrada, el sistema generará diagramas de construcciones lógicas con una gran similitud con los sistemas basados en reglas. De este modo, es posible representar y categorizar las diferentes condiciones que conforman el modelo sucesivamente.

Dado un árbol de regresión, éste estará formado por nodos internos, nodos de probabilidad, nodos hojas y arcos. Los primeros determinan un test a comprobar para alguno de los atributos de los datos. Los nodos de probabilidad proponen la ocurrencia de un evento aleatorio asociado a la tipología del problema. Los nodos hojas serán aquellos que aparecerán en la parte final del árbol y que proporcionan el valor obtenido para la variable objetivo. Finalmente, las ramas comunican los nodos de los árboles en función de las características de éstos.

La profundidad de los árboles generados resulta en una consecuente variación de los resultados obtenidos. Ha de ser un parámetro a controlar debido a la posibilidad de generar situaciones de *overfitting*, es decir, casos donde el modelo se ha adaptado en exceso a los datos proporcionados, no captando la generalidad del problema para otros casos no conocidos en la fase de entrenamiento.

A continuación, pasamos a describir tres tipos de árboles de regresión que podrán ser utilizados para el análisis de regresión aquí estudiado.

- **M5P**. Se trata de una reconstrucción del algoritmo M5 de Quinlan para inducir árboles a partir de modelos de regresión. El M5P combina un árbol de regresión convencional con la posibilidad de tener funciones de regresión lineal en cada una de los nodos del árbol.
- **M5Rules.** Genera una lista de decisión para problemas de regresión utilizando el algoritmo *divide-and-conquer*. En cada iteración, se construye un árbol modelo utilizando el método M5 de Quinlan y transforma la mejor hoja en una regla.
- **Random Forest**. Se trata de una combinación de árboles de regresión donde cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos, de modo que el resultado final se calcula a partir de una media ponderada del conjunto de árboles.

Los resultados serán dependientes del número de árboles a considerar, así como de su profundidad.

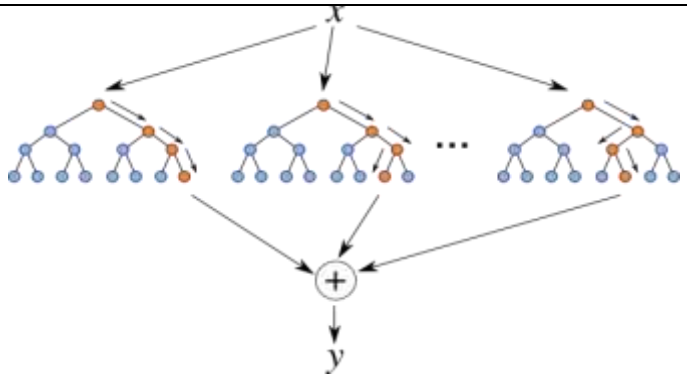


Figura 12. Representación gráfica de un Random Forest

Red neuronal.

Metodología inspirada en el funcionamiento del sistema nervioso animal. Está conformado por un sistema de nodos interconectados, conocidos comúnmente como redes neuronales. Está compuesta por una sucesión de capas constituidas por distintas unidades, denominadas neuronas. Cada una de ellas recibe unos valores de entrada a partir de las interconexiones con las neuronas de la capa previa, para generar una salida a proporcionar a la capa posterior. Se puede observar en la *Figura 13* un esquema sencillo de una red neuronal.

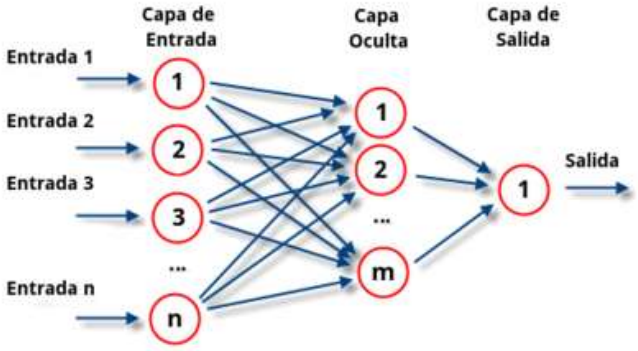


Figura 13. Esquema general de una red neuronal

Lógica difusa.

Dadas las características de la lógica difusa o *fuzzy logic*, tiene como principal beneficio el tratamiento de la imprecisión inherente a problemas como los aquí tratados. Permite, en algunas situaciones, evitar problemas que otros tipos de modelados no consiguen manejar correctamente, dada una alta imprecisión en los datos utilizados.

Máquinas de vector soporte (SVM).

Una máquina de vector soporte (SVM, *Support Vector Machine*) es un conjunto de algoritmos de aprendizaje supervisado aplicado a todo tipo de problemas, tanto de regresión como de clasificación. Si

bien su desarrollo principal se ha considerado para problemas de clasificación binaria, que se explicará debidamente en el apartado correspondiente de modelos de clasificación multinomial, también es posible aplicarla a casos de regresión.

Para ello, la idea radica en la realización de un mapeo de los datos de entrenamiento $x \in X$ a un espacio F de mayor dimensionalidad mediante $\varphi: \mathrm{X} \rightarrow \mathrm{F}$, de modo que sea aplicable una regresión lineal en dicho nuevo espacio. El caso de clasificación multinomial de SVM se denomina comúnmente como método *One-vs-All*, donde se enfrenta cada una de las clases (*One*) ante al resto (*All*).

Resultados.

El caso en estudio parte de tres variables independientes (potencia, voltaje, intensidad) y una variable dependiente u objetivo (consumo eléctrico). Como cabe esperar por la definición de éstas, existe una alta relación entre las tres variables independientes y el consumo eléctrico.

Es por ello que las técnicas de pre-procesado no han requerido filtrar ninguna de ellas mediante ninguna de las técnicas de selección de atributos de *Weka* mostrado en la tarea anterior. Así, los modelos se han generado finalmente a partir de tres variables independientes.

En la *Tabla 5* se muestran los resultados obtenidos para los distintos modelos considerados en el estudio del consumo energético. Para la comparación de los modelos, se han considerado las medidas MAE (en inglés, *mean absolute error*) y MAPE (en inglés, *mean absolute porcentaje error*).

Tabla 5. Resultados obtenidos para los distintos algoritmos considerados para el modelado de regresión de predicción de consumo energético

MODELO		MAE	MAPE	CORRELACIÓN
Regresión lineal multivariante		0.14	0.26	0.91
Árboles de decisión	M5Rules	0.11	0.17	0.92
	M5P	0.11	0.17	0.92
	Random Forest	0.11	0.17	0.91
Red neuronal		0.11	0.17	0.92
SVM (SVR)		0.12	0.24	0.91

Como se puede observar, se han tenido en cuenta todos los modelos indicados anteriormente, a excepción del modelo basado en la lógica difusa, al no considerarse que las propiedades proporcionadas por ésta sean adecuadas para los datos disponibles.

A través de dichos datos, los resultados han sido similares con respecto a las tres medidas consideradas, por lo que la decisión final tomada para dicha estimación ha estado influenciada por otras consideraciones. En concreto:

- *La gran correlación entre las variables independientes y el consumo eléctrico.*

- *El reducido número de variables independientes.*
- *El gran número de datos disponibles para el entrenamiento del modelo.*

Dichas características hacen que el método de regresión lineal multivariante sea mucho más interpretable, por lo que dicha mejora en la capacidad de interpretabilidad del modelo compensa el ligero empeoramiento de los resultados. Se trata de un modelo de regresión con una mayor eficiencia siempre y cuando el modelo no sea excesivamente complejo, como ocurre en este caso.

2.- Algoritmo de predicción del confort desarrollado.

De modo análogo al caso de predicción del consumo energético, se muestra en la *Figura 14* un esquema del procedimiento realizado para la generación del modelo de predicción del confort.

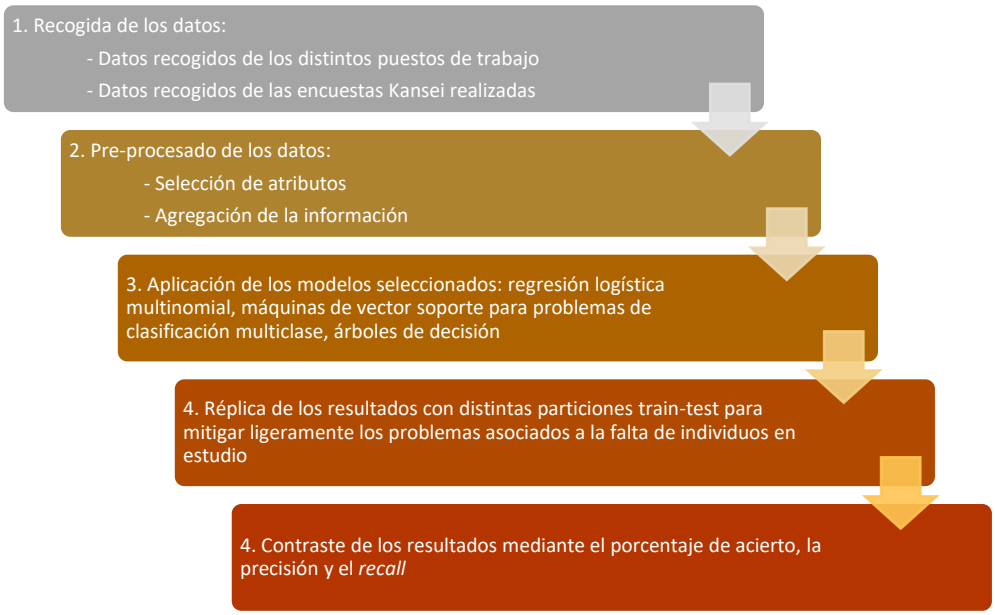


Figura 14. Esquema del proceso desarrollado para la generación del modelo de predicción del confort

Las dos primeras tareas del algoritmo, recogida de datos y pre-procesado de los datos, se han explicado en los apartados anteriores de este informe técnico. A partir de los datos ya pre-procesados, se han aplicado los siguientes modelos de clasificación multinomial. Estos modelos de clasificación tienen como característica principal la existencia de una variable dependiente u objetivo que toma un número finito de valores, denominadas habitualmente como clases o categorías.

Dentro de este tipo de modelos, podemos distinguir dos modalidades en función del número de clases posibles. En caso de encontrarnos con solo dos clases, se trataría de un modelo de clasificación binomial. En caso de existir más de dos, como es el caso que nos ocupa, nos encontramos ante un modelado de clasificación multinomial. En concreto, este caso se centra en un estudio de clasificación con cinco categorías distintas.

A continuación, pasamos a desarrollar cada uno de los modelos tratados para este caso particular.

Regresión logística multinomial.

Como se ha explicado anteriormente, la regresión lineal es una de las técnicas más habituales para los problemas de regresión. Sin embargo, existen adaptaciones a los casos de clasificación como el aquí estudiado.

La idea principal de la regresión logística multinomial se basa en la construcción de una función lineal predictiva que genera una puntuación a partir de un conjunto de pesos que son combinados linealmente con las variables independientes mediante el producto escalar:

$$SC(X_i, k) = \beta_k \cdot X_i,$$

donde X_i denota las variables independientes en forma de vector para dicha observación i, β_k el vector de pesos correspondiente a la clase k, y $SC(X_i, k)$ la puntuación generada para la observación i y la clase k.

Máquinas de vector soporte (SVM).

Como se ha indicado brevemente en el apartado de modelos de regresión, las máquinas de vectores soporte (SVM) han sido diseñadas en su origen para problemas de clasificación. Más concretamente, se ha estudiado en un inicio para problemas de clasificación binaria, donde se separan ambas clases en dos espacios mediante un hiperplano de separación dado como un vector entre dos puntos, denominado vector soporte, generando dichos espacios lo más amplios posibles. Así, los nuevos valores a clasificar corresponderán a la clase que pertenezca su espacio.

Árboles de decisión.

Se trata del último modelo aquí considerado para el caso de clasificación multinomial. En este caso, todo su desarrollo es análogo a los árboles de regresión presentados anteriormente, donde los nodos hoja, en lugar de proporcionar un valor continuo, solo tiene la posibilidad de generar un valor entre las diferentes clases del problema de clasificación multinomial correspondiente. Así, siguen existiendo los mismos tipos de nodos y de arcos, basados de nuevo en construcciones lógicas.

Resultados.

Este segundo caso parte de un total de cinco variables independientes (temperatura, humedad, CO_2, ruido, luminosidad) y una variable dependiente u objetivo (grado de confort). Del mismo modo que en el caso de la predicción del consumo eléctrico, existe una gran asociación entre dichas variables con el confort del usuario a predecir, tal y como se ha podido apreciar a partir de los análisis realizados de las encuestas Kansei a lo largo del paquete de trabajo *PT2 - Elaboración y procesado de encuestas de confort* y toda la documentación asociada a éste que se ha proporcionado a lo largo del proyecto.

Para reforzar dicha afirmación, se han aplicado distintos métodos de selección de atributos a través de la herramienta *Weka,* mostrados en la tarea anterior dando lugar a no existir ninguna razón para la eliminación de alguna variable o para un cambio de variables en busca de una reducción de la dimensionalidad (análisis de componentes principales). De este modo, los modelos generados han tomado cinco variables independientes como entrada.

En la *Tabla 6* se muestran los resultados obtenidos para los distintos modelos considerados en el estudio del confort. En este caso, se han aplicado distintas métricas a las utilizadas en el caso del consumo energético, al encontrarnos ante un caso de clasificación, y no de regresión. En concreto, se han considerado el porcentaje de acierto, la precisión y el *recall*.

La precisión muestra la proporción de falsos positivos, mientras que el *recall* lo hace con la proporción de falsos negativos. Véase que dichas métricas, cuanto más cercanas a 1, mejor actuación del modelo de clasificación.

Tabla 6. Resultados obtenidos para los distintos algoritmos considerados para el modelado de clasificación de predicción de confort

MODELO	PORCENTAJE	PRECISIÓN	RECALL
Regresión logística multinomial	0.31	0.23	0.24
SVM	0.35	0.28	0.28
Árbol de decisión*	0.5*	0.34*	0.46*

Nótese que los mejores resultados han sido obtenidos mediante el árbol de decisión, si bien se tratan de resultados engañosos, ya que dicho modelo toma en prácticamente todas las ocasiones el mismo valor. Dado que el número de individuos estudiado es muy reducido, los resultados obtenidos no presentan una fiabilidad suficiente como para seleccionarlos en función a criterios tales como las métricas anteriores. En caso de disponer en un futuro de una mayor cartera de usuarios, sería posible desarrollarlo de una forma más precisa.

Mediante la comparación de las características de los modelos analizados para este problema concreto de clasificación multinomial, se ha considerado que el modelo que mejor se ajusta son las **máquinas de soporte vectorial (SVM)**. Se trata de uno de los modelos más extendidos en los procesos de clasificación, siendo su comportamiento más adecuado en este tipo de situaciones que el resto de los casos analizados, al perder la regresión logística parte de sus propiedades al trasladarse al caso de clasificación, y no adaptarse de forma adecuada los árboles de decisión ante situaciones con un número reducido de variables.

En resumen, este modelo ha sido seleccionado como preferible sobre los demás por las siguientes razones:

- *La gran correlación entre las variables independientes y el grado de confort como se ha mostrado en las encuestas Kansei.*

- *El gran número de datos disponibles para el entrenamiento del modelo.*

Nótese que el segundo punto se refiere a los datos recogidos de las plantas del edificio mediante los sensores habilitados para tal fin, los cuales son suficientes para el análisis. La falta de un mayor número de trabajadores es la causa de la pobreza de los resultados, como se ha resaltado previamente.

T4.3: Diseño e implementación del sistema de visualización.

Sistema de procesado.

En esta sección se describen los componentes e interacciones que tienen lugar en el sistema desde el momento en el que una muestra es capturada en un dispositivo recolector, hasta el momento en el que se visualiza.

Dado que el sistema está compuesto de múltiples componentes interconectados, se presenta a continuación una descripción breve de la funcionalidad y características de cada uno de ellos.

Apache Kafka es un sistema de mensajería distribuido y escalable según el paradigma *publish-subscribe* que además ofrece persistencia. Los mensajes se publican y se agrupan en tópicos. Un *consumidor* (cualquier aplicación cliente) puede suscribirse a un tópico para escuchar todos los mensajes publicados en dicho tópico por uno o varios *productores*.

Apache Spark es un framework de procesamiento *batch* (ejecución por lotes) para conjuntos de datos masivos capaz de desplegarse sobre diversos gestores distribuidos de recursos (por ejemplo, Hadoop YARN o Mesos). También proporciona una API denominada *Spark Streaming* orientada al procesamiento en tiempo real. Apache Spark se integra fácilmente con Apache Kafka.

Apache Cassandra es una base de datos NoSQL perteneciente a la subcategoría de bases de datos *Column Family* (familia de columnas). Cassandra presenta una arquitectura P2P (sin maestros ni esclavos) que facilita la administración y además sirve de base para que Cassandra sea escalable horizontalmente de manera prácticamente lineal. Cassandra es especialmente adecuada para cargas de trabajo no-transaccionales con altos volúmenes de escritura, como es el caso en el que nos encontramos.

Graphite es un componente software especializado en el almacenamiento de datos de series temporales. Está compuesto por *Carbon* (un conjunto de *daemons* para recepción, cacheado, distribución de la carga y agregación de datos) y *Whisper* (una base de datos similar a RRD que almacena muestras de series temporales en disco).

A continuación, se presenta un diagrama que representa la visión global del flujo de procesamiento.

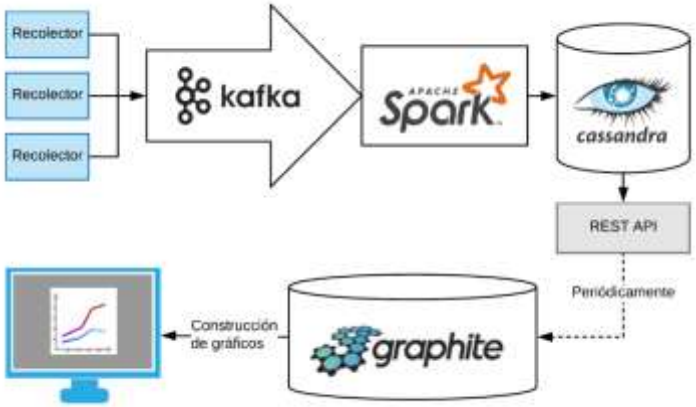


Figura 15: Sistema de procesado desde la captura del dato hasta su visualización.

Las muestras son inicialmente recogidas en los recolectores de datos. Estos dispositivos están repartidos por el entorno de despliegue y disponen de varios sensores (uno para cada variable monitorizada). El software instalado en los recolectores almacena temporalmente las muestras y las publica de manera periódica en un tópico de la cola de ingestión Apache Kafka.

Existe una aplicación desarrollada sobre la API de Spark Streaming que está en ejecución de manera continuada y que consume los mensajes publicados en el tópico anteriormente mencionado de Apache Kafka (el mismo en el que publican los recolectores). Cada vez que recibe un mensaje, le aplica un conjunto de filtros de limpieza y transformación para posteriormente lanzar una petición de inserción a la base de datos Cassandra.

Diseño e implementación del sistema de visualización.

En esta sección se presenta una descripción del diseño del sistema de visualización implementado. Dicho sistema ofrece la capacidad de mostrar la evolución de las variables capturadas por el sistema de manera multidimensional. El sistema es configurable y acepta la modificación de las escalas temporales o las frecuencias de refresco.

El sistema de visualización está organizado de manera jerárquica en dos componentes:

- **Vistas**: Una vista es una agrupación de componentes de vista que presentan una temática común (por ejemplo, una planta de un edificio). Una vista define el rango temporal por defecto que aplica a sus componentes contenidos y proporciona controles para forzar la actualización simultánea de todos ellos.
- **Componentes de Vista**: Cada componente de vista representa el estado instantáneo o la evolución temporal de una variable (por ejemplo, la evolución de la humedad capturada por un sensor en las últimas 6 horas). Un mismo elemento de esta categoría puede estar contenido en

distintas vistas. Los componentes de vista que representan una serie temporal permiten a su vez filtrar la escala temporal visualizada de manera individual.

En la siguiente figura se puede observar un diagrama que representa la organización explicada previamente.

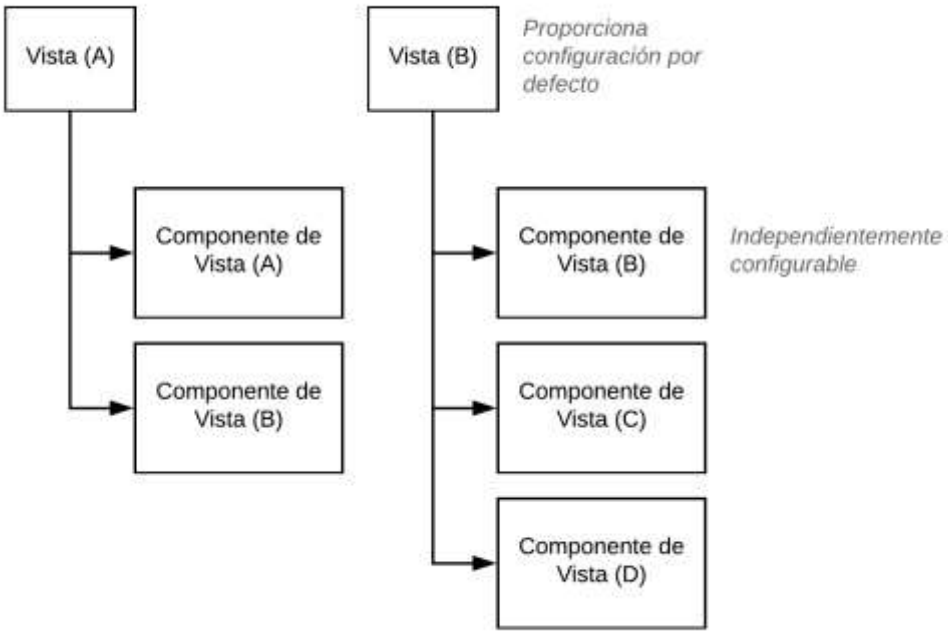


Figura 16: Organización jerárquica del sistema de visualización.

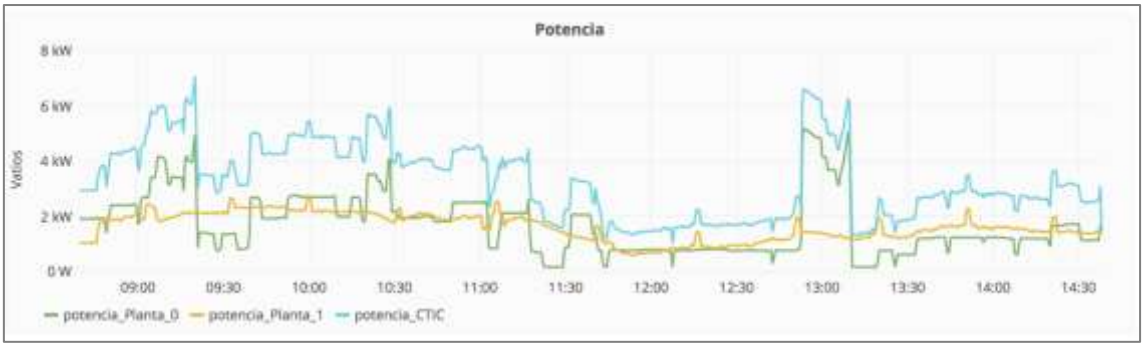
En la implementación final se definen seis vistas:

- Una vista que da información sobre las variables de confort global del edificio (*Confort*).
- Dos vistas de variables de confort, una para cada planta del edificio (*Confort Planta 0* y *Confort Planta 1*).
- Una vista de consumo global de energía del edificio (*Energía*).
- Dos vistas de consumo pormenorizado de energía para cada planta del edificio (*Energía Planta 0* y *Energía Planta 1*).

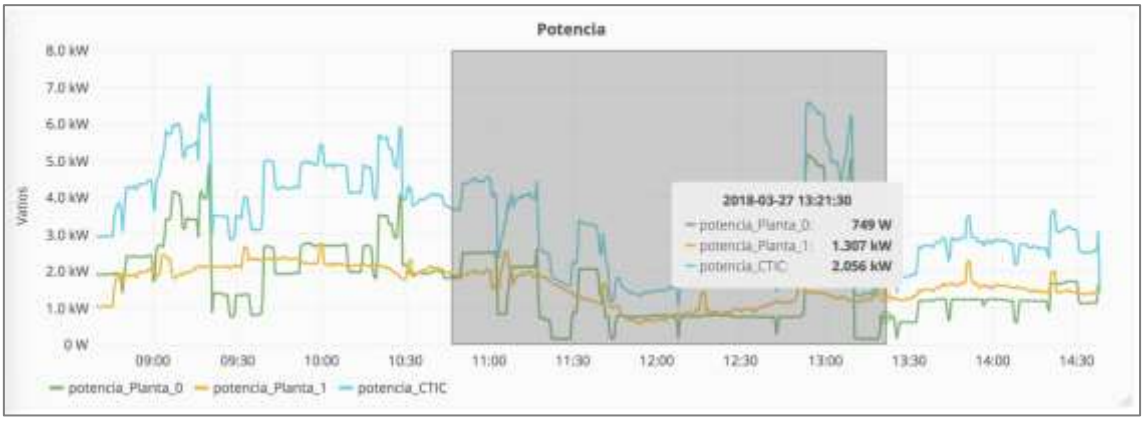
A continuación, se pasa a describir los componentes de vista contenidos en cada una de las vistas. Solo se describirá un miembro de cada uno de los pares (*Confort Planta 0*, *Confort Planta 1*) y (*Energía Planta 0*, *Energía Planta 1*) debido a que son equivalentes entre sí, la única diferencia notable es que el origen de datos reside en los recolectores instalados en una planta distinta. Téngase en cuenta que dichos recolectores siguen el mismo diseño general y son a su vez técnicamente equivalentes.

Vista *Energía Edificio*

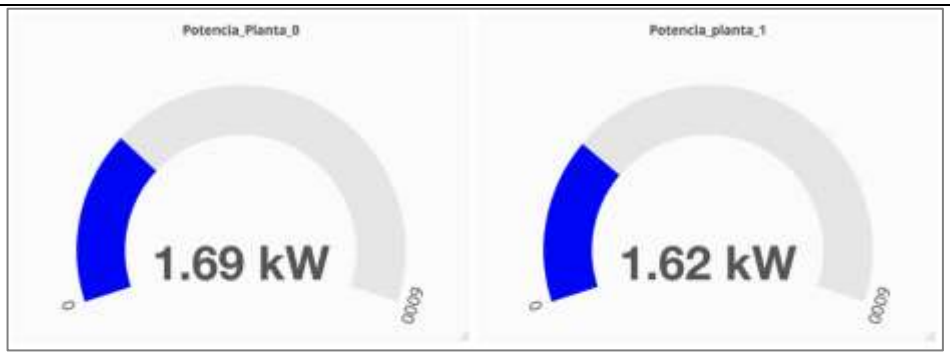
El componente principal de esta vista es el componente de *Potencia*, que muestra la evolución temporal de la potencia consumida global, así como las potencias de cada una de las plantas del edificio. La potencia consumida global se define como la suma de las potencias de los sistemas de iluminación y los SAI de cada una de las plantas.



A modo de demostración se puede ver a continuación un detalle en el que el usuario selecciona un intervalo de la gráfica con el objetivo de restringir el intervalo temporal. También se puede observar la ventana de detalle que aparece al desplazar el cursor por encima de la serie temporal.



Los componentes *Potencia Planta 0* y *Potencia Planta 1* muestran el valor instantáneo del consumo total de cada una de las plantas.

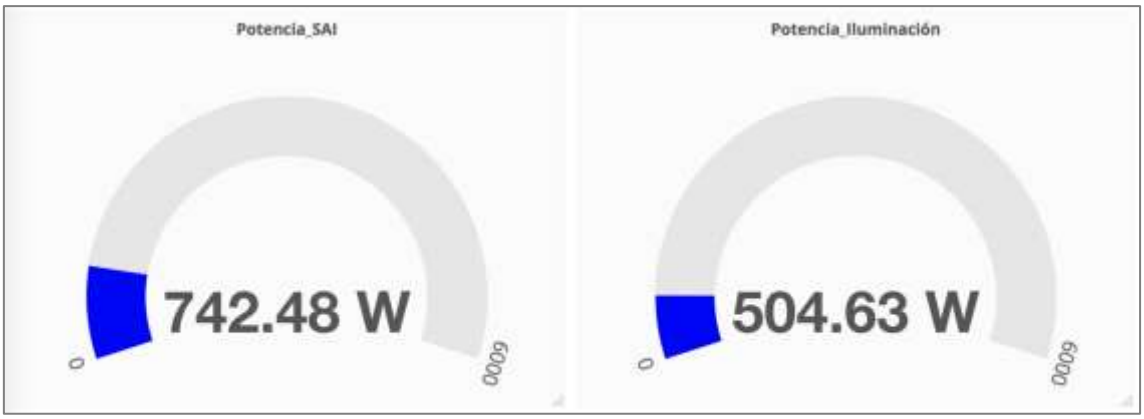


Finalmente, el componente *Potencia Total* muestra el valor instantáneo del consumo total de todos los elementos del edificio.

Vista *Energía Planta*

El componente *Potencia Planta* muestra la evolución de la potencia consumida global en planta, así como la potencia consumida por el SAI y el sistema de iluminación. La potencia global de planta se define como la suma de las series de iluminación y SAI.

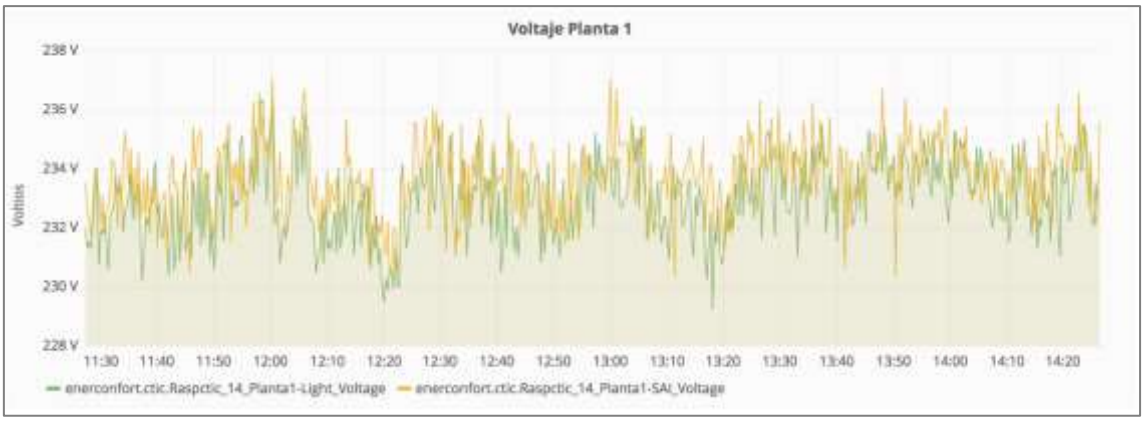
Los componentes de *Potencia SAI* y *Potencia Iluminación* proporcionan los valores instantáneos de las potencias representadas en el componente previo.



El componente *Corriente Planta*, muestra, de manera equivalente al componente de potencia, el amperaje consumido por el SAI y el sistema de iluminación, así como la serie temporal resultado de sumar ambos consumos.

Los componentes *Corriente SAI* y *Corriente Iluminación* proporcionan el valor instantáneo de los dos consumos de corriente.

Finalmente, se muestran los componentes que muestran la evolución temporal del voltaje, así como sus valores instantáneos.

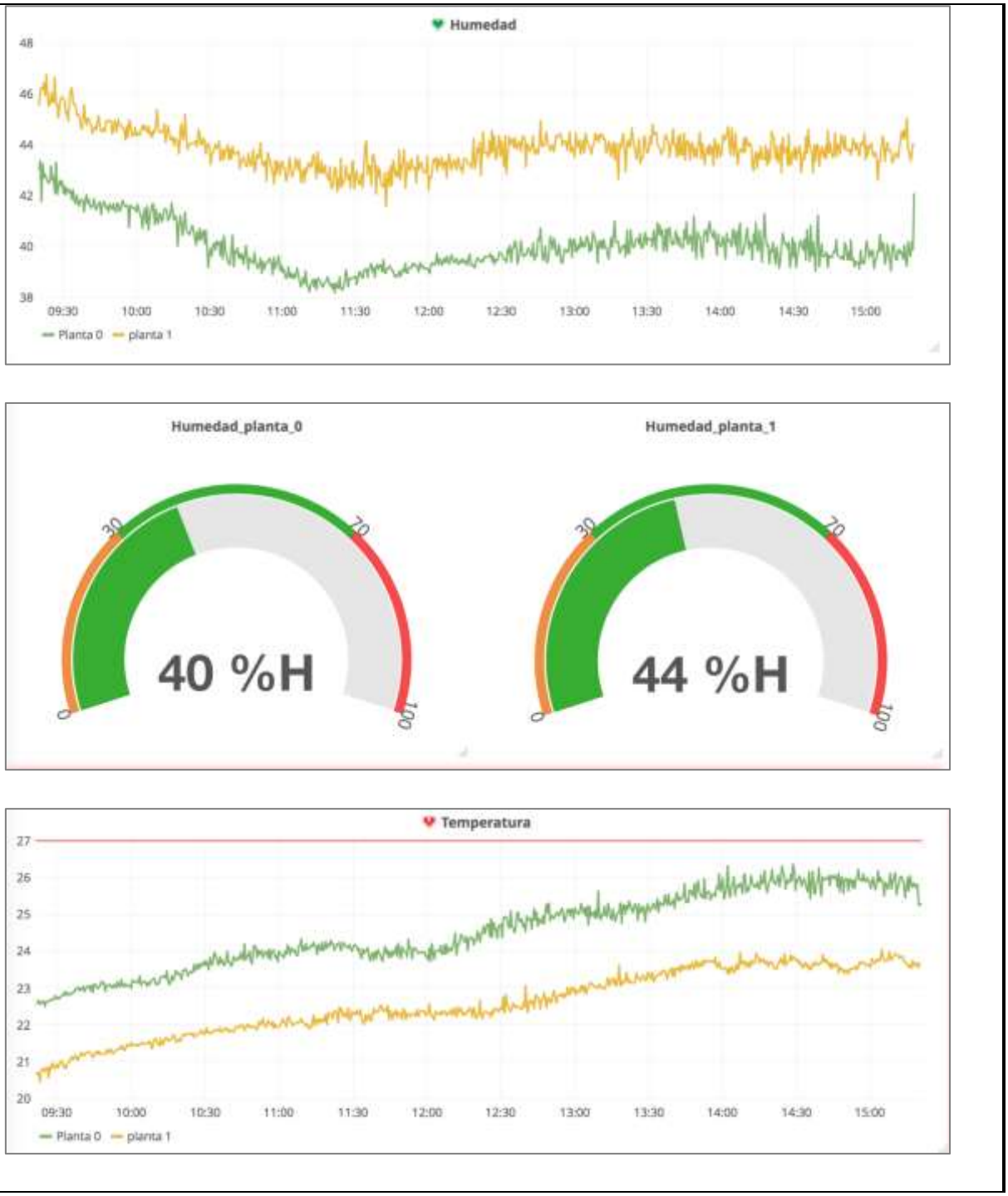


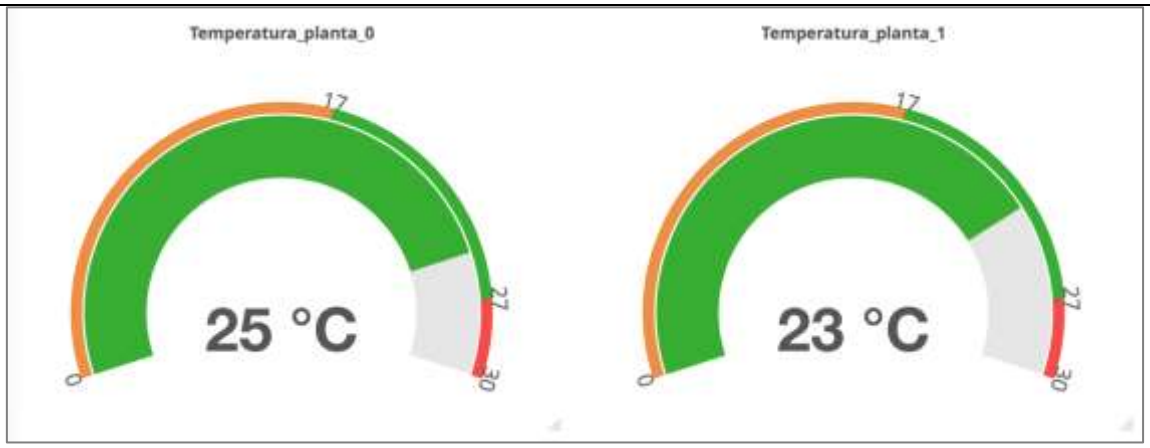
Vista *Confort Edificio*

Esta vista contiene cinco grupos de componentes para cada una de las variables de confort analizadas (humedad, temperatura, CO2, ruido y luminosidad). Cada grupo contiene a su vez un componente de serie temporal con la evolución de las variables en cada una de las plantas, así como dos componentes de valor instantáneo en cada planta.

El valor global de una variable en una planta se define como la media de los valores proporcionados por los recolectores instalados en dicha planta. Es decir, la temperatura que figura para la planta 1 es la media de temperaturas capturadas por los seis dispositivos instalados en la planta 1.

Para evitar una sobrecarga excesiva de figuras en esta sección se muestran los componentes correspondientes a humedad y temperatura. Los componentes de CO2 (ppm), ruido (dB) y luminosidad (lux) son equivalentes con las salvedades obvias en cuanto a escalas y unidades.


Humedad
48
46
44
42
40
38
09:30
10:00
10:30
11:00
11:30
12:00
12:30
13:00
13:30
14:00
14:30
15:00
Planta 0
planta 1
Humedad_planta_0
30
70
0
100
40 %H
Humedad_planta_1
30
70
0
100
44 %H
Temperatura
27
26
25
24
23
22
21
20
09:30
10:00
10:30
11:00
11:30
12:00
12:30
13:00
13:30
14:00
14:30
15:00
Planta 0
planta 1

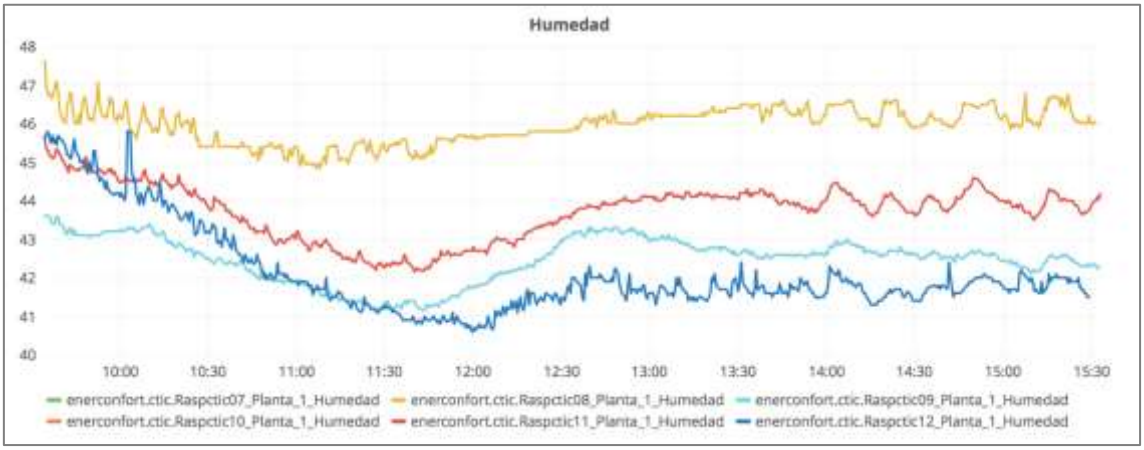


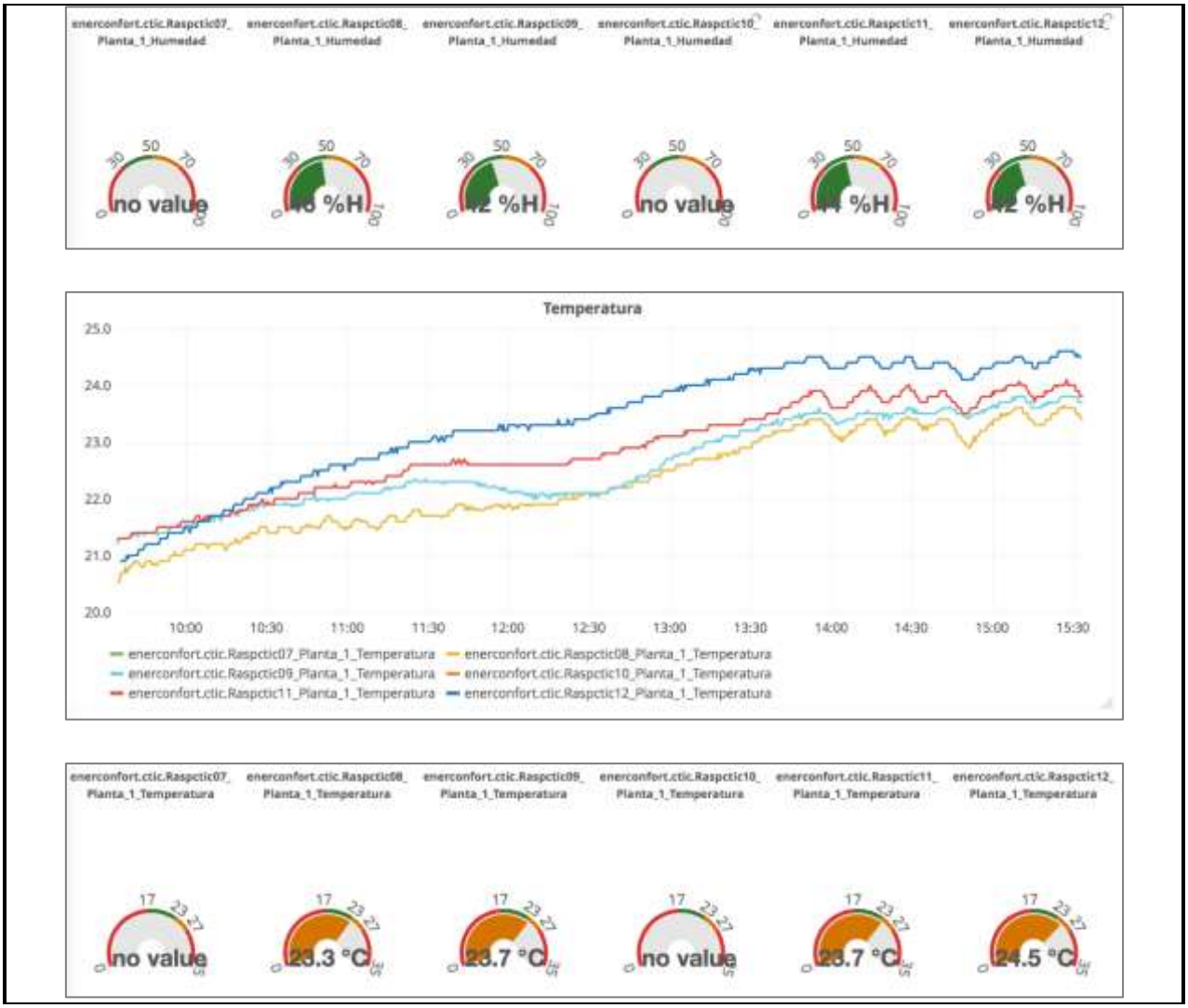
Vista *Confort Planta*

La vista de confort de planta presenta una vista pormenorizada de los datos capturados por los recolectores (cuya plataforma hardware es un ordenador embebido Raspberry Pi interconectado con un conjunto de sensores).

Al igual que en la vista de *Confort de Edificio*, se muestran las cinco variables de confort. Para cada variable se muestra un componente de serie temporal que contiene la evolución de las medidas de todos los dispositivos, así como un componente de valor instantáneo para cada uno de ellos.

De manera equivalente a la vista anterior, y para evitar una sobrecarga de figuras, sólo se muestran los componentes correspondientes a humedad y temperatura.





2. RESULTADOS CONSEGUIDOS

Durante desarrollo del proyecto se han conseguido todos los resultados planificados:

- Se ha analizado la problemática del consumo energético a partir de la recopilación de la normativa europea y nacional en relación a edificios y se han fijado los límites mínimos y máximos exigidos en relación a las condiciones ambientales en edificios de oficinas.
- Se han analizado las diferentes metodologías existentes en relación con las tecnologías que se están aplicando en este proyecto, y se han seleccionado aquellas tecnologías que más se adaptan al objetivo de ENERCONFORT.
- Se han elaborado las encuestas Kansei para poder establecer de forma subjetiva el confort de los usuarios del edificio de oficinas seleccionado como caso de uso. Se han analizado dichos resultados y llegado a fórmulas predictivas que permiten conocer el confort actual de los trabajadores y los factores que influyen en el mismo.
- Se han diseñado un sistema de adquisición y almacenamiento de variables que ha permitido centralizar la captación de datos procedente de sensores instalados en cada uno de los puntos clave seleccionados en el caso de uso.
- Se ha diseñado e implementado un sistema de procesado y análisis de datos adaptado a las diferentes fuentes de energía.

- Se han procesado los datos recogidos por los sensores y los obtenidos por las encuestas para integrarlos y poder desarrollar los algoritmos de optimización para la minimización del consumo energético y la maximización del confort.
- Se ha diseñado e implementado un sistema de visualización que muestra cómo influyen las diferentes variables en cada usuario, mostrando también las soluciones que se deben adoptar para maximizar el confort y minimizar el consumo energético.