ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA ELÉCTRICA Y ELECTRÓNICA

DETECCIÓN DE PÉRDIDAS NO TÉCNICAS EN CLIENTES ESPECIALES CON TELEMEDICIÓN, BASADA EN INTELIGENCIA ARTIFICIAL CON APLICACIÓN EN LA EMPRESA ELÉCTRICA AMBATO.

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE MAGISTER EN REDES ELÉCTRICAS INTELIGENTES

AUTOR:

LLAGUA ARÉVALO JOSÉ LUIS

DIRECTOR:

DR. PATRICIO ANTONIO PESÁNTEZ SARMIENTO

CODIRECTOR:

DR. PAÚL FABRICIO VÁSQUEZ MIRANDA

Quito, 04 de septiembre de 2023

AVAL

Certificamos que el presente trabajo fue desarrollado por Llagua Arévalo José Luis, bajo nuestra supervisión.

Firmado electrónicamente por:
PATRICIO ANTONIO
PESANTEZ SARMIENTO

Dr. Patricio Antonio Pesántez Sarmiento

DIRECTOR DEL TRABAJO DE TITULACIÓN

Firmado electrónicamente por:
PAUL FABRICIO
VASQUEZ MIRANDA

Dr. Paúl Fabricio Vásquez Miranda

CODIRECTOR DEL TRABAJO DE TITULACIÓN

DECLARACIÓN DE AUTORÍA

Yo, Llagua Arévalo José Luis, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración dejo constancia de que la Escuela Politécnica Nacional podrá hacer uso del presente trabajo según los términos estipulados en la Ley, Reglamentos y Normas vigentes.

Firmado electrónicamente por:
JOSE LUIS LLAGUA AREVALO

LLAGUA ARÉVALO JOSÉ LUIS

DEDICATORIA

Con todo el Amor del mundo para mi futura esposa CRISTINA ELIZABETH, todo el esfuerzo y sacrificio para alcanzar este logro, sin duda te lo mereces más que YO. Te Amo Mucho. Otro triunfo que Dios mediante festejaremos juntos.

Para mis hijos, para que con el trascender del tiempo, sepan que el trabajo dedicado y perseverante, siempre rinde sus frutos.

DYLAN STEVE, siempre serás el motivo por el que se creó un profesional, una vida y un hombre libre, líder decidido y amoroso. Orgulloso de ti, Siempre hijo mío.

DIDIER SEBASTIÁN, llegaste a este mundo para regresarme a la vida y renacer en nuestro hogar con el amor más puro que he podido sentir, Gracias a ti soy una persona, un hombre y un Padre ESPECIAL; porque seres como tú, hacen FAMILIAS ESPECIALES

A mi padre, GIOVANNI, gracias por seguir creyendo en mí y seguir brindándome tu apoyo, orgulloso siempre de llevar tu apellido.

A mi madre, MIRIAM DEL ROCÍO, por favor no me faltes nunca, mi eterna novia.

Para mis hermanas, hermanos, sobrinos y tíos que han sido como mis segundos padres, son importantes e indispensables en mi día a día.

Con todo el amor del mundo...

José Luis

Ser BUENO es fácil, lo difícil es ser JUSTO

AGRADECIMIENTO

A Dios, por las oportunidades brindadas.

A mi director, el Dr. Patricio A. Pesántez; quien colaboró incansablemente en el desarrollo de este proyecto.

A mi futura esposa, Cristina Elizabeth, sin duda sigues siendo mi inspiración, fortaleza y paz. Por siempre mi Beta.

A mis hijos, Dylan Steve y Didier Sebastián, por ser la fuerza y el motivo necesario para seguir adelante, a pesar de todas las adversidades.

A mis padres, Miriam del Rocío y Giovanni, su apoyo y motivación son dignas de personas fuertes de corazón y pensamiento.

A mi Hermana, Adriana del Rocío, la vida nos sigue preparando para grandes logros y batallas vencidas.

Al Msc. Luis Chiza, por su colaboración desinteresada para culminar con este trabajo.

Al departamento de la FIEE y a su secretaria, la Ing. Mónica Guerra por el apoyo brindado durante toda la formación académica de la Maestría.

A la EPN, por abrirme las puertas de tan grandiosa institución.

A Docentes, compañeros y colaboradores externos que hicieron gratificante y lleno de buenos y no tan buenos momentos el camino a conseguir este gran objetivo.

ÍNDICE DE CONTENIDO

RESUMEN

La EEASA como empresa distribuidora, almacena una gran cantidad de datos generados por los medidores inteligentes colocados en los clientes especiales; esta información histórica es analizada para detectar consumos anómalos que no son reconocidos fácilmente y son una parte significativa dentro de las pérdidas no técnicas de energía en la rendición de cuentas. La ejecución de técnicas adecuadas de clusterización con sus índices de validación del Machine Learning y el uso de Redes Neuronales del Deep Learning, trabajan a la par con el objetivo de detectar curvas engañosas que registran consumos de energía menores a los reales.

Dado ese contexto, se desarrolla una metodología que consiste en clasificar los consumos de energía en base a los días de la semana y feriados, que presentan comportamientos similares en los consumos de energía, para luego ser agrupados y finalmente ser llevados como conjuntos de datos para el aprendizaje, prueba y validación de una red neuronal clasificadora que está densamente conectada y poder identificar las curvas diarias descritas por estos clientes; obteniendo como resultado, la detección de patrones anómalos de consumo de energía.

PALABRAS CLAVE: Subregistro de Energía, Curvas Fraudulentas, Índices de Validación, Redes Neuronales, Pérdidas No Técnicas

ABSTRACT

The EEASA, as a distribution company, stores a large amount of data generated by smart meters placed in special customers; This historical information is analyzed to detect anomalous consumption that is not easily recognized and is a significant part of the non-technical energy losses in the accounting. The execution of adequate clustering techniques with their Machine Learning validation indices and the use of Deep Learning Neural Networks, work hand in hand with the objective of detecting misleading curves that register less than real energy consumption.

Given this context, a methodology is developed that consists of classifying energy consumption based on weekdays, weekends and holidays, which present similar behaviors in energy consumption, to then be grouped and finally taken as data sets to the learning, testing and validation of a classifying neural network that is densely connected and being able to identify the daily curves described by these clients; obtaining as a result, the detection of abnormal patterns of energy consumption.

KEYWORDS: Energy Subregistration, Fraudulent Curves, Validation Indices, Neural Networks, Non-technical Losses

ACRÓNIMOS

Adam: Estimación Momentánea Adaptativa

AMI: Medidor Inteligente Automático

AMR: Automatic Meter Reading

ARN: Red Neuronal Artificial

byPass: Conexión Eléctrica en Paralelo, Puente Eléctrico

CSV: Valores Separados por Comas

DBA: Promedio del Baricentro Dinámico

df: Data Frame

DTW: Deformación Dinámica en el Tiempo

EEASA: Empresa Eléctrica de Ambato S. A.

ETL: Extracción, Transformación y Carga

K: Valor de Grupos

k: valores de grupos

KM: K-medias

K-Means: Técnica de Clusterización

KNIME: Plataforma para Gestión de Datos

KW: Kilovatios

max: máximo

min: mínimo

NaN: No es un Número

NN: Redes Neuronales

PReLU: Unidad Lineal Rectificada Paramétrica

ReLU: Unidad Lineal Rectificada

ROC: Característica Operativa Del Receptor

SC: Coeficiente de Silueta

Soft-DTW: Suavizado del DTW

SVM: Máquinas de Soporte Vectorial

WCSS: Suma de Cuadrados Dentro del Clúster

1. INTRODUCCIÓN

Para una sociedad que depende altamente de la disponibilidad, eficiencia y confiabilidad de la electricidad, se hace indispensable para las empresas eléctricas del sector tener una buena administración de la producción y distribución de energía. Uno de los grandes problemas que preocupa a esta industria son las pérdidas eléctricas, que se produce mayormente en el segmento de distribución [1]. En el transporte de energía, las pérdidas son la diferencia entre la electricidad que ingresa a la red y la que es entregada para el consumo final, y son reflejo del nivel de eficiencia de la infraestructura en transmisión y distribución. El poder reducir las pérdidas es fundamental para incrementar la eficiencia de la distribución de energía, y en muchos casos puede apoyar incluso a mejorar la sostenibilidad financiera de las empresas de distribución. En el Ecuador, a nivel general, en los últimos años, en el sector eléctrico ecuatoriano se ha mejorado sustancialmente el nivel de pérdidas, bajando del 22% en el año 2006 al 12 % en 2019 [2], pero actualmente se ha incrementado al 13,02 %.

El concepto de pérdidas eléctricas incluye también la electricidad entregada pero no facturada, que se traduce directamente en pérdidas financieras y sirve como indicador del desempeño operacional de las empresas eléctricas. Las pérdidas eléctricas se encuentran divididas en dos tipos: pérdidas técnicas y no técnicas. Las pérdidas técnicas son aquellas que se dan al transportar la energía eléctrica, debidas al calentamiento natural de los elementos por el paso de la corriente eléctrica, además de la magnetización. Las pérdidas técnicas pueden reducirse hasta cierto punto, con inversiones en infraestructura, no obstante, son inevitables, ya que se deben a procesos físicos.

Las pérdidas no técnicas, por su parte, pueden clasificarse en tres tipos; 1) Robo, hecho por el cual un usuario realiza conexiones ilegales a la red eléctrica y consume la electricidad sin pagarla; 2) Fraude, son aquellas causadas por mano humana que al manipular medidores y/o cableados reduce la lectura del consumo de energía dando como resultado una factura inferior por el servicio consumido y 3) Errores en facturación y medición, los cuales pueden ser causados por una lectura errónea del consumo por causas humanas, medidores antiguos o mal calibrados, falta de equipos de medición o por el mantenimiento deficiente de equipos causando una lectura inapropiada de los consumos o incluso por errores en los sistemas comerciales.

Tradicionalmente los robos de electricidad se identifican a través de inspecciones con cuadrillas en áreas donde se presume que hay tendencia al fraude, ya sea con base a información histórica (estadísticas de consumo), o mediciones realizadas a nivel de subestaciones, donde se evidencia desbalance entre consumo de energía y facturación. Tradicionalmente la detección de patrones de consumos irregulares o sospechosos en el sector eléctrico relacionados con pérdidas no técnicas se ha realizado mediante revisión de información; sin embargo, comúnmente las empresas distribuidoras de energía no cuentan con perfiles de usuarios fraudulentos que puedan servir de prueba para realizar la validación. Se requiere identificar conjuntos o clases de usuarios representativos de comportamientos que pueden ser considerados como anómalos, para así, basados en estas clases identificar a los posibles infractores, por lo tanto, se plantea el uso de técnicas de Inteligencia Artificial (IA) para la detección de pérdidas no técnicas [3].

La EEASA cuentan con medidores de energía en la mayoría de los clientes especiales y esto les permite computar el consumo a ser facturado (comerciales y grandes clientes). Esta transformación requirió de una gran inversión y trajo importantes oportunidades comerciales, en particular como fuente de información para la detección de pérdidas no técnicas. A esta infraestructura de medición inteligente se la conoce por la sigla AMR (Automatic Meter Reading) o al término telemedición que es el más común en la distribuidora. Esta información ha sido la principal fuente de análisis para determinar qué clientes deben recibir una inspección. Algunas de las estrategias utilizadas en el pasado para generar una inspección estaban vinculadas a la detección de una disminución de consumo por debajo de algún umbral o seleccionar registros de consumos con muy poca varianza.

Sobre finales de la primera década del siglo XXI se comienza a investigar el uso de aprendizaje automático (o reconocimiento de patrones) para el análisis de los datos de los clientes y la identificación de posibles fraudes [4]. En los últimos 10 años, la detección automática de pérdidas no técnicas ha sido un campo muy activo a nivel académico. Al igual que en otras áreas del análisis de datos, recientemente ha evolucionado hacia técnicas de aprendizaje profundo y aprendizaje automático. El propósito de este proyecto es presentar una metodología de detección de pérdidas no técnicas utilizando las lecturas de los consumos cada 10 minutos de energía, utilizar varios índices de clasificación que garanticen la conformación de grupos, para finalmente llegar a implementar las técnicas de inteligencia artificial para el desarrollo del reconocimiento de patrones de curvas fraudulentas y no fraudulentas de los clientes que industriales que posee la EEASA.

1.1 Pregunta de investigación

El modelo propuesto pronosticará o clasificará los perfiles de demanda de los potenciales clientes especiales que estén sub registrando energía, lo que ayudará a disminuir las pérdidas no técnicas.

1.2 Objetivo General

Crear un modelo de Aprendizaje Automático o Deep Learning, basado en redes neuronales artificiales, para identificar consumos anormales de energía de clientes especiales y de esta manera disminuir las pérdidas comerciales.

1.3 Objetivos Específicos

- Construir una base de datos y clasificar un número determinado de clientes especiales con el fin de caracterizarlos según los datos de consumo de energía registrados en la EEASA durante el periodo del 31 de mayo del 2020 hasta el 31 de mayo del 2022.
- Procesar los datos de energía consumida obtenidos de cada cliente especial para que sean capaces de ser utilizados en el entrenamiento de la red neuronal artificial.
- Desarrollar un algoritmo capaz de leer y entrenar la red neuronal para clasificar o pronosticar patrones de comportamiento anómalo de clientes especiales.
- Realizar pruebas de validación, con registros de consumo de energía completamente desconocidos para el modelo propuesto.
- Validar el modelo en la base de datos que previamente se construyó de los clientes especiales de la EEASA.
- Estimar el impacto económico y de energía debido a las pérdidas no técnicas de los clientes especiales de la EEASA que conformaron la base de datos para el desarrollo del proyecto.

1.4 Alcance

Aprender a manipular el “Sistema de Telemedición para clientes especiales Easymetering AMI Solutions” de la EEASA, de tal forma que se pueda conectar en modo remoto con una de las computadoras que tenga acceso a los datos generales de los clientes especiales y así descargar los registros de energía consumida por cada cliente, para así consolidar la Base de Datos de 100 clientes que corresponde al 35% de la totalidad de clientes especiales con telemedición, para trabajar en el aprendizaje automático.

Acondicionar los datos obtenidos de los clientes, de tal forma que describan la misma estructura para poder ensamblarlos en un solo archivo XLSX que servirá como base de datos; para el proceso de entrenamiento se utilizará el 70% de los datos de los clientes, seleccionados indistintamente y no consecutivos, el restante 30%, se dividirá en dos para las pruebas 15% y la validación del modelo el otro 15%.

El modelo de la red neuronal artificial se desarrollará con KNIME, el cual es un programa basado en Python pero con interfaz gráfica, donde: los datos de entrada serán los registros de consumo de energía durante 2 años en intervalos de horas de los clientes especiales de la EEASA y se especificará tres incidencias: el número de capas que tiene la red, el número de neuronas de cada capa que estarán densamente conectadas y la función de activación que se usará en cada una de las capas, donde se podrán utilizar las funciones Sigmoide o ReLU. Con estas especificaciones y con la adecuación de los datos y el algoritmo programado, se puede definir generalmente como será el modelo de la ARN teniendo como salida el valor de la clasificación que predecirá si el cliente está sub registrando energía.

Una vez entrenada la red neuronal y con los resultados obtenidos durante los procesos de validación y prueba con registros de clientes especiales desconocidos para el modelo, se tendrá el caso de uso con consumos de energía nuevos para la red neuronal.

Se clasificará los consumos de energía que pueden ser identificados como registros anormales, logrando obtener valores normalizados de energía que serán registros de consumo de energía no facturados y por consiguiente valores económicos que darán un valor porcentual a la reducción de las pérdidas no técnicas que posea la distribuidora.

1.5 Marco Teórico

1.5.1. Estado del Arte

Los avances para la identificación de pérdidas no técnicas han sido innegables recientemente, los modelos de identificación de pérdidas no técnicas se valen de informaciones para detectar patrones o encontrar áreas donde hay grandes cantidades de pérdidas no técnicas de energía; así como también metodologías basadas en mediciones en campo a los centros de transformación de la red de distribución y técnicas basadas en inteligencia artificial. por lo que es necesario investigar métodos alternativos que tengan mayor flexibilidad y se adapten fácilmente al contexto del problema.

En la reseña [5] se muestra, uno de los métodos más tradicionales de identificación de Pérdidas no Técnicas con el Escalón de Consumo, el cual resulta del análisis del histórico de consumo del usurario; cuando el consumo del cliente es reducido significativamente por algunos meses consecutivos, caracteriza el escalón de consumo. La clasificación de clientes sospechosos a partir del análisis de sus datos de factura realizado por un especialista de la empresa de distribución es una práctica poco eficiente [5]. Los trabajos documentados de este tipo apuntan que las sospechas con este método solo tienen una precisión del 13% a 15% [6].

Para el método de la referencia [7], se fundamenta en el uso de medidores portátiles en miniatura para la inspección previa. Estos contadores se instalan discretamente en el ramal de conexión del usuario y registran su consumo durante algún tiempo para luego comparar las lecturas del medidor miniatura y el del usuario y si hay grandes diferencias entre las medidas, se realiza la inspección. En proyectos piloto, la tasa de éxito en el reconocimiento de fraudes fue del 100%, teniendo como desventaja la considerable cantidad que se necesita de equipos y personal técnico para realizar las mediciones sospechosas.

Otros métodos utilizados para detectar las Pérdidas No Técnica de Energía, son los que utilizan inteligencia artificial para señalar posibles fraudes o defectos en los contadores de energía en los sistemas de distribución, implicando que, con la ayuda de programas informáticos con capacidad de aprendizaje automático es viable aumentar la posibilidad de encontrar contadores de energía donde existen estas pérdidas.

Las metodologías que utilizan inteligencia artificial pueden detectar automáticamente nuevos patrones, así como analizar modelos ya conocidos por la experiencia humana, mejorando así la tasa de éxito esperada en la inspección en sitio, como los trabajos citados

a continuación: Máquinas de Vector de Soporte (SVM), es una técnica de aprendizaje automático que se ha utilizado con éxito en la detección y localización de Pérdidas No Técnicas de Energía, especialmente porque tiene la capacidad de brindar soluciones de gran generalidad a problemas de clasificación de patrones, debido a que la clasificación es una función no lineal, su entrada son las características de las muestras y la salida es la clase a la que pertenecen. La formulación original de SVM trata con problemas de clasificación binaria, pero puede extenderse a problemas de clasificación con múltiples clases a un alto costo computacional. La SVM puede realizar aprendizaje supervisado o no supervisado. El método de clasificación se utilizó para entrenar y probar el clasificador SVM utilizando los datos empíricos. El conjunto de datos constaba de 3156 clientes. Este conjunto de datos se particionó aleatoriamente en un conjunto de entrenamiento y de prueba. Aproximadamente dos tercios de la base de datos formaron un conjunto de entrenamiento y el tercio restante un conjunto de prueba. El conjunto de capacitación consistió en 2102 clientes, de los cuales aproximadamente el 78% eran casos limpios y el 22% fraudulentos. En comparación, el equipo de prueba estaba compuesto por 1052 clientes, de los cuales el 75% estaban limpios y el 25% eran fraudulentos [8].

La metodología empleando redes neuronales artificiales, busca realizar el procesamiento de datos imitando una red neuronal natural, como el cerebro humano. Para la activación de la técnica, se requieren un historial de consumo del cliente, se comprobó que el aprendizaje de esta metodología implementada mejora significativamente la tasa de éxito en la clasificación de clientes identificados como sospechosos. Una de las limitaciones de esta técnica es que la identificación de consumidores irregulares depende en gran medida de los datos utilizados en la formación, que se obtienen a través de las inspecciones en campo a los contadores de energía, que caracteriza el aprendizaje supervisado. Por lo tanto, las redes neuronales aprenden a identificar solo los patrones de clientes irregulares dentro del grupo de consumidores considerados fraudulentos. Esta metodología puede ser criticada por limitar el aprendizaje del reconocimiento de patrones a una base de entrenamiento que no forma una muestra estadística significativa. Una alternativa sería obtener datos de inspecciones realizadas al azar para obtener una medida estadística más significativa o emplear métodos de aprendizaje no supervisados. Los resultados obtenidos del sistema propuesto no fueron muy satisfactorios, debido a que del 100% de los resultados obtenidos como fraudulentos, sólo el 65.03% de los clientes preseleccionados para ser inspeccionados en campo, resultaron con fraude [9].

El K-medias (K-Means) es un algoritmo que pertenece a las técnicas de agrupamiento no jerárquico o particionales y es el más popular dentro de esta categoría y los datos usados para activar esta técnica corresponden a 5000 perfiles de clientes industriales de una empresa brasileña, donde 280 de ellos presentan pérdidas no técnicas. También se aportan algunas características que consideran relevantes como: demanda facturada, demanda contratada, demanda máxima, entre otras. Los datos se dividieron en 50% para el entrenamiento y el 50% para la prueba. En los resultados obtenidos se observó que los algoritmos mejoran su desempeño cuando se incluyen ciertas características de los clientes. Este enfoque presenta algunas debilidades, una de ellas es que no toma en cuenta el desequilibrio en el agrupamiento de las clases del entrenamiento [10].

La aplicación de los algoritmos de lógica difusa es una técnica también aplicada para el agrupamiento es una de las más utilizada para determinar perfiles de carga. En la referencia [11], se emplea para identificar perfiles de consumo sospechosos comparándolos con perfiles de consumo regulares; este esquema se compone de dos pasos: el primer paso, se utiliza el algoritmo C-medias difuso para agrupar los clientes dentro de las clases correspondientes, es decir, con perfiles similares. Posteriormente, mediante una matriz de pertenencia difusa y la distancia euclidiana a los centros de agrupamiento se clasifican los clientes en fraudulentos o no fraudulentos. Para desarrollar el algoritmo se utilizaron los datos históricos de consumo de los últimos seis meses de 20126 clientes de un área residencial y cinco atributos para el estudio (promedio de consumo, máximo consumo, deviación estándar, cantidad de inspecciones y promedio de consumo del área residencial). La tasa de éxito obtenida fue de 74.5% [11].

Un enfoque particularmente prometedor para la detección y cuantificación de pérdidas no técnicas se basa en métodos de estimación de estados. Tales métodos tienen una larga historia de éxito en sistemas de transmisión y desde la década de 90 su aplicación a redes de distribución ha sido investigada [12].

En el artículo de la referencia [13] se propone una metodología basada en la estimación del estado para un sistema de distribución y la prueba de error grueso basada en la colinealidad, capaz de no solo identificar la ocurrencia de pérdidas no técnicas, sino también de estimar su magnitud. Para ese propósito, se emplea un estimador de estado trifásico que procesa las mediciones disponibles en tiempo real y carga los valores de pronóstico.

La estimación del estado sirve como un filtro para pequeños errores que hacen que este método sea más robusto, el estimador de estado puede detectar los nodos con datos de

demanda inconsistentes. La estimación de estado busca obtener la mejor estimación posible, en el sentido de los mínimos cuadrados ponderados, donde su resolución proporciona los errores en las mediciones, las mismas que son normalizadas y ayuda a tener un indicio en primera instancia de que nodos presentan inconsistencias. Sin embargo, aunque los errores normalizados suministren los datos inconsistentes en los nodos, no es posible discernir precisamente qué nodos presentan anomalías en su sistema de medición de manera efectivamente. Por lo que adicional se puede hacer un análisis de error grueso a los resultados del estimador de estado y detectar consumos fraudulentos [13].

En este trabajo se propone desarrollar la combinación de dos metodologías para la Identificación de Pérdidas no técnicas de Energía en los sistemas de distribución, para lo cual se clasificará las curvas diarias de consumo de energía de los clientes dividiéndolas en curvas características con respecto a los días de la semana, fines de semana y feriados, para luego emplear tres índices de evaluación de agrupamiento de los clientes, segmentando de manera adecuada a la data histórica de dos años y finalmente entrenar las redes neuronales diseñadas para clasificar los clientes fraudulentos y los que no tienen esta condición; esperando obtener una alta precisión en el reconocimiento de clientes que subregistran energía.

1.5.2. Pérdidas no técnicas en Distribución

Las pérdidas no técnicas es la diferencia entre las pérdidas globales y pérdidas técnicas. Las pérdidas no técnicas, a su vez, son causadas principalmente por el robo de energía, además de otros problemas, tales como, errores de lectura del medidor, que pueden ser causados intencionalmente o no; consumidores clandestinos, no registrados en el sistema del distribuidor y fallos en la actualización de bases de datos y registros.

1.5.3. Pérdidas No Técnicas: Clasificación

Las pérdidas no técnicas pueden ocurrir de 4 maneras diferentes, con intensidades que varían de acuerdo con factores que van desde lo cultural a lo técnico [5].

1.5.3.1. Hurto

Diferenciado por el desvío directo de la energía de las redes eléctricas por el consumidor encubierto, por lo tanto, la energía utilizada por este consumidor no se tiene en cuenta. Estas conexiones ilegales generalmente se realizan en el alimentador de bajo voltaje o en el transformador de servicio, en cuyo caso las conexiones están expuestas, lo que hace posible la identificación visual [5]. En Ecuador, este tipo de irregularidad a menudo ocurre en áreas de riesgo, lo que hace que la inspección y la lucha contra el robo de energía sean más complejas, y puede haber problemas de seguridad para los técnicos del distribuidor.

1.5.3.2. Estafa

El fraude se da regularmente cuando el consumidor está registrado por la distribuidora, pero realiza cambios en el conexionado que afectan las marcas de su medidor de energía o utiliza campos magnéticos (imanes) que realicen interferencia cerca del medidor para evitar la rotación del disco. Estos cambios pueden ser rústicos y causar daños intencionados en el medidor o tener cargas en paralelo con el medidor, lo que hace que registre un consumo menor que el real; recalcando que actualmente la mayoría de los medidores son electrónicos, de radio frecuencia, lo que hace que este tipo de fraudes sean menos probables.

1.5.3.3. Dificultades de Facturación

Pueden ocurrir debido a varios factores, pero consisten principalmente en errores de lectura del medidor. Estos errores de facturación pueden ser involuntarios debido al mal estado o al posicionamiento de los medidores, lo que dificulta la lectura o incluso a los problemas de administración y sistemas de la empresa que causan errores de facturación. Sin embargo, hay casos de favoritismo intencional por parte del empleado del distribuidor, por parentesco o amistad con el consumidor y, en casos donde la factura de energía es alta, problemas como el pago de sobornos a los empleados para que registren valores menores que al valor real [6].

1.5.3.4. Dificultades en los Cobros

Estos son los casos que corresponden al impago de facturas por parte de los consumidores. Este tipo de pérdida es conocida por la compañía, pero aún trae grandes pérdidas al distribuidor [6].

1.5.4. Detección de Pérdidas: Métodos

Las técnicas para el cálculo de pérdidas no técnicas de energía eléctrica se pueden dividir en dos clases principales: indirecta y directa. La técnica indirecta trata de estimar las pérdidas técnicas y obtener las pérdidas no técnicas a través de la diferencia entre las pérdidas totales y las pérdidas técnicas estimadas, mientras que las técnicas directas buscan detectar directamente las pérdidas no técnicas [5].

Los métodos directos pueden utilizar los datos históricos para definir las normas de consumo y detectar comportamientos anómalos, o el uso de datos en tiempo real para detectar la ocurrencia de pérdidas no técnicas en el sistema [5].

1.5.4.1. Método Indirecto

La estimación de pérdidas técnicas admite, además de obtener pérdidas no técnicas por la diferencia entre pérdidas globales y pérdidas técnicas, tener la eficiencia del sistema y detectar las necesidades de mejoras. La precisión de la estimación de pérdidas técnicas depende de la información útil de la red y, por lo tanto, puede suceder en situaciones de:

- ***Alto grado de conocimiento:*** en este caso hay mucha información sobre la red y la carga que es posible calcular las pérdidas técnicas con buena precisión.
- ***Bajo nivel de conocimiento:*** cuando no hay mucha información, las medidas disponibles se utilizan para la estimación, generalmente subestación y algunos dispositivos de red, además de las comparaciones con sistemas similares.
- ***Caso híbrido:*** cuando existe un alto grado de conocimiento de solo una parte de la red, se utilizan otras metodologías.

1.5.4.2. Método Directo

Los métodos directos pueden usar datos históricos para definir patrones de consumo y detectar comportamientos anómalos, o usar datos en tiempo real o casi en tiempo real para detectar la ocurrencia de pérdidas no técnicas en el sistema.

1.5.5. Métodos que emplean los Datos Históricos

Los métodos que usan datos históricos tienen una estructura similar, divididos en tres etapas: preparación, clasificación e investigación.

La minería de datos o exploración de datos es el campo de la estadística y las ciencias informáticas donde su proceso intenta descubrir padrones en grandes bases de datos, lo que ha permitido el desarrollo de diversos algoritmos que abarcan distintas técnicas de aprendizaje: supervisados y no supervisados, las cuales incluyen tareas de agrupamiento, clasificación y regresión. A partir de la información almacenada en base de datos se puede obtener patrones de consumo, que reflejan el comportamiento de los consumidores [7]. Los patrones se pueden agrupar para crear perfiles o segmentos de consumo que podrían utilizarse en el pronóstico y control de carga, también nos permite identificar de manera automática irregularidades en el consumo y la detección de las Pérdidas no Técnicas.

1.5.6. Proceso de Minería de Datos

Se define como un conjunto de técnicas que permiten explorar de manera automática o semiautomática grandes bases de datos, es la fase más importante del proceso de descubrimiento de conocimiento en base de datos, más conocido como (Knowledge Discovery in Databases), el proceso comprende cuatro etapas recopilación; preparación; Data Mining; interpretación y evaluación [8].

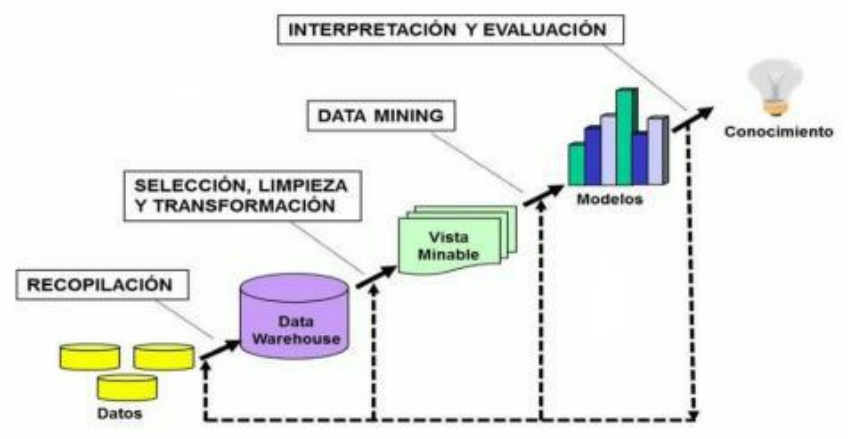


Figura 1.1: Secuencia del Data Mining [8]

1. Recopilación de datos: Es la fase donde se toman los datos que se desean analizar procedentes de diferentes fuentes y se integran en un mismo y único repositorio de datos, denominado almacén de datos, más conocido como data warehouse. Es una tecnología diseñada especialmente para organizar grandes volúmenes de datos de procedencia generalmente estructurada [8].

2. Preparación de los datos: toma los datos almacenados en data warehouse y los transforma en un subconjunto de datos consistente para extraer el conocimiento implícito en ellos [8].

Esta etapa consta de varias partes:

- *Selección:* el objetivo es seleccionar el subconjunto de datos, diferenciando las variables objetivo (variables a predecir, calcular o inferir) y las variables independientes (las cuales sirven para ayudar en el proceso de las variables objetivo), para esto se aplican técnicas de muestreo adecuadas [8].
- *Limpieza de datos:* se tratan todos los datos que puedan influir en un análisis inexacto y resultados incorrectos. Identificación y tratamiento de valores atípicos (outiers), datos erróneos e irrelevantes, la existencia de datos incompletos. Los valores atípicos son observaciones aisladas cuyo comportamiento se diferencia claramente del comportamiento medio del resto de las observaciones [8].
- *Transformación:* preparación de los datos de entrada acorde a la técnica de minería de datos que se aplicará, en esta fase se aplican técnicas de transformación como son: reducción o aumento de la dimensión, desratización o numeración, normalización de rango, escalado [8].

3. Minería de datos: obtención de un modelo en el que están implícitos los patrones de comportamiento observados. Se distinguen entre técnicas predictivas, las cuales se utilizan para predecir el valor desconocido de uno o varios atributos para uno o varios registros y técnicas descriptivas, generan modelos que, de alguna forma, describen los datos [8].

4. Interpretación y evaluación: consiste en evaluar la calidad de los modelos y realizar una interpretación de estos para obtener el conocimiento buscado, para validar los resultados se utilizan intervalos de confianza, Bootstrap, análisis ROC y evaluación de modelos [8].

1.5.7. Análisis de Conglomerados

Es el método más popular en el aprendizaje no supervisado. Este consiste en dividir una población heterogenia de objetos en grupos homogéneos, de tal forma que los objetos de cada grupo sean similares o guarden una relación entre ellos y diferentes al resto de objetos de otros grupos. La medida de similitud está basada en los atributos que describen a los objetos. Para poder establecer los diferentes grupos de objetos similares entre sí, es necesario elegir una función de distancia y calcular con ella la distancia entre los individuos. Las medidas de distancia más empleadas son: distancia Euclidiana, distancia Manhattan, distancia Minkowski [9].

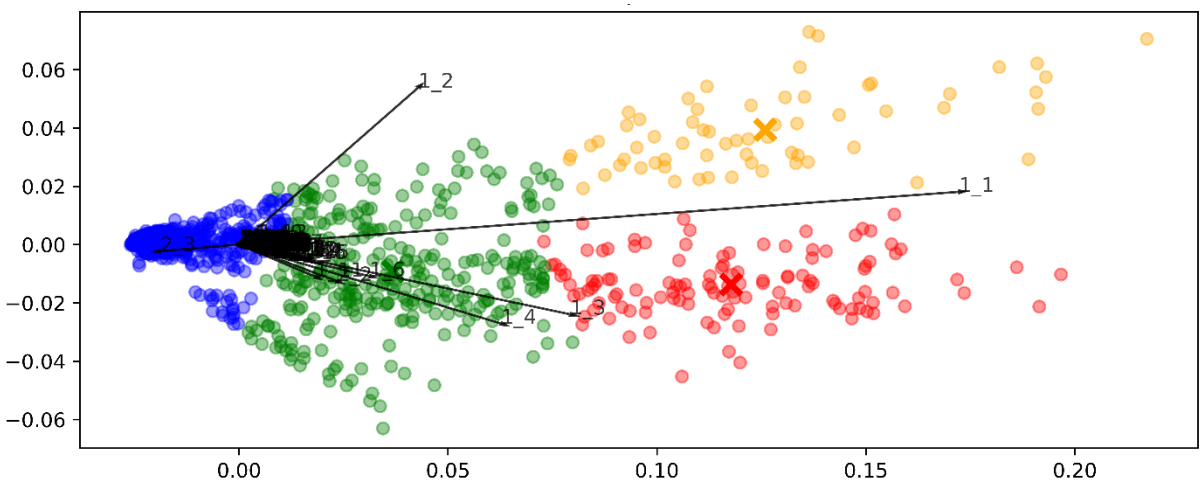


Figura 1.2: Agrupamiento o Clusterización

1.5.7.1. K vecinos más cercanos

Se desarrolla en torno a una modificación de la técnica k-Means, motivado por los comentarios y conclusiones de [10], [11], [12], y [13] que demuestran la potencial mejora de agrupamiento de series temporales basados en la métrica de deformación dinámica del tiempo (Dynamic Time Warping, DTW).

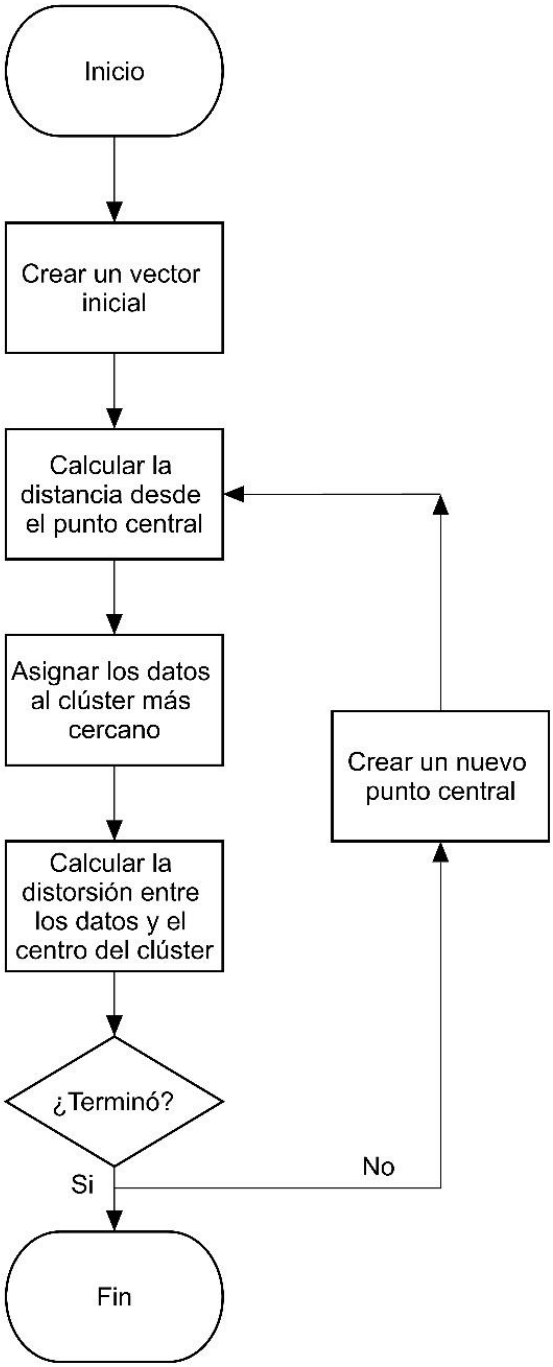


Figura 1.3: Proceso de Clusterización con KM [14]

El KM clásico con distancia euclidiana agrupa un conjunto de datos de $x(m)$ * (m = 1, …, M) muestras en k = 1, …, K agrupamientos mediante un proceso iterativo. Una primera suposición es hecha por K agrupamientos con centros $c(k)$. Los K centros clasifican las muestras en el sentido de que la muestra $x(m)$ pertenezca al agrupamiento k si la distancia $\|x(m) - c(k)\|$ es la mínima para todas las K distancias. Los centros estimados se utilizan para clasificar las muestras en conglomerados y se recalculan sus valores $c(k)$ del tipo $c(n)$ * (n = 1, …, N). El procedimiento se repite hasta la estabilización de los centros de los grupos como se muestra en la Figura 1.3. El número óptimo de agrupaciones no se conoce a priori y la calidad de la agrupación depende del valor de K.

Mientras que la modificación del KM de [10] permite que la agrupación de series temporales mediante DTW sea superior a la agrupación clásica de métrica euclidiana debido a su capacidad de capturar distorsiones temporales. DTW se formula como un problema de optimización dado por la Ecuación 1.1.

$$DTW(x, c) = \min(\pi) \sqrt{\sum_{(i,j)\in\pi} d(x_i, c_j)^2}$$

Ecuación 1.1: Minimización de la Distancia DTW

Dónde $\pi = [\pi_0, \ldots, \pi_k]$ es un camino que cumple las siguientes propiedades:

- Es una lista de pares de índices $\pi_k = (i_k, j_k)$ con $0 \leq i_k \leq m$ y $0 \leq j_k \leq n$
- $\pi_0 = (0, 0)$ y $\pi_K = (m-1, n-1)$
- Para todos $k > 0$, $\pi_k = (i_k, j_k)$ está relacionado con $\pi_{k-1} = (i_{k-1}, j_{k-1})$ como sigue:
    - $i_{k-1} \leq i_k \leq i_{k-1} + 1$
    - $j_{k-1} \leq j_k \leq j_{k-1} + 1$

En la Figura 1.4 (b), se puede ver la ruta en color azul para un par de series temporales que pueden ser $x^{(m)}$ y $c^{(n)}$ de modo que la distancia euclidiana entre series temporales alineadas es mínima, como se ve en Figura 1.4 (a).

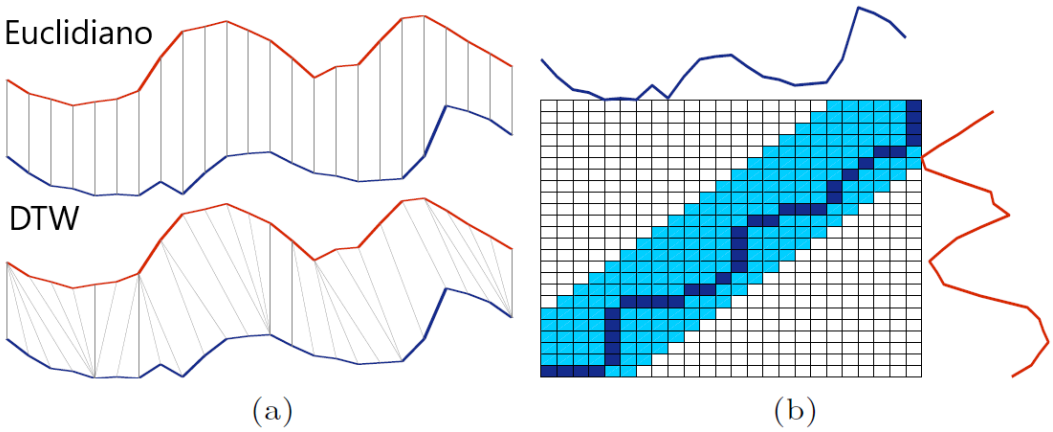


Figura 1.4: Cálculo de similitud: (a) alineación de series temporales Euclidiano (arriba) y DTW (abajo), (b) matriz de similitud cruzada calculada con DTW [12].

1.5.7.2. Índices de Validación

Uno de los mayores inconvenientes del algoritmo KM es que el número de grupos debe ser especificado de antemano, y si no se posee suposiciones válidas sobre la cantidad de conglomerados, el agrupamiento puede ser difícil de abordar. Los indicadores de validez son formas para cuantificar la calidad del agrupamiento. Existen dos tipos principales de indicadores: internos y externos. Los índices internos se calculan únicamente mediante la representación interna de los resultados de la agrupación, mientras que los índices externos comparan las agrupaciones generadas con información de agrupación externa, es decir etiquetas. El trabajo se enfoca en el interno ya que es más útil en la aplicación real. Puesto que, si ya tenemos una buena información de agrupación externa, ya no necesitamos hacer la agrupación [15].

Se experimentará con dos índices internos: la suma de cuadrados dentro del clúster (Within Cluster Sum of Saquéis - WCSS), y el coeficiente de silueta (Silhouette Coefficient - SC).

La Suma de Cuadrados dentro del clúster (WCSS)

La función objetiva implícita en KM mide la suma de las distancias de todos los puntos dentro de un clúster respecto al centroide del clúster, denominado WCSS. Esto se calcula como $\|x^{(m)} - c^{(k)}\|$, donde un conjunto de datos de $x^{(m)}$ * (m = 1,…,M) muestras puede tener k=1,…,K agrupamientos. Los centros de los agrupamientos son $c^{(k)}$. Por definición, esto

está orientado a maximizar el número de clústeres y, en casos límite, cada punto de datos se convierte en su propio centroide de clúster [16].

El Coeficiente de Silueta (SC)

Es otra medida de calidad de la agrupación, y se aplica a cualquier agrupación, no solo a KM. El índice de silueta para un patrón individual p en el conjunto de datos está definido por la Ecuación 1.2.

$$SC(p) = \frac{distinción(p) - consistencia(p)}{max\{distinción(p) - consistencia(p)\}}$$

Ecuación 1.2: Coeficiente de Silueta

La consistencia es la distancia promedio entre *p* y todos los demás patrones en el mismo grupo. La distinción es la distancia promedio entre *p* y todos los patrones restantes que no están en el mismo grupo. El rango de SC es [-1;1], donde "1" es la mejor agrupación y "-1" la peor agrupación. Mientras que WCSS es equiparable para los mismos datos con diferentes *k*, su número no es comparable entre diferentes soluciones de agrupación en diferentes datos y, por lo tanto, no tiene un umbral absoluto. Por otro lado, el coeficiente de silueta tiene un rango fijo y, por ende, se puede usar como métrica general para comparar la calidad de la agrupación, independientemente de los datos o la cantidad de grupos [16].

1.5.8. Redes Neuronales

Originalmente, las redes neuronales fueron propuestas teóricamente por [26], diseñando el primer modelo de redes neuronales (NN). Es así como se encuentra basado en modelos y algoritmos matemáticos, que, a causa de la fecha, no pudo ser probado por las carencias en recursos computacionales de la época.

Las primeras definiciones de esta metodología surgen a partir de la similitud que estas poseen con las redes neuronales biológicas [27], como se observa en la Figura 1.5; teniendo en cuenta que el cerebro humano es aún más complejo, dado que hasta la fecha no se conoce en su totalidad sus funciones. No obstante, se asemejan, según el autor, en 6 características principales: aprendizaje, adaptación, generalización, paralelismo masivo,

robustez, almacenamiento asociativo de información y procesamiento de información espaciotemporal.

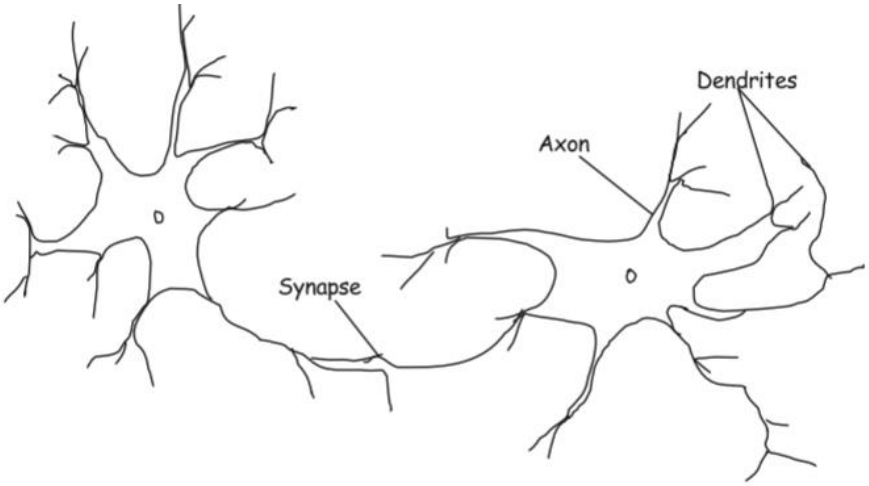


Figura 1.5: Red Neuronal Biológica [27]

Una neurona genérica consiste en una capa de entrada (*input Layer*), capas ocultas y una final, que tiene como característica, neuronas de salida (*output Layers*) [28], tal como se ve en la Figura 1.6.

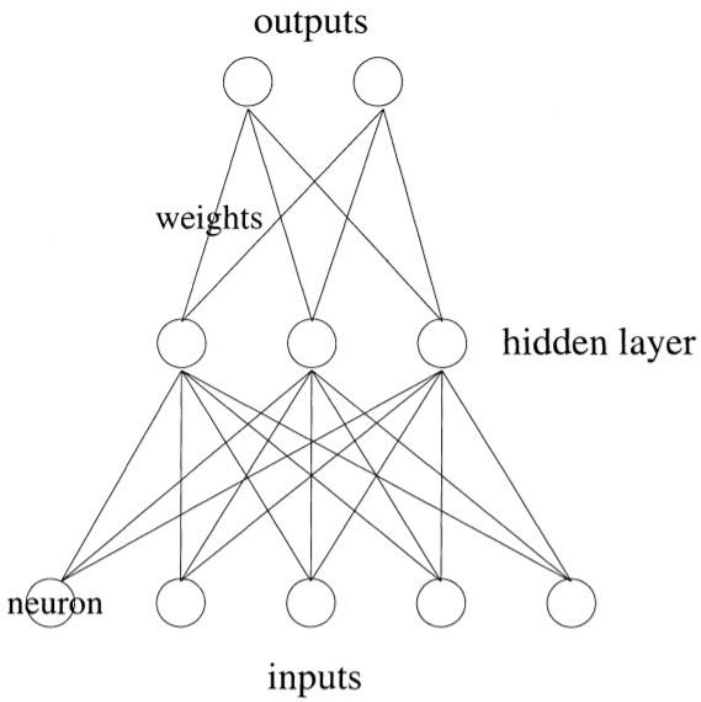


Figura 1.6: Arquitectura de una red neuronal [28]

Según [29], existen tres partes esenciales para formar la arquitectura de una neurona: neuronas, conexiones con sus pesos, parámetros y sesgos. La primera está construida sobre la base de funciones que contienen pesos y sesgos a la espera de la entrada de datos, para así, realizar los cálculos determinados. Luego, pasan por una función de activación para filtrar los datos a un rango predeterminado. Los pesos, en términos generales, son valores que la red aprende para generalizar un problema. El sesgo corresponde a valores que representan lo que la red supone que se deberían sumar a la multiplicación de los pesos con los datos, aprendiendo a detectar los sesgos óptimos [30].

1.5.8.1. Redes Neuronales Densamente Conectadas

En una red neuronal artificial la unidad análoga a la neurona biológica es referida como un "nodo" o "elemento de procesamiento". Un nodo tiene muchas entradas (dendritas) y combina, usualmente a través de una suma, los valores de estas entradas. El resultado es un nivel de actividad interna para el nodo. Las entradas combinadas son luego modificadas por una función de transferencia. Esta función de transferencia puede ser de tipo umbral lo que hará, que sólo pase información si el nivel de actividad combinado llega a un cierto nivel, o puede ser una función continua de la combinación de las entradas. El valor salida de la función de transferencia es generalmente pasado directamente hacia la ruta de salida del nodo.

La ruta de salida de un nodo puede ser conectada a entradas de otros nodos por medio de ponderaciones que corresponden (análogamente) a la resistencia sináptica de las conexiones neuronales. Como cada conexión posee una correspondiente ponderación o peso, las señales de las líneas de entrada hacia un nodo son modificadas por estos pesos previamente antes de ser sumadas. Es decir, la función de suma es una sumatoria ponderada. En sí mismo, este modelo simplificado de una neurona no es muy interesante; los efectos interesantes resultan de las maneras en que las neuronas sean interconectadas.

Una red neuronal consiste en muchos nodos unidos o conectados de la manera en que se muestra en la Figura 1.7, donde los nodos son usualmente organizados en grupos llamados "capas". Una red típica consiste en una secuencia de capas con total o aleatorias conexiones entre capas sucesivas.

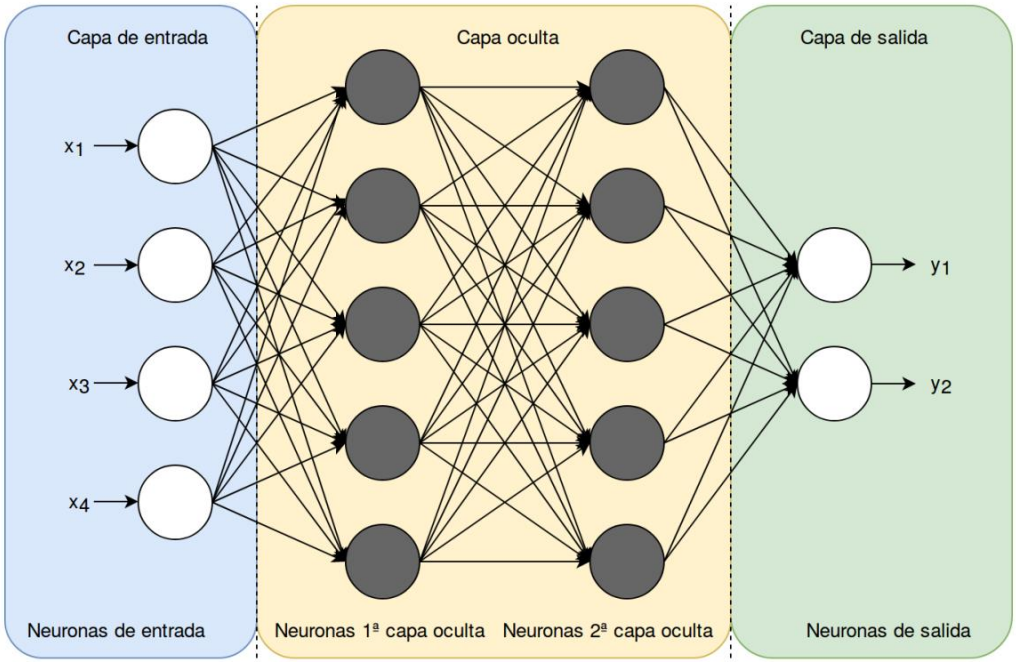


Figura 1.7: Red Neuronal Densamente Conectada [31]

Existen usualmente dos capas con conexiones hacia el mundo exterior: un búfer de entrada donde los datos se le presentan a la red, y un búfer de salida que contiene la respuesta de la red a una entrada dada. Las capas intermedias se las denomina "capas ocultas" [31].

1.5.8.1.1. Perceptrón [32]

Un tipo de neurona artificial muy conocida es el perceptrón. Los perceptrones fueron desarrollados entre 1950 y 1960 por el científico Frank Rosenblatt, inspirado por los trabajos anteriores de Warren McCulloch y Walter Pitts.

La Figura 1.8 muestra el esquema de un perceptrón. Éste toma n entradas binarias y produce una única salida y, también binaria.

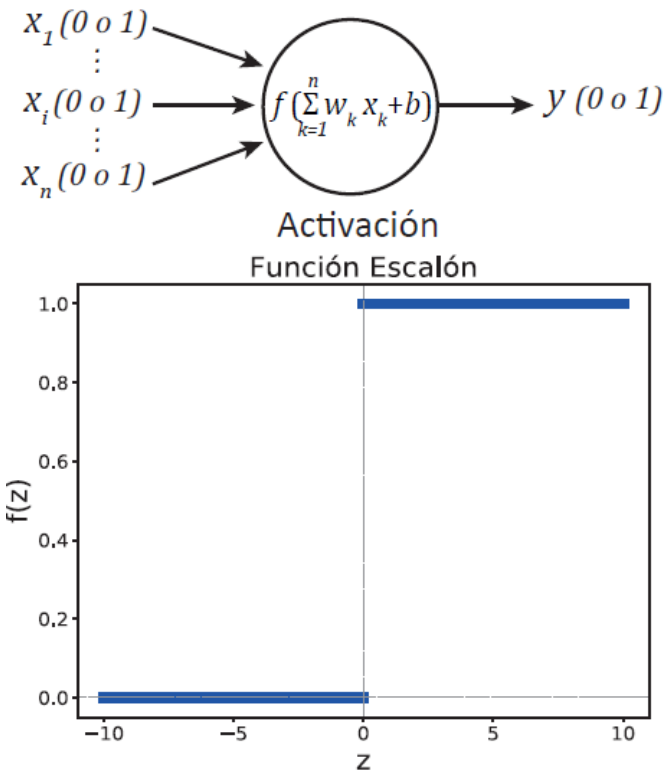


Figura 1.8: Esquema del perceptrón (arriba) y su función de activación (abajo).

Matemáticamente un perceptrón f se define de la siguiente manera:

$$f(x,w,b) = \begin{cases} 0 \ si \ \sum_{k=1}^{n} w_k * x_k \ \leq \ b' \\ 1 \ si \ \sum_{k=1}^{n} w_k * x_k \ > \ b' \end{cases}$$

Ecuación 1.3: Definición de Perceptrón

Donde x y w son vectores cuyas componentes son las entradas $\{x_1, \ldots, x_n\}$ y los pesos $\{w_1, \ldots, w_n\}$ respectivamente, y b' es el valor umbral (constante real).

Es posible reescribir la Ecuación 1.3, sabiendo que $\sum_{k=1}^{n} w_k * x_k = w * x$, sumando $-b'$ a ambos lados de las desigualdades y definiendo $b \equiv -b'$ como el sesgo del perceptrón:

$$f(x,w,b) = \begin{cases} 0 \ si \ w * x + b \ \leq 0 \\ 1 \ si \ w * x + b \ > 0 \end{cases}$$

Ecuación 1.4: Simplificación de Perceptrón

El sesgo b suele pensarse como una medida de cuan fácil es hacer que el perceptrón produzca como salida un 1 (neurona activada) o un 0 (neurona en reposo). Por ejemplo, para un perceptrón con un sesgo muy grande y positivo, es extremadamente fácil que produzca un 1. En cambio, si el sesgo es grande y negativo, entonces es difícil que el perceptrón produzca aquel valor como salida.

Por último, vale la pena mencionar que usualmente al perceptrón se lo puede considerar como un dispositivo de toma decisiones en base a evidencia empírica. De esta manera, al variar los valores de los parámetros (pesos y sesgo) se pueden obtener diferentes modelos de toma de decisiones. Asimismo, es plausible que si se considera una red de perceptrones se puedan tomar decisiones de mayor complejidad y más sutiles que con un solo perceptrón.

1.5.8.1.2. Perceptrón Multicapa [32]

Como se mencionó anteriormente los modelos que se desprenden del Perceptrón se basan en los principios de corrección de error planteados por el algoritmo de la regla delta para entrenar a estos sistemas. En un principio el desarrollo del Perceptrón llevó a la generación de un nuevo tipo de red cuya modificación principal respecto a la estructura del Perceptrón se basa en el uso de varias capas de neuronas artificiales, en vez de usar una sola capa. Este hecho significativo no hubiera servido de nada sin el cambio de la función de activación de las neuronas artificiales pasando de una función no diferenciable como era la activación logística a una función diferenciable y no lineal como lo es la sigmoide.

El recurso de este tipo de función de activación introdujo un nuevo paradigma en el procesamiento de los sistemas neuronales permitiendo a las redes neuronales aprender las variaciones no lineales de los distintos tipos de ambientes, que, en su mayoría, presentan variaciones del tipo no lineal.

En este momento se puede entender la importancia de este suceso ya que la mayor parte del tiempo el flujo de datos en las redes de comunicaciones sucede aleatorio y discontinuo. Este tipo de características son precisamente las que el nuevo sistema neuronal nos permitirá asimilar. Este sistema neuronal considerado también una red neuronal se conoce en la literatura como Perceptrón Multicapa debido a que es parte del principio del Perceptrón simple. En la Figura 1.9 se muestra la arquitectura de esta red neuronal.

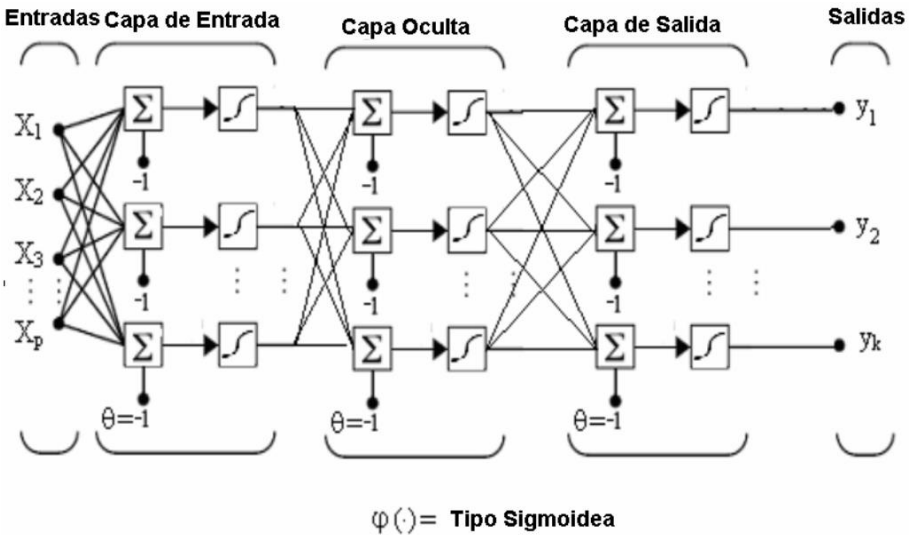


Figura 1.9: Arquitectura del Perceptrón Multicapa

Al observar la arquitectura del Perceptrón Multicapa se ve que las múltiples entradas conectadas en la primera capa son mapeadas en las salidas en función de las distintas capas de neuronas intermedias y de los parámetros libres de la red. Se puede semejar a una caja negra que realiza una operación sobre las entradas, produciendo un rango de salidas en función de los parámetros libres.

Se infiere de la arquitectura que el algoritmo de entrenamiento de una red con tales características deberá ser fraguado con el fin de que los cambios en los parámetros libres del error sean mínimos en las unidades básicas de la estructura; de manera que el conjunto de cambios produzca un error global que tienda al mínimo. Se buscará entonces, el límite en el cual la configuración de los parámetros libres produzca errores mínimos.

Tomando en cuenta este razonamiento, la evolución al Perceptrón Multicapa tuvo que basar su éxito en el diseño del algoritmo de entrenamiento que lograra minimizar el error al modificar adecuadamente los pesos y umbrales. La historia marco como primer paso, estudiar la forma de minimizar el error en una capa de neuronas lineales, que se conoce también como filtro lineal. El análisis de este tipo de red neuronal que posee elementos lineales nos permite deducir un algoritmo más complejo para entrenar a una red como la Perceptrón Multicapa que posee elementos no lineales.

1.5.8.2. Optimización de Hiperparámetros en Aprendizaje Profundo

Son los parámetros que modificamos manualmente en una red neuronal. Son muy importantes para lograr un buen desempeño y encontrar los valores óptimos harán que se adapten al problema propuesto. Los hiperparámetros hacen referencia a la serie de cálculos que se desarrollan en el análisis de redes neuronales profundas. En efecto, de entre todos los hiperparámetros, se detalla los que influyen en el entrenamiento de una red neuronal y cómo afectan en la optimización.

1.5.8.2.1. Número de Capas

Existen tres tipos de capas (L*ayers*) en una red neuronal, todas contienen una o más neuronas, como se mostró en la Figura1.7:

- ***Capa de entrada (Input Layer):*** Esta capa contiene neuronas que representan los datos que la red neuronal usara para entrenar. El número de neuronas de esta capa depende del número de características que tengan los datos.
- ***Capa oculta (Hiden Layer):*** Una red neuronal puede tener varias capas de este tipo, cada una de estas capas contiene neuronas, en una red neuronal tradicional cada una de las neuronas de una capa están conectadas con todas las neuronas de la siguiente capa.
- ***Capa de salida (Output Layer):*** Esta capa es la que se encarga de entregar los resultados, si estamos resolviendo un problema de clasificación esta capa tendrá un numero de neuronas igual al número de clases que existan en los datos. El resultado es una lista de probabilidades para cada clase.

1.5.8.2.2. Learning Rate [33]

La tasa de aprendizaje es la frecuencia con la que un algoritmo actualiza las estimaciones, este hiperparámetro está dado por formas algebraicas de como actualizar los pesos en las conexiones entre capas, como se muestra en la Figura 1.10, donde el valor de $n,$ está definido por la multiplicación del $n * {}^{\partial\, E_{total}}/_{\partial\, w}$, que es la importancia que le damos al error para actualizar cada peso, es decir, la rapidez o cómo de abruptos son los cambios en los pesos.

$$w_1^+ = w_1 - \eta \frac{\partial E_{total}}{\partial w_1} \qquad w_5^+ = w_5 - \eta \frac{\partial E_{total}}{\partial w_5}$$

$$w_2^+ = w_2 - \eta \frac{\partial E_{total}}{\partial w_2} \qquad w_6^+ = w_6 - \eta \frac{\partial E_{total}}{\partial w_6}$$

$$w_3^+ = w_3 - \eta \frac{\partial E_{total}}{\partial w_3} \qquad w_7^+ = w_7 - \eta \frac{\partial E_{total}}{\partial w_7}$$

$$w_4^+ = w_4 - \eta \frac{\partial E_{total}}{\partial w_4} \qquad w_8^+ = w_8 - \eta \frac{\partial E_{total}}{\partial w_8}$$

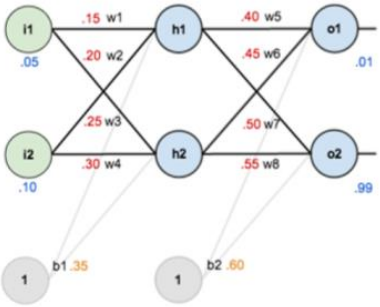


Figura 1.10: Definición de Learning Rate [33]

Así, un n muy alto, hará que los cambios en los pesos sean muy grandes de una iteración a otra, lo que tiene el problema de que podemos llegar a saltarnos nuestro mínimo, definiéndolo mejor en la Figura 1.11

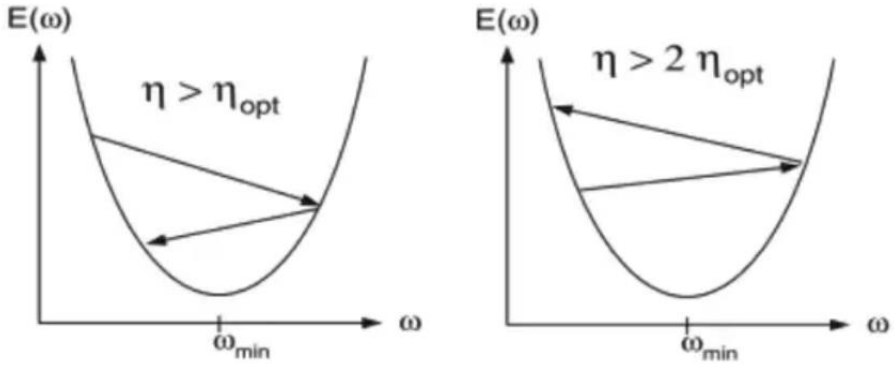


Figura 1.11: Cambios en el Learning Rate [33]

1.5.8.2.3. Batch Size [33]

Es el número de datos que tiene cada iteración de un ciclo (*Epoch*), esto es útil porqué la red neuronal actualiza los parámetros de peso y Bias más veces, también cuando se tienen grandes cantidades de datos se necesitan computadoras con más memoria y la red neuronal tarda más en ejecutar cada ciclo, si dividimos los ciclos en iteraciones con un numero de datos más pequeño ya no es necesario cargar todos los datos en la memoria al mismo tiempo y la red neuronal se entrena más rápido.

1.5.8.2.4. Epoch

Este es el número de veces que se ejecutaran los algoritmos de forwardpropagation y backpropagation. En cada ciclo (Epoch) todos los datos de entrenamiento pasan por la red neuronal para que esta aprenda sobre ellos. Si se especifica el parámetro Batch size cada ciclo (Epoch) tendrá más ejecuciones internas, estas ejecuciones se llaman iteraciones, en cada iteración se ejecutan los algoritmos de forwardpropagation y backpropagation, de esta manera la red neuronal actualiza más veces los parámetros de peso y Bias.

1.5.8.2.5. Lost Function

La función de perdidas, también conocida como función de costo, es la función que nos dice que tan buena es la red neuronal, un resultado alto indica que la red neuronal tiene un desempeño pobre y un resultado bajo indica que la red neuronal está haciendo un buen trabajo. Esta es la función que optimizamos o minimizamos cuando realizamos el backpropagation.

Existen varias funciones matemáticas que pueden usarse, la elección de una depende del problema que se esté resolviendo. Algunas de estas funciones son:

- ***Cross Entropy:*** Esta función se usa para problemas de clasificación.
- ***Mean Squared Error:*** Esta función se usa para problemas de regresión.

1.5.8.2.6. Optimizer

Es el algoritmo de optimización, el más famoso es el de gradiente descendiente y el más usado con el nombre Adam (*Adaptive Moment Estimation*) y es un algoritmo de optimización que combina las ventajas de los algoritmos RMSprop y Momentum para mejorar el proceso de aprendizaje de un modelo de red neuronal. Al igual que Momentum, Adam utiliza una estimación del momento y de la magnitud de los gradientes anteriores para actualizar los parámetros del modelo en cada iteración. Sin embargo, en lugar de utilizar una tasa de aprendizaje constante para todos los parámetros, Adam adapta la tasa de aprendizaje de cada parámetro individualmente en función de su estimación del momento y de la magnitud del gradiente. Esto permite que el modelo se ajuste de manera

más eficiente y efectiva a los datos de entrenamiento, lo que puede llevar a una mayor precisión de la predicción en comparación con otros métodos de optimización.

1.5.8.2.7. Funciones de Activación

Las funciones de activación se encuentran en cada neurona de una red neuronal y la utilidad más importante que tienen es indicar cuando una neurona se activa o se apaga, dependiendo de la función de activación que se use, la neurona tendrá ciertos límites, recordemos que primero se calcula la función de la neurona y el resultado de esta función se le pasa a la función de activación, esta última busca si los datos tienen los patrones que busca la neurona o no los tiene. Las funciones de activación más usadas son [32]:

- ***Logística Sigmoide***. $\delta : \mathbb{R} \rightarrow [0,1]$ tal que,

$$f(z) = \sigma(z) \equiv \frac{1}{1 + e^{-z}}$$

Ecuación 1.5: Función Sigmoide

- ***Tangente Hiperbólica (TanH):*** $\tanh : \mathbb{R} \rightarrow [1,1]$ tal que,

$$f(z) = \tanh(z) \equiv \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}} = 2\sigma(2z) - 1$$

Ecuación 1.6: Función Tangente Hiperbólica

- ***Softmax o Exponencial Normalizada:*** $\sum : \mathbb{R} \rightarrow [-1,1]^{k}$ tal que,

$$f(z) = \sum(z)$$

Ecuación 1.7: Función Softmax

Donde:

$$\sum(\boldsymbol{z})_{\boldsymbol{j}} \equiv \frac{e^{z_j}}{\sum_{i=1}^{k} e^{z_i}}$$

Ecuación 1.8: Sumatoria de función Softmax

para $\forall_j$ en $[1,k]$.

- ***Unidad Lineal Rectificada (ReLU):*** $\max: \mathbb{R} \rightarrow \mathbb{R}^{+}[1,1]$ tal que,

$$f(z) = \max(o, z) \equiv \begin{cases} z \ si \ z \ > 0 \\ 0 \ si \ z \ \leq 0 \end{cases}$$

Ecuación 1.9: Función ReLU

- ***Leaky ReLU:*** $f : \mathbb{R} \rightarrow \mathbb{R}$, tal que,

$$f(z) = \max(o, z) \equiv \begin{cases} z \quad si \quad z > 0 \\ 0.01z \ si \quad z \leq 0 \end{cases}$$

Ecuación 1.10: Función Leaky ReLU

- ***Unidad Lineal Rectificada Paramétrica (PReLU):*** $f : \mathbb{R} \rightarrow \mathbb{R}$, tal que,

$$f(z) = \begin{cases} z \ si \ z \ > a \\ 0 \ si \ z \ \leq a \end{cases}$$

Ecuación 1.11: Función PReLU

Si $a \ \leq 1$, $f(z) = \max(z, az)$, la cual se conoce como neurona o unidad **Maxout**.

- ***Thresholded ReLU:*** $f : \mathbb{R} \rightarrow \mathbb{R}^{+}$, tal que,

$$f(z) = \max(o, z) \equiv \begin{cases} z \ si \ z \ > a \\ 0 \ si \ z \ \leq a \end{cases}$$

Ecuación 1.12: Función Thresholded ReLU

Donde $a \ \epsilon \ \mathbb{R}$

Estas son las funciones de activación más comunes en las neuronas, en la Figura 1.12 se observan la forma de estas funciones.

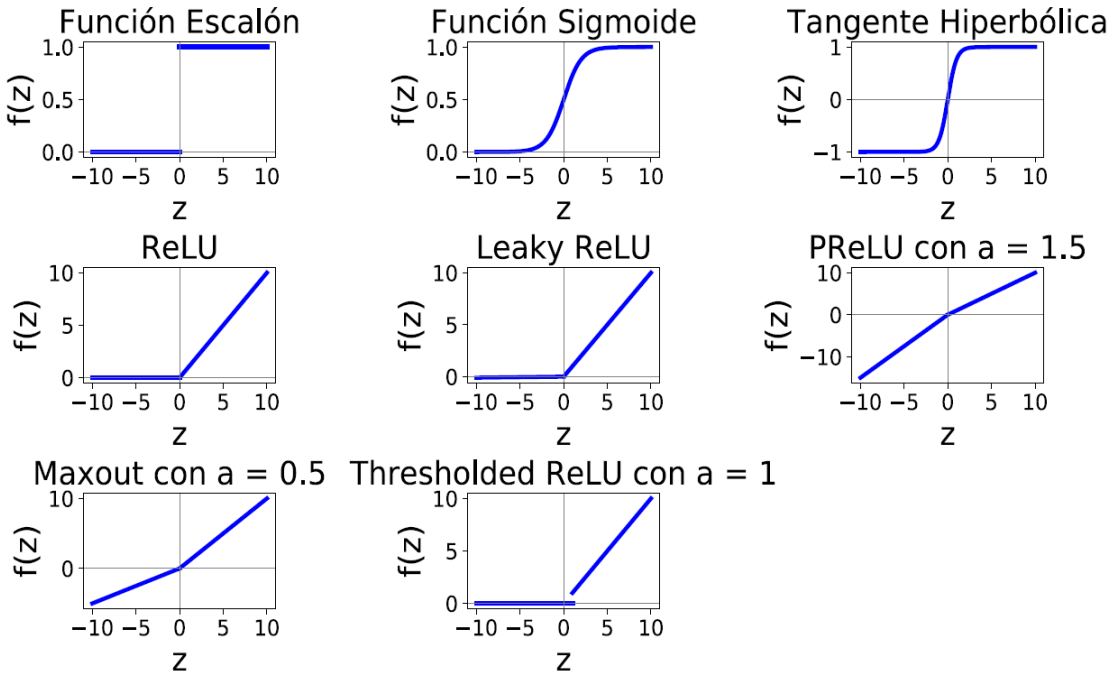


Figura 1.12: Gráficas de las distintas funciones de activación

1.5.9. Entornos de Trabajo

Se dará una breve explicación de los entornos de trabajo utilizados en el desarrollo de este trabajo.

1.5.9.1. EASY METERING

Es la plataforma que maneja la Empresa Eléctrica Ambato Regional Centro Norte S.A. que es la distribuidora quien nos proporcionará los datos históricos para ser analizados. La interfaz de usuario se la puede observar en la Figura 1.13, donde se puede observar que para ingresar se necesita de un usuario y una clave entregada por la distribuidora, solo los usuarios que tienen clave de acceso exclusivo pueden ingresar a la data histórica de todos los clientes.

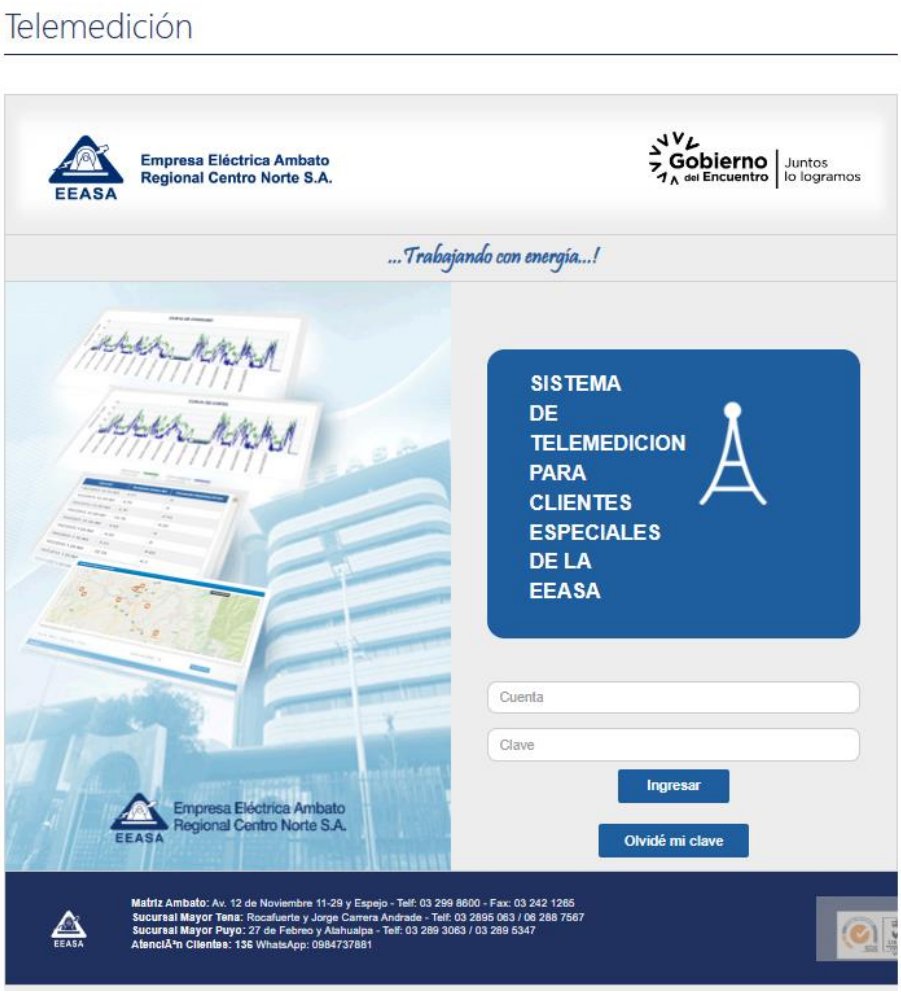


Figura 1.13: Interfaz de usuario de la plataforma EASY METERING

La plataforma está trabajando con la distribuidora desde el 2014 y es más conocida por el nombre completo “Sistema de Telemedición para clientes especiales Easymetering AMI Solutions”. Los clientes especiales que poseen la clave de ingreso único tienen acceso solo a los últimos 3 meses de la data propia.

1.5.9.2. Excel

Es una herramienta de Microsoft muy eficaz para obtener información con significado a partir de grandes cantidades de datos. También funciona muy bien con cálculos sencillos y para realizar el seguimiento de casi cualquier tipo de información. La clave para desbloquear todo este potencial es la cuadrícula de las celdas. Las celdas pueden contener

números, texto o fórmulas. Los datos se escriben en las celdas y se agrupan en filas y columnas. Esto permite sumar datos, ordenarlos y filtrarlos, ponerlos en tablas y crear gráficos muy visuales.

1.5.9.3. JupyterLab

Es una interfaz de usuario basada en la web para Proyecto Jupyter de Anaconda y está totalmente integrado en Adobe *Experience Platform*. Proporciona un entorno de desarrollo interactivo para que los científicos de datos trabajen con Jupyter Notebooks, código y datos.

JupyterLab es el último entorno de desarrollo interactivo basado en web para blocs de notas, código y datos. Su interfaz flexible permite a los usuarios configurar y organizar flujos de trabajo en ciencia de datos, computación científica, periodismo computacional y aprendizaje automático. Un diseño modular invita a las extensiones para ampliar y enriquecer la funcionalidad.

1.5.9.4. KNIME Analytics Platform

KNIME Analytics Platform es un software de código abierto pensado para gestionar datos. Intuitivo, abierto y actualizado. KNIME hace que la comprensión de los datos y el diseño de los flujos de trabajo y los componentes reutilizables sean accesibles para todos.

KNIME pertenece a una nueva generación de herramientas dominadas como Plataformas de Data Science y Machine Learning. Estas herramientas permiten a científicos de datos expertos, analistas o usuarios de negocio interactuar con sus datos y crear, desplegar y gestionar sus modelos de analítica avanzada. Las herramientas integran las funcionalidades principales para realizar proyectos de minería de datos: importación de datos, preparación de datos, exploración de datos, modelado, evaluación y despliegue. Dentro de estas herramientas KNIME se encuentra en el grupo de líderes del último diagrama de Gartner conocido también como la “navaja suiza” del mercado.

Como puntos fuertes para robustecer el uso de esta plataforma, podemos destacar:

- ***Facilidad de uso:*** a través de interfaz visual la programación de aplicaciones en KNIME es altamente intuitiva. Conectando visualmente nodos que encapsulan distintas funciones e integrando módulos automatizados de Machine Learning, Deep Learning y modelos preprogramados; facilita la analítica avanzada a los usuarios de negocio sin experiencia en este ámbito. Al mismo tiempo brinda un ecosistema óptimo para desarrolladores avanzados con la posibilidad de integrar programación en Python o R.

- ***Extensas funcionalidades:*** KNIME ofrece el ciclo completo de Data Mining. Con la posibilidad de conectarse a múltiples y heterogéneas fuentes de datos. pudiéndose unificar datos provenientes de distintas Bases de Datos, archivos y servicios web diversos como Azure, etc. con muy poco esfuerzo. Con una gran variedad de nodos para el preprocesamiento la herramienta ofrece las condiciones óptimas para la generación de procesos de ETL automatizada. Finalmente ofrece los principales algoritmos y métodos de evaluación para la generación de modelos potentes. Adicionalmente dispone de múltiples extensiones (Text Processing, Big Data con Spark y Hadoop, Deep Learning con TensorFlow y Keras y muchas más) que empoderan la herramienta aún más.

- ***Bajo costo de adquisición e implantación:*** KNIME cuenta con una versión gratuita “KNIME Analytics Platform” para el uso personal, así como una versión de pago “KNIME Server” para el uso en organizaciones que quieren llevar sus actividades de data Mining a un nuevo nivel.

En la Figura 1.14 se muestra el entorno de *KNIME Analytics Platform*, donde se observa las diferentes ventanas y entorno de trabajo. Al costado derecho tenemos la ayuda directa del desarrollador y la comunidad *KNIME Hub* y al mismo tiempo la pestaña *Description*, que se activa cada vez que seleccionamos un Nodo, desplegándose la descripción completa del nodo y como configurarlo, en el centro, el espacio de trabajo dónde se extraen los nodos y se realiza los diferentes diagramas de programación, en el costado izquierdo las ventanas *KNIME Explorer*, donde se guardan directamente los Grupos de trabajo y se los puede abrir con solo arrastrarlos al entorno de trabajo, la ventana *Workflow Coach*, que es la ventana donde nos aparecerán sugerencias de nodos mientras se desarrolla un trabajo, sugeridas a partir de las estadísticas del uso de los nodos por la comunidad KNIME; finalmente tenemos una de las ventanas más importantes del programa, el *Node Repository* que como su nombre lo indica, tiene todos los nodos para ser utilizados para un programador.

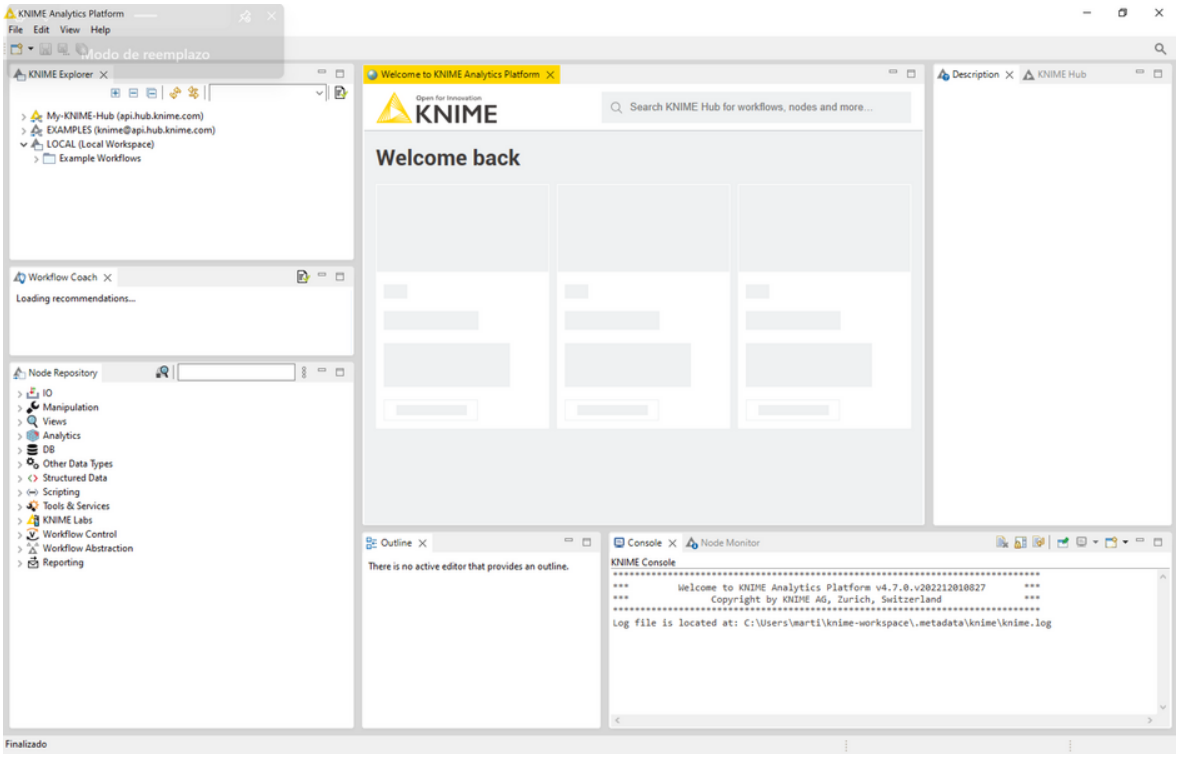


Figura 1.14: Entorno de KNIME Analytics Platform

2. METODOLOGÍA

2.1. Área de Concesión de los Clientes Especiales de la EEASA.

El área de concesión de la EEASA, se encierra a gran parte de la zona central del País en una superficie de aproximadamente 40.805 km² y aproximadamente 832.075 habitantes, como se puede observar en la Figura 2.1, que incluye a las provincias de Tungurahua y Pastaza, en su totalidad; los cantones: Palora, Huamboya y Pablo Sexto en la provincia de Morona Santiago y la parte sur de la provincia de Napo, que incluye su capital Tena y los cantones de Archidona y Carlos Julio Arosemena Tola [34].

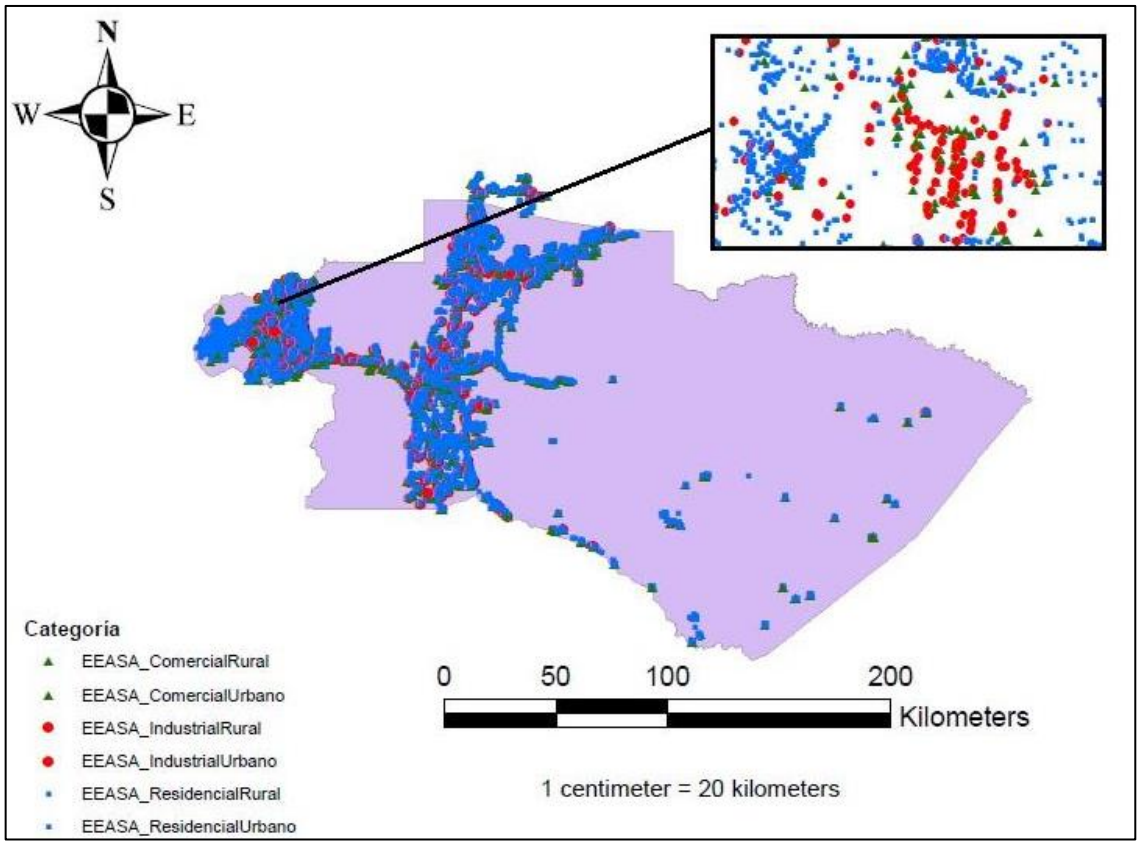
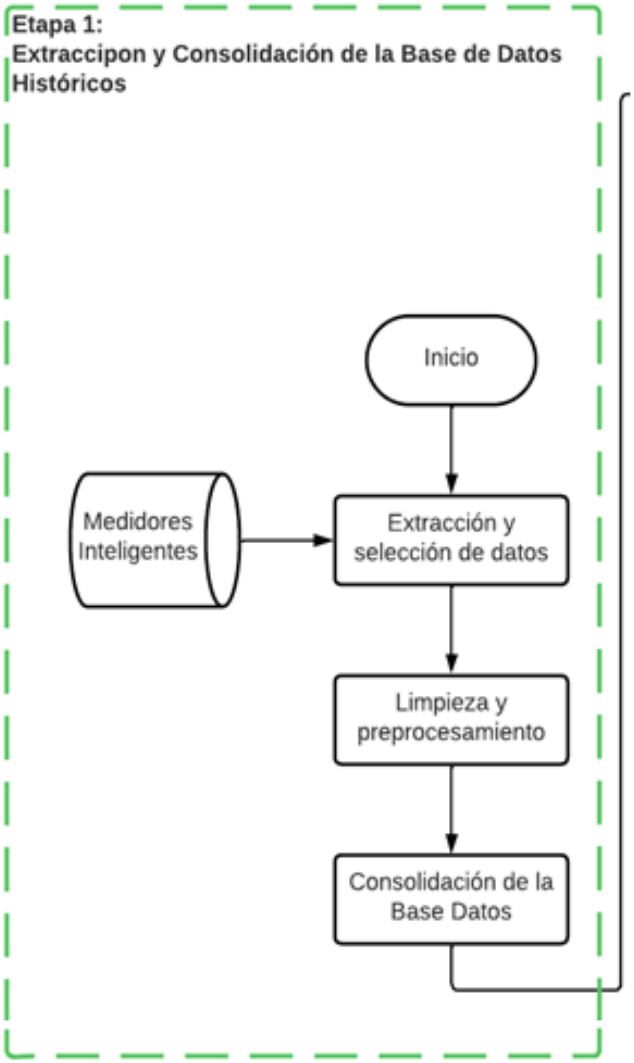


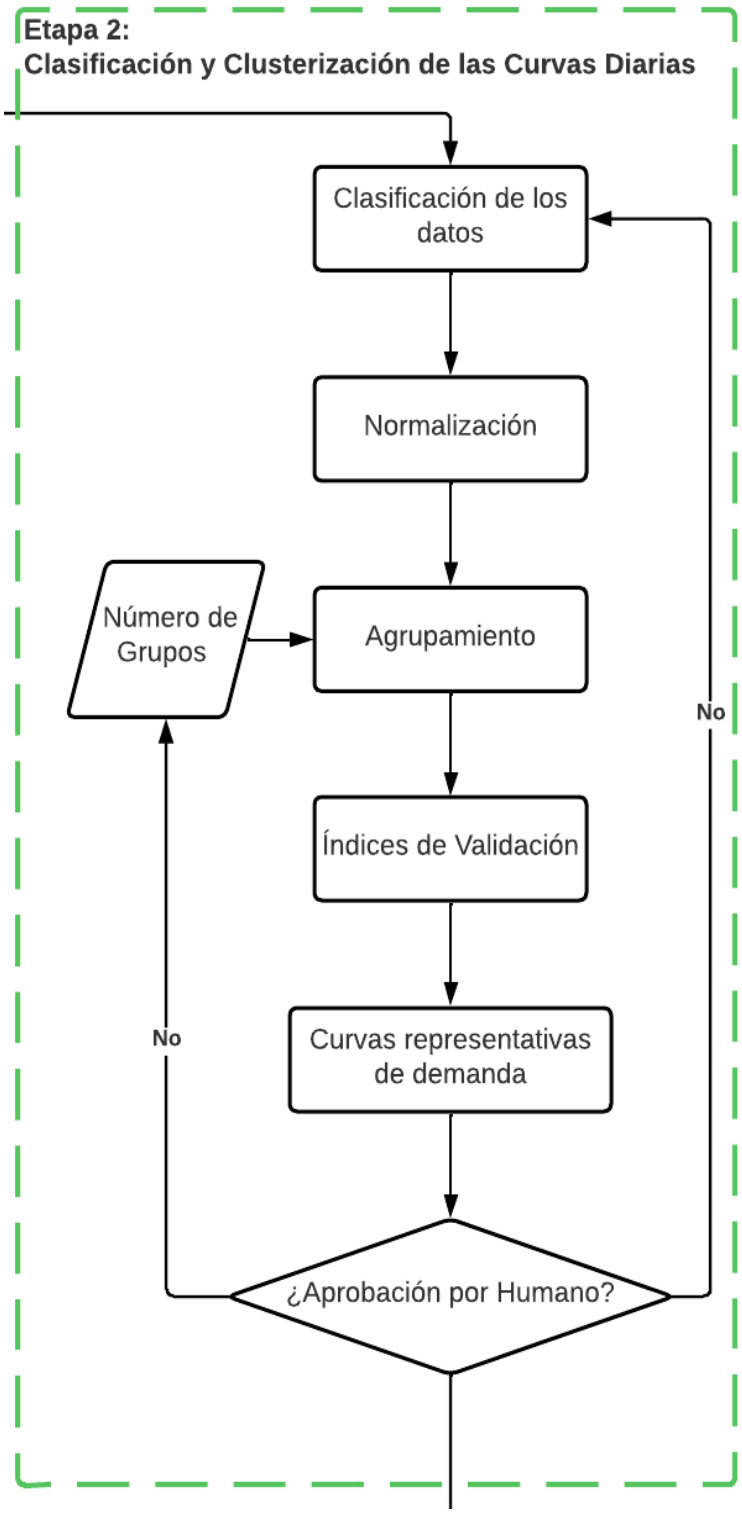
Figura 2.1: Área demográfica de los clientes especiales

A partir de los datos históricos de los clientes de telemedición industrial y comercial como se observa en la Figura 2.1 para los dos años de datos históricos, se conformó una matriz de datos, para el caso de estudio se tomó en cuentan todos los días.

2.2. Propuesta Metodológica

La metodología propuesta consiste en la evaluación de las curvas diarias históricas de dos años para la identificación de pérdidas no técnicas mediante el aprendizaje profundo de la red neuronal de clasificación. Este método nos ayudará a clasificar las curvas con anomalías y típicas con base en el aprendizaje de los datos históricos descargados de los clientes de la empresa distribuidora, identificando patrones considerados con subregistro de clientes que podrían ser sospechosos de fraude de energía. La metodología se resume en la Figura 2.2, donde se muestra tres etapas claras para el desarrollo del presente trabajo.




Etapa 2:
Clasificación y Clusterización de las Curvas Diarias
Clasificación de los datos
Normalización
Número de Grupos
Agrupamiento
Índices de Validación
Curvas representativas de demanda
¿Aprobación por Humano?
No
No

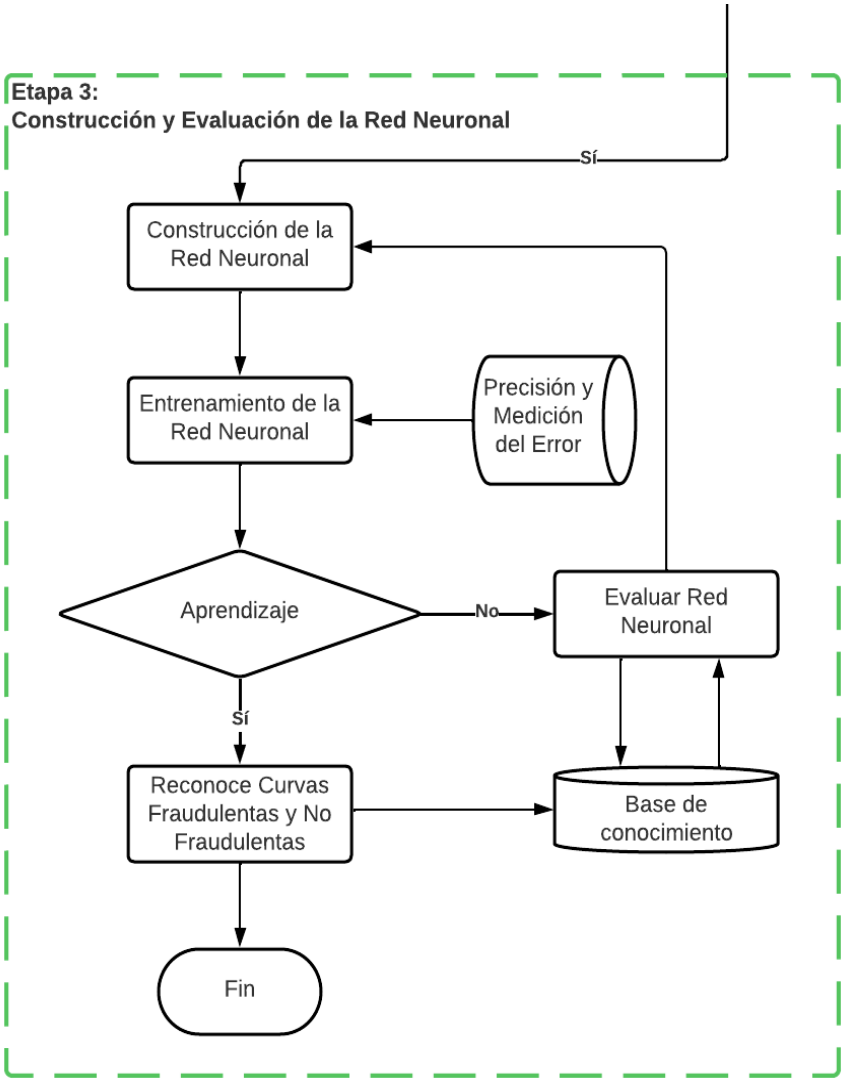


Figura 2.2: Diagrama de Flujo de la metodología propuesta

2.2.1. Extracción y selección de datos

De los datos históricos de los clientes especiales (industriales y comerciales) que poseen telemedición como se observa en la Figura 2.1 para el periodo de tiempo comprendido entre el 31 de mayo del 2020 al 31 de mayo del 2022 se conformó una matriz de datos, para lo cual se hizo una preselección de los clientes más representativos a sus características de consumo e históricos de medición, exponiendo las zonas de concesión que posee la distribuidora en la Tabla 2.1, excluyendo las zonas 6, 9, 13 y 14 por no contar con el proceso de telemedición o los datos históricos necesarios para el estudio, según el departamento de Sistemas de Medición de la EEASA. El valor de la columna de código nos ayudara a identificar la zona a la que pertenecen los clientes elegidos.

Tabla 2.1: Zonas de la concesión de la EEASA

Código	Descripción	Abreviatura
1	AMBATO	AMB
2	PELILEO	PEL
3	PILLARO	PIL
4	BAÑOS	BAÑ
5	PATATE	PAT
6	AGENCIAS	AGE
7	PASTAZA	PAS
8	PALORA	PAL
9	PUYO	PUY
10	QUERO	QUE
11	TENA	TENA
12	ARCHIDONA	ARCH
13	TISALEO	TIS
14	SUCUMBIOS	SUCUM

Para la selección de los clientes se obtiene una matriz de información en formato Excel como muestra la Figura 2.3 con detalles determinantes, como son las dos columnas finales que poseen las características de “Está en Telemedición” y el “Total Consumo KW”, que serán importantes en el momento de seleccionar los 100 de los 297 clientes que cumplen con las condiciones de tener datos históricos mayores a dos años y que no hayan sido motivo de cambio en el proceso de telemedición.

maag_codigo AGENC	maac_codigo CuentaDel Clier	rfg_fact_potencia FactorDePotencia	c_cliente NombreDelCliente	ma_medidores_mame Direccion	maac_telemetria EstáEnTelemedición?	TOTAL CONSUMO kW
2	271045	0.93457379	GARZON RAMOS DIANA GISELA	COTALO	N	
2	272238	0.87726274	OÑA JAMI JORGE	EL CARMEN-BENITEZ	N	
2	273085	0.93193794	SAILEMA MASAQUIZA EDISON JAVIER	SALASAKA CENTRO PELILEO	N	
3	45372	0.99998089	VALLE LEON CARLOS FABIAN	MIRABELES C SN Y RUMINAHUI PILLARO NUEVO	N	
3	89866	0.85442778	BARONA ROSALES FRANCISCO MIGUEL	TASINTEOPILLARO	N	
3	96201	0	CORPELEC DEL ECUADOR CELEC EP	PUCARA SAN JOSE DE POALOPILLARO	N	
3	163260	0.97862266	MOYA SEGUNDO	ROCAFUERTEATIPILLAHUAZOPILLARO	S	133577
3	161495	0.91571208	SANCHEZ JACOME MARTHA SUSANA	SECTOR ROCAFUERTE PILLARO	S	102978
3	111638	0.99999563	ARCOS ARCOS EDGAR OSWALDO	SUCRE Y NAVAREZPILLARO	N	
3	118497	0.99018722	GADM SANTIAGO DE PILLARO COLISEO DEPORTE	SUCRE Y RODRIGUEZ DE GUZMANPILLARO	N	
3	160597	0.9726604	AVALOS JAVIER	SECTOR ROCAFUERTE PILLARO	S	91920
3	132797	0.88030461	PROAÑO CHILUIZA MARCO VINICIO	LA ELEVACION PILLARO	S	89473
3	161860	0.94893711	CASTRO NARVAEZ NATALIA DEL LOURDES	MIRABELES Y LAS MANZANAS PILLARO NUEVO	S	78942
3	169382	0.90908245	JACOME JACOME CIRO ERNESTO	BARRIO ROCAFUERTE	S	66063
3	279187	0.99732445	MAYORGA GORDON KLEVER ESTUARDO	GUAPANTE GRANDE COCHA VERDE	S	51816
3	233071	0.92676571	SAN JUAN MERCADO	ROCAFUERTE Y FLORES PILLARO	S	49737
3	225793	0.99303409	TIENDAS INDUSTRIALES ASOCIADAS TIA S.A.	SUCRE ENTRE URBINA Y ROCAFUERTE	S	41411
3	167991	0.98961467	CASTRO NARVAEZ NATALIA DE LOURDES	CALLATE LA Y - PILLARO	S	39988
3	164287	0.95817621	VILLACRES CAMPAÑA RODRIGO MARCELO	SECTOR ROCAFUERTE PILLARO	S	39107
3	163788	0.9983273	RAMIREZ RAMIREZ MARIA MERCEDES	PAMPA MIA PILLARO	N	
3	232124	0.97209642	MAYORGA GORDON KLEVER ESTUARDO	HUAPANTE GRANDE COCHA VERDE PILLARO	S	38293
3	168980	0.89425499	RODRIGUEZ PAZMIÑO CARLOS GUILLERMO	ROCAFUERTE	S	35545
3	166551		ESTADIO SANTIAGO DE PILLARO	ROCAFUERTE Y HEROES DEL 41PILLARO	N	
3	166852	0.99994998	DIRECCION DISTRITAL DE SALUD 18D04 SAN PEDF	LAS UVILLAS SAN MIGUELITO PILLARO	N	
3	165292	0.99965529	RUIZ DIAZ JUAN EDUARDO	WILSON GOMEZ Y FLORIDA PILLARO NUEVO	S	32656
3	104759	0.99998628	CNT EP ...	BOLIVAR B253 CARLOS TAMAYOPILLARO	S	23181
3	168450	0.96898206	COMPANIA PROALISAN LIMITADA	HUAPANTE GRANDE PILLARO	N	
3	168642	0.99859914	GARCIA PAZMIÑO JOSE ADOLFO	PENILEO PILLARO	N	
3	99792	0.93979342	DIRECCION DISTRITAL DE SALUD N18D04 SAN PEI	LOS FRUTALES Y MANZANAS PILLARO NUEVO	S	22080
3	167984	0.95143152	DIRECCION DISTRITAL DE EDUCACION 18D05 SAN	HUAPANTE GRANDE PILLARO	S	15992
3	169516	0.99947615	CANAR SIZA WILLIAM MARCELO	LA FLORIDA Y LAS ROSAS - PILLARO	N	
3	198807	0.82340709	COMPAÑIA PROALISAN CIA LTDA	HUAPANTE GRANDE PILLARO	N	
3	200989	0.825897	ARCOS ARCOS FLAVIO EDWIN	EL CENSO SAN MIGUELITO PILLARO	N	
3	131378	0.99736563	RUIZ DIAZ JUAN EDUARDO	PENILEO PILLARO	S	14164
3	217146	0.97165546	AVICOLA MAYORGA	HUAPANTE GRANDE COCHA VERDE PILLARO	N	
3	219394	0.93484961	AMORES ARCOS WIDMO FABIAN	EMILIO MARIA TERAN PILLARO	N	
3	221826	0.75857981	NARANJO GARCES FAUSTO RODRIGO	QUILLAN ALEMANIA PILLARO	N	
3	222273	0.51101095	NARANJO GARCES FAUSTO RODRIGO	QUILLAN ALEMANIA PILLARO	N	
3	224928	0.99962937	LUZURIAGA ROBALINO LUIS ANTONIO	QUILLANPATA SAN MIGUELITO PILLARO	N	

Figura 2.3: Matriz de Clientes Especiales de la EEASA

Con el filtrado de los clientes que poseen telemedición, los clientes que cumplen con el medidor AMI por más de dos años y con el método de Muestra Estratificada realizada en Excel, se determina la cantidad de individuos de cada zona, que deben ser extraídos para conformar los 100 clientes que serán parte de este estudio, como se puede observar en la Tabla 2.2, donde especifica el número de clientes elegidos de los que se extraerán los datos históricos.

Tabla 2.2: Muestra Estratificada de los Clientes de la EEASA

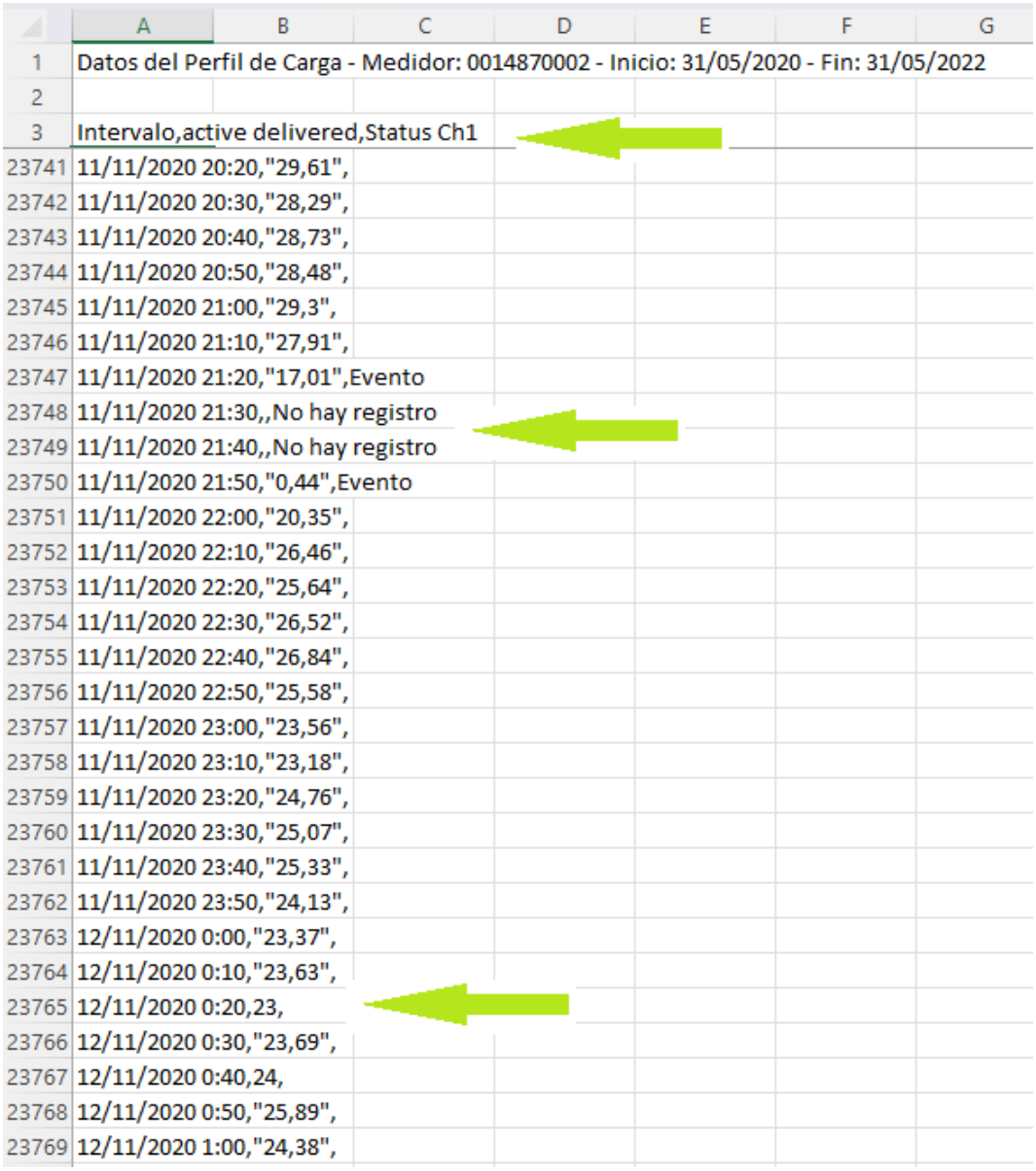
Zona	Descripción	Clientes	Muestra
1	AMBATO	142	60
2	PELILEO	43	10
3	PILLARO	13	3
4	BAÑOS	11	5
5	PATATE	8	2
7	PASTAZA	43	8
8	PALORA	2	2
10	QUERO	14	3
11	TENA	19	5
12	ARCHIDONA	2	2
TOTAL		**297**	**100**

La base de datos proveniente de los medidores inteligentes posee una gran dimensión y muy poca organización, ya que la plataforma “Sistemas de Telemedición para clientes especiales EASYmetering AMI Solutions”, que es el medio utilizado por la EEASA desde su desarrollo en el 2014; almacena registros relacionados con los clientes de electricidad como: voltaje [V], corriente [A], potencia [kW], energía activa [kWh] y energía reactiva [kVArh]. Siendo un gran inconveniente el gran número de conglomerados para el almacenamiento de todas estas variables, ya que no tienen una estructura jerárquica para la adquisición de los datos, lo cual genera un trabajo adicional en la consolidación de la data para cada cliente, en la Figura 2.4 se muestra como es la plataforma de Telemedición, de donde podemos elegir los parámetros deseados como el intervalo de tiempo, las unidades y el formato en que desea descargar la información generada, es importante mencionar que hay clientes a los que se deben descargar hasta 5 archivos de Excel para completar la data del cliente de los dos años propuestos.

Figura 2.4: Plataforma de Telemedición de la EEASA para Extracción de Datos

En este trabajo se seleccionan las variables de energía activa, y voltajes menores o iguales a 600 V, catalogados como bajo voltaje según el regulador [7]. En la Figura 2.5 se muestran los datos de uno de los archivos provenientes de los medidores inteligentes en formato CSV, donde podemos observar que en la primera posición se almacena la estampa de tiempo cada 10 minutos, en la segunda la energía activa entregada durante ese lapso y finalmente en algunos casos el estado del canal por el que se envía estos datos. Es importante mencionar que no todos los registros se encuentran completos, debido a fallos

que se provocan en la comunicación entre los AMI y la Base de Datos de la EEASA; además no todas las columnas contienen el mismo formato, como se ve en este caso en la fila 23764, que el valor de la energía entregada no se encuentra entre comillas.



	A	B	C	D	E	F	G
1	Datos del Perfil de Carga - Medidor: 0014870002 - Inicio: 31/05/2020 - Fin: 31/05/2022						
2							
3	Intervalo,active delivered,Status Ch1						
23741	11/11/2020 20:20,"29,61",						
23742	11/11/2020 20:30,"28,29",						
23743	11/11/2020 20:40,"28,73",						
23744	11/11/2020 20:50,"28,48",						
23745	11/11/2020 21:00,"29,3",						
23746	11/11/2020 21:10,"27,91",						
23747	11/11/2020 21:20,"17,01",Evento						
23748	11/11/2020 21:30,,No hay registro						
23749	11/11/2020 21:40,,No hay registro						
23750	11/11/2020 21:50,"0,44",Evento						
23751	11/11/2020 22:00,"20,35",						
23752	11/11/2020 22:10,"26,46",						
23753	11/11/2020 22:20,"25,64",						
23754	11/11/2020 22:30,"26,52",						
23755	11/11/2020 22:40,"26,84",						
23756	11/11/2020 22:50,"25,58",						
23757	11/11/2020 23:00,"23,56",						
23758	11/11/2020 23:10,"23,18",						
23759	11/11/2020 23:20,"24,76",						
23760	11/11/2020 23:30,"25,07",						
23761	11/11/2020 23:40,"25,33",						
23762	11/11/2020 23:50,"24,13",						
23763	12/11/2020 0:00,"23,37",						
23764	12/11/2020 0:10,"23,63",						
23765	12/11/2020 0:20,23,						
23766	12/11/2020 0:30,"23,69",						
23767	12/11/2020 0:40,24,						
23768	12/11/2020 0:50,"25,89",						
23769	12/11/2020 1:00,"24,38",						

Figura 2.5: Forma como se extraen los datos de la plataforma de Telemedición

2.2.2. Limpieza y depuración de los Datos

Para la limpieza de los datos el trabajo es minucioso, cliente por cliente, ya que cada data tenía sus propias particularidades, lo cual no permitía tratarlos a todos los archivos extraídos de forma general. En la Figura 2.6, se muestra el tratamiento de limpieza y

depuración de los datos extraídos y como terminan finalmente para ser agregados a la consolidación de la Data Completa; donde los datos de la columna “C” son los que contendrán los 105264 datos de energía entregada cada 10 minutos durante los 731 días de los dos años, siendo estos los datos analizados durante este estudio. En este proceso, verificamos las inconsistencias, para lo que se realiza el siguiente procedimiento [8]:

1. Valores de consumo anómalos y cortes por mora son detectados y reemplazados en base a la información de días similares.

2. En la fase de depuración, los valores perdidos se detectan y reemplazan mediante técnicas de interpolación. La interpolación toma en cuenta todos los puntos y correlaciona el comportamiento completo del conjunto usado.

	A	B	C	D	E	F	G	H	I	J	K	L
1	DATE	TIME	kWh-Del	0	105265	ULTIMA FILA			14	DATOS FALTANTES		
2	31/5/2020	00:00:00	5.17	ok	VERDADERO	0	144	31/5/2020	ok	0	0	
3	31/5/2020	00:10:00	5.1	ok	VERDADERO		144	31/5/2020	ok	0		
4	31/5/2020	00:20:00	5.23	ok	VERDADERO		144	1/6/2020	ok	0		
5	31/5/2020	00:30:00	5.17	ok	VERDADERO		144	2/6/2020	ok	0		
6	31/5/2020	00:40:00	5.17	ok	VERDADERO		144	3/6/2020	ok	0		
7	31/5/2020	00:50:00	4.91	ok	VERDADERO		144	4/6/2020	ok	0		
8	31/5/2020	01:00:00	4.85	ok	VERDADERO		144	5/6/2020	ok	0		
9	31/5/2020	01:10:00	5.23	ok	VERDADERO		144	6/6/2020	ok	0		
10	31/5/2020	01:20:00	5.23	ok	VERDADERO		144	7/6/2020	ok	0		
11	31/5/2020	01:30:00	5.17	ok	VERDADERO		144	8/6/2020	ok	0		
12	31/5/2020	01:40:00	5.23	ok	VERDADERO		144	9/6/2020	ok	0		
13	31/5/2020	01:50:00	5.17	ok	VERDADERO		144	10/6/2020	ok	0		
14	31/5/2020	02:00:00	5.29	ok	VERDADERO		144	11/6/2020	ok	0		
15	31/5/2020	02:10:00	5.23	ok	VERDADERO		144	12/6/2020	ok	0		
16	31/5/2020	02:20:00	5.17	ok	VERDADERO		144	13/6/2020	ok	0		
17	31/5/2020	02:30:00	5.17	ok	VERDADERO		144	14/6/2020	ok	0		
18	31/5/2020	02:40:00	4.91	ok	VERDADERO		144	15/6/2020	ok	0		
19	31/5/2020	02:50:00	5.23	ok	VERDADERO		144	16/6/2020	ok	0		
20	31/5/2020	03:00:00	5.23	ok	VERDADERO		144	17/6/2020	ok	0		
21	31/5/2020	03:10:00	5.23	ok	VERDADERO		144	18/6/2020	ok	0		
22	31/5/2020	03:20:00	5.23	ok	VERDADERO		144	19/6/2020	ok	0		

Figura 2.6: Depuración y Limpieza de los datos

En el caso, donde los datos eran cero, tenían la particularidad de que estaban seguidos de valores NaN, mostrando que había fallas de conexión con el servidor y el medidor AMI, por lo cual se interpolaban los datos para el segmento descrito.

2.2.3. Consolidación de la Data Completa

Con los datos de los clientes procesada y depurada se procede a consolidar la base de Datos de los 100 clientes, conservando sus nombres asignados por la zona de su ubicación y el lugar en el que ocupan en el escalafón de consumo de energía; como se muestra en

la Figura 2.7, donde en las primeras columnas tenemos la fecha y hora de lectura de la energía entregada y luego en orden cronológico los valores que tiene cada cliente, formando el histórico de los dos años de lecturas de las AMI.

	A	B	BE	BF	BG	BH	BI	BJ	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV	BW	BX
1	DATE	TIME	_55	1_56	1_57	1_58	1_59	1_60	2_2	2_3	2_4	2_5	2_6	2_7	2_8	2_9	2_10	3_1	3_2	3_3	4_1	4_2
53063	3/6/2021	11:30:00	5.70	13.34	17.39	9.61	9.68	9.81	117.18	63.17	34.74	19.80	5.30	17.89	10.43	4.94	6.34	10.76	8.38	8.51	27.53	22.18
53064	3/6/2021	11:40:00	7.58	13.58	17.81	9.14	10.33	9.82	133.14	60.09	35.44	21.56	4.70	14.03	9.45	4.85	6.50	10.70	8.07	8.68	30.24	23.94
53065	3/6/2021	11:50:00	5.73	12.70	17.39	10.44	9.96	9.81	164.64	56.64	35.88	21.29	3.58	12.68	11.01	4.79	5.53	10.80	8.56	7.79	29.67	33.42
53066	3/6/2021	12:00:00	5.19	13.12	17.01	11.54	10.43	9.81	164.22	58.59	35.31	22.30	3.01	13.52	9.76	5.02	5.71	10.72	8.86	8.10	33.58	33.30
53067	3/6/2021	12:10:00	5.14	12.97	16.97	10.76	10.66	9.81	171.36	51.25	34.40	24.17	2.93	12.77	7.70	5.02	5.83	11.10	7.92	8.02	36.41	36.45
53068	3/6/2021	12:20:00	5.52	12.54	15.96	10.03	10.60	9.81	156.45	51.06	35.12	22.79	3.20	12.94	8.86	5.11	5.31	11.02	8.43	6.32	38.68	35.63
53069	3/6/2021	12:30:00	5.44	12.17	16.17	10.49	10.02	9.82	138.81	51.44	34.46	23.16	2.96	18.82	10.58	4.97	5.56	10.67	7.87	6.09	37.83	36.86
53070	3/6/2021	12:40:00	4.96	11.82	16.34	9.82	9.72	9.81	142.17	58.71	34.15	22.36	2.84	30.66	10.96	5.18	5.88	10.69	7.83	6.31	38.12	31.53
53071	3/6/2021	12:50:00	4.88	13.48	17.39	10.23	11.05	9.81	158.13	51.06	33.71	19.56	3.37	29.82	9.89	5.20	5.56	10.60	7.88	6.64	38.87	25.55
53072	3/6/2021	13:00:00	5.87	12.38	17.56	9.60	11.12	9.81	135.45	58.21	33.99	15.70	3.80	30.16	10.89	4.88	4.28	11.06	6.23	6.92	38.56	34.30
53073	3/6/2021	13:10:00	5.36	12.38	16.46	10.01	9.44	9.81	152.04	48.86	33.39	12.89	6.13	30.66	9.28	4.81	4.07	10.57	6.29	7.27	34.40	25.67
53074	3/6/2021	13:20:00	4.47	12.15	16.04	9.80	9.91	9.81	159.81	46.54	31.59	14.39	7.73	20.83	7.07	4.78	4.06	7.20	6.65	6.64	32.29	22.71
53075	3/6/2021	13:30:00	5.01	12.24	16.30	9.08	10.62	9.82	177.45	46.79	30.93	13.90	8.10	23.02	6.00	4.90	3.77	3.07	7.05	5.87	28.29	25.74
53076	3/6/2021	13:40:00	3.80	12.20	15.96	9.33	10.72	9.82	194.88	45.92	30.37	16.94	8.10	26.80	7.25	5.02	3.18	2.40	5.87	5.31	17.92	32.60
53077	3/6/2021	13:50:00	5.16	13.26	16.00	9.00	12.01	9.81	161.07	45.23	31.41	16.66	7.40	27.72	9.99	4.72	3.77	7.54	6.24	5.56	28.13	35.06
53078	3/6/2021	14:00:00	5.93	12.97	15.16	9.58	10.63	9.82	113.19	42.09	33.01	16.91	5.64	23.27	9.67	4.65	3.58	8.08	5.43	6.76	30.68	35.78
53079	3/6/2021	14:10:00	4.17	11.24	16.00	8.78	9.44	9.82	103.95	42.15	34.27	18.61	2.82	25.79	9.55	4.88	3.44	7.48	5.27	7.59	29.80	35.75
53080	3/6/2021	14:20:00	4.32	11.83	16.17	8.27	9.53	9.82	125.16	41.34	34.30	19.45	2.74	26.63	10.06	4.83	4.11	8.70	5.77	7.60	27.66	36.92
53081	3/6/2021	14:30:00	5.06	10.98	15.67	8.47	9.35	9.82	121.59	46.48	35.37	19.78	3.00	22.60	7.64	4.91	4.21	9.19	4.75	7.36	27.22	38.02
53082	3/6/2021	14:40:00	4.33	11.03	15.58	9.73	10.02	9.82	119.49	47.48	34.56	20.92	4.04	23.60	7.50	4.79	3.91	9.70	4.80	7.29	15.75	36.67
53083	3/6/2021	14:50:00	4.70	11.52	15.08	9.19	8.40	9.82	148.47	46.17	33.93	21.48	4.56	27.55	9.35	4.76	4.10	9.63	4.81	7.07	13.89	18.33
53084	3/6/2021	15:00:00	4.89	12.39	15.54	10.28	8.39	9.82	166.32	47.11	34.43	22.87	3.71	27.55	6.89	4.84	4.36	9.43	5.20	7.07	25.33	14.11
53085	3/6/2021	15:10:00	5.57	12.97	14.74	9.39	9.62	9.82	165.69	46.54	34.43	24.44	2.59	26.88	7.49	4.87	3.89	9.59	5.41	6.31	25.14	7.78
53086	3/6/2021	15:20:00	5.49	11.30	14.53	8.52	9.58	9.83	138.81	45.73	35.03	25.04	3.64	23.60	8.51	4.69	4.28	9.12	5.02	6.07	25.64	4.91
53087	3/6/2021	15:30:00	5.67	12.29	12.85	8.96	9.71	9.83	149.73	46.73	34.90	23.09	3.83	19.99	7.24	4.66	4.30	8.23	4.88	6.14	26.74	5.17

Figura 2.7: Forma en como se ve la Data Completa

2.2.4. Reducción y Clasificación de la Data

La data completa tiene una resolución de cada diez minutos, pero para obtener una reducción de datos más efectiva y sin perder información importante en los registros de energía, se suman todos los valores de energía dentro de cada hora, obteniendo una reducción de “105264 filas × 101 columnas” a “17544 filas × 101 columnas”, el valor extra de la columna se refiere a los valores de Tiempo que tiene la data, como se muestra en la Figura 2.8. En el Anexo A, se muestra la forma de cómo se realizó la reducción del df.

	TIME	1_1	1_2	1_3	1_4	1_5	1_6	1_7	1_8	1_9	...	10_2	10_3	11_1	11_2	11_3	11_4	11_5	12_1	12_2	12_3
0	2020-05-31 00:00:00	309.96	15.69	125.81	57.96	6.04	46.94	19.53	3.36	5.17	...	2.01	0.001275	4.12	5.42	10.3005	1.3545	2.424	2.286	2.08	2.409
1	2020-05-31 00:10:00	300.78	15.69	125.81	59.01	6.04	45.89	19.53	3.36	5.10	...	2.04	0.001350	4.01	5.50	11.3715	1.1565	2.034	2.754	2.24	2.403
2	2020-05-31 00:20:00	304.02	15.44	125.81	58.38	6.27	45.78	19.32	3.36	5.23	...	2.02	0.001350	3.91	4.94	10.8990	1.3950	2.148	2.358	2.22	2.421
3	2020-05-31 00:30:00	301.32	15.44	125.81	58.17	6.50	46.04	17.75	3.36	5.17	...	1.95	0.001350	3.74	5.04	12.2850	0.9585	2.100	2.358	2.07	2.412
4	2020-05-31 00:40:00	304.02	15.69	125.81	57.96	6.16	45.68	16.70	3.15	5.17	...	1.91	0.001350	3.74	5.40	10.7415	1.3680	2.316	2.340	1.97	2.418
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
105259	2022-05-31 23:10:00	381.78	283.36	120.12	73.92	66.23	44.15	24.78	20.79	17.20	...	1.47	0.001200	4.17	4.18	13.2930	2.5110	2.580	2.691	2.23	1.908
105260	2022-05-31 23:20:00	392.04	279.13	120.12	73.50	65.78	43.10	24.89	18.69	18.65	...	1.43	0.001200	4.18	4.07	13.1985	2.5425	2.358	2.781	2.26	1.947
105261	2022-05-31 23:30:00	380.16	296.31	120.12	72.45	65.32	43.00	24.99	17.64	16.95	...	1.49	0.001200	4.01	4.43	13.5450	2.2860	2.676	2.565	2.28	1.875
105262	2022-05-31 23:40:00	392.58	295.31	120.12	69.09	64.87	42.26	24.68	21.21	17.45	...	1.45	0.001275	4.03	4.81	13.7970	2.7315	2.238	2.880	2.26	1.824
105263	2022-05-31 23:50:00	383.94	285.85	120.12	61.95	64.75	41.21	25.20	21.00	17.58	...	1.49	0.001200	3.79	4.64	13.5135	2.1285	2.130	2.664	2.23	1.956

105264 rows × 101 columns

	Timestamp	1_1	1_2	1_3	1_4	1_5	1_6	1_7	1_8	1_9	...	10_2	10_3	11_1	11_2	11_3	11_4	11_5	12_1
0	2020-05-31 00:00:00	1821.96	93.89	754.86	349.44	37.28	276.95	109.63	19.95	30.75	...	11.99	0.008025	23.04	31.36	66.5595	7.2045	13.170	14.553
1	2020-05-31 01:00:00	1833.84	93.89	754.92	282.45	37.51	280.72	113.41	21.84	30.88	...	12.03	0.008025	22.36	31.74	67.7880	7.1505	11.574	14.319
2	2020-05-31 02:00:00	1817.64	94.89	754.92	265.86	36.95	269.75	116.46	21.63	31.00	...	12.04	0.008025	22.87	32.24	66.6855	7.1820	10.728	13.797
3	2020-05-31 03:00:00	1854.36	93.64	754.98	259.77	37.63	276.63	101.13	19.32	31.26	...	12.04	0.008100	23.34	30.78	67.8510	6.9525	10.842	13.095
4	2020-05-31 04:00:00	1818.18	92.89	755.04	270.27	37.51	276.37	100.18	17.43	30.62	...	12.08	0.008100	22.61	30.78	66.6540	6.9435	10.182	14.166
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
17539	2022-05-31 19:00:00	2411.64	1559.24	720.54	724.92	651.96	273.02	145.96	341.67	106.10	...	9.35	1.257750	65.50	30.84	82.2780	20.5380	37.452	32.337
17540	2022-05-31 20:00:00	2380.86	1793.79	720.72	723.24	666.79	266.96	144.82	367.08	95.38	...	9.04	0.252150	57.81	29.93	82.1835	19.1160	27.474	31.527
17541	2022-05-31 21:00:00	2370.60	1872.48	720.78	610.89	526.68	268.23	145.87	309.75	94.56	...	8.88	0.007425	42.18	27.72	82.0575	16.4745	19.878	27.414
17542	2022-05-31 22:00:00	2199.96	1693.95	720.78	499.38	404.58	257.83	148.49	146.16	100.17	...	8.80	0.007350	26.31	26.83	80.8290	15.9030	20.262	23.580
17543	2022-05-31 23:00:00	2310.12	1735.77	720.72	428.40	393.30	257.66	149.01	120.12	104.84	...	8.82	0.007350	24.59	26.79	81.1440	14.5530	14.622	16.524

17544 rows × 101 columns

Figura 2.8: Reducción de la data a horas

La clasificación de los datos se realiza con conocimientos previos sobre la forma en que los clientes especiales de la EEASA registran sus consumos de energía, conociendo de antemano que la mayoría de los sectores industriales trabajan en jornadas laborales típicas entre los días lunes y viernes, que los fines de semana son pocos los clientes industriales que trabajan plenamente y que los días de feriado son una opción muy poco favorable para laborar; el comportamiento de las curvas diarias generadas en base al consumo de energía suministrada, es similar en la mayoría de clientes especiales, afectando claramente su

forma los días entre semana, como muestra la Figura 2.9, donde se observa que de lunes a viernes la gráfica presenta los picos más altos de consumo, diferenciándose de los consumos con los fines de semana y feriados; por lo que es indispensable realizar la clasificación de estas curvas. Para realizar esta primera clasificación se utilizan “estampas de tiempo”, que se definen con el empleo de la librería “time” y el desarrollo de líneas de código para cumplir con la clasificación [9].

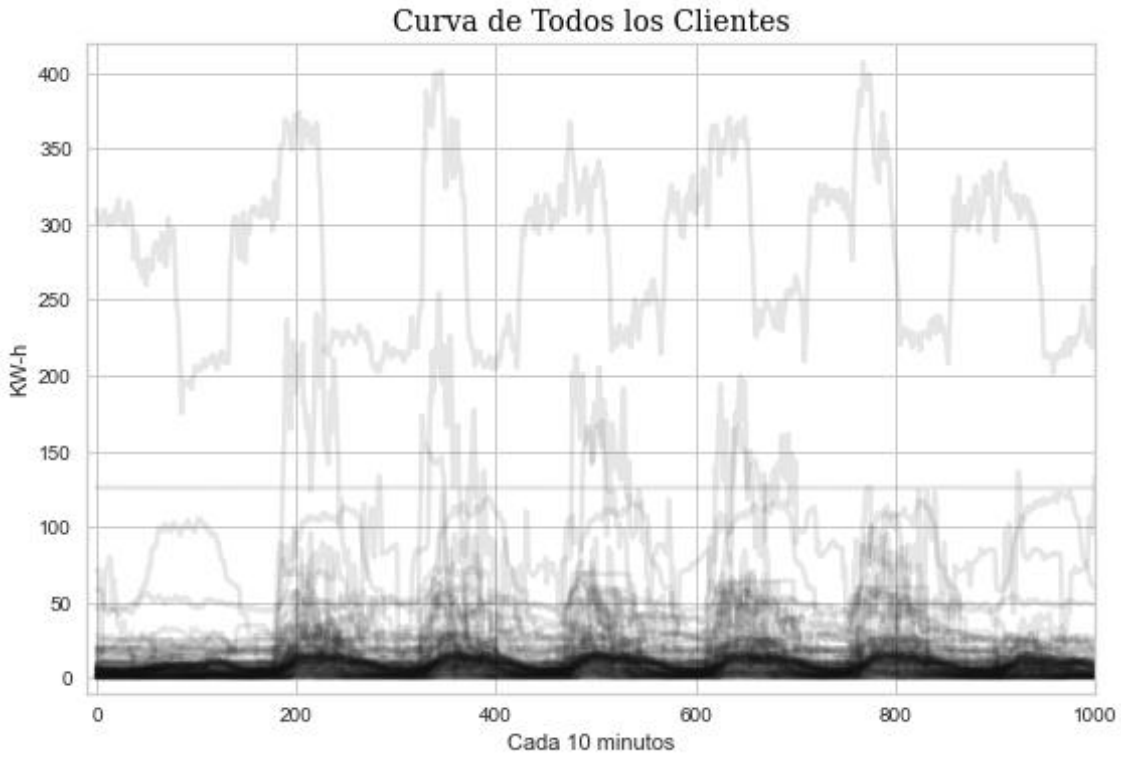


Figura 2.9: Curvas de todos los clientes con intervalos de tiempo = 10 minutos

En la Tabla 2.3, se puede observar un resumen de la clasificación realizada, donde se detalla los días de cada grupo, los datos dentro de cada grupo y la matriz formada para la agrupación con los índices k-medias.

Tabla 2.3: Data Frame formados después de la clasificación

Clasificación	Días	Datos por Día	Data Frame
Grupo de lunes a viernes	500	24	12000 x 100
Grupo de fines de semana	173	24	4152 x 100
Grupo de feriados	58	24	1392 x 100

En la Figura 2.10 se muestra las curvas diarias del Grupo de lunes a viernes, después de realizar la clasificación. Observar las curvas de los otros grupos en el Anexo B.

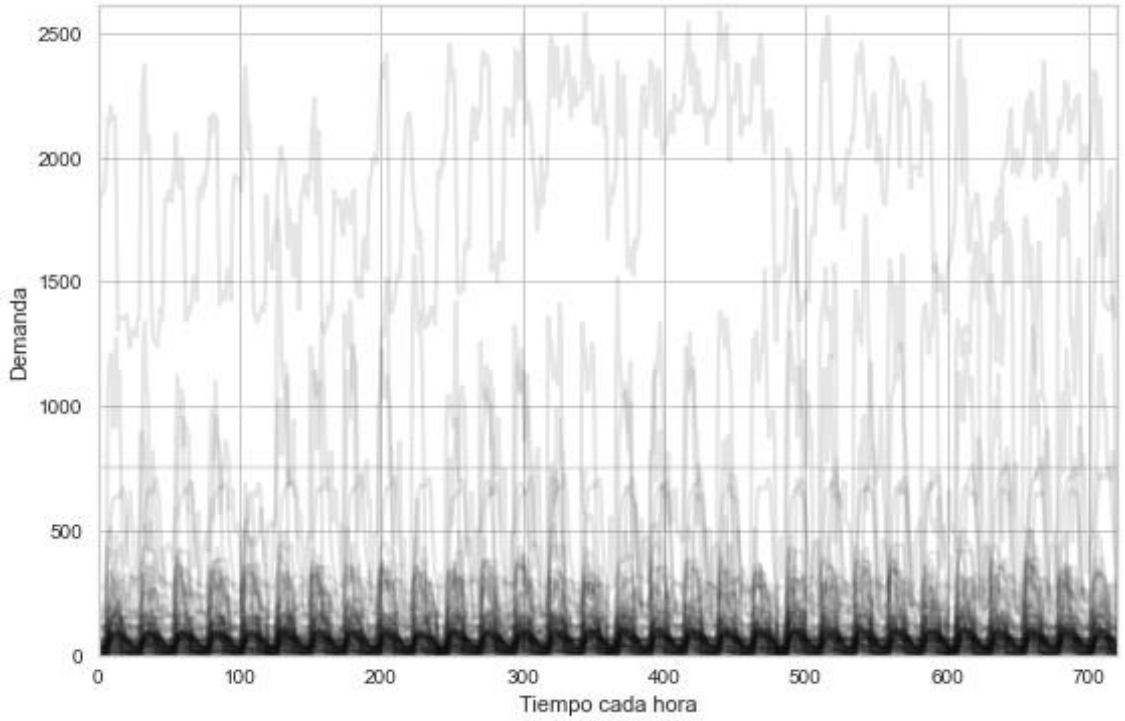


Figura 2.10: Curvas Diarias del grupo de lunes a viernes

2.2.5. Normalización de Datos

Para la normalización de datos se decidió utilizar la normalización mínimo-máximo para normalizar los consumos de todos los clientes:

$$NL = {L - \min(L)} \Big/ {\max(L) - \min(L)} \qquad \textbf{Ecuación 2.1}$$

donde L representa el consumo de cada hora de un cliente, $min(L)$ y $max(L)$ son los valores mínimo y máximo en el conjunto de datos de cada cliente, respectivamente. Los datos normalizados se utilizan luego para el análisis estadístico [10].

Los consumos de la Figura 2.8, se normalizaron utilizando la metodología mínimo máximo de manera que todos los valores cayeron dentro del intervalo [0,1]. La Figura 2.11 muestra los datos normalizados y la Figura 2.12 las curvas con los datos normalizados del Grupo de lunes a viernes. Ver Anexo B para observas las curvas normalizadas de los otros grupos.

	1_1	1_2	1_3	1_4	1_5	1_6	1_7	1_8	1_9	1_10	...
0	0.539168	0.077785	0.973451	0.143986	0.007778	0.328478	0.061492	0.028455	0.096691	0.004792	...
1	0.556499	0.078268	0.973684	0.059006	0.008040	0.301289	0.068024	0.028825	0.096882	0.004792	...
2	0.575737	0.078075	0.973607	0.023715	0.007267	0.295086	0.053556	0.032890	0.096278	0.004949	...
3	0.574523	0.077688	0.973762	0.019481	0.007506	0.250580	0.045076	0.098300	0.093830	0.004635	...
4	0.578336	0.069424	0.973762	0.023998	0.007767	0.257431	0.045076	0.109017	0.096278	0.004792	...
...	...	...	...	...	...	...	...	...	...	...	...
11995	0.753380	0.599671	0.929148	0.775833	0.706149	0.282138	0.105998	0.598300	0.334573	0.524588	...
11996	0.743501	0.690255	0.929381	0.773574	0.722989	0.249447	0.104488	0.643016	0.300499	0.523771	...
11997	0.740208	0.720645	0.929458	0.622530	0.563890	0.256298	0.105879	0.542129	0.297893	0.462058	...
11998	0.685442	0.651696	0.929458	0.472614	0.425243	0.200194	0.109350	0.254250	0.315724	0.164148	...
11999	0.720797	0.667847	0.929381	0.377188	0.412434	0.199277	0.110039	0.208426	0.330568	0.111532	...

12000 rows × 100 columns

Figura 2.11: Datos normalizados del grupo de lunes a viernes

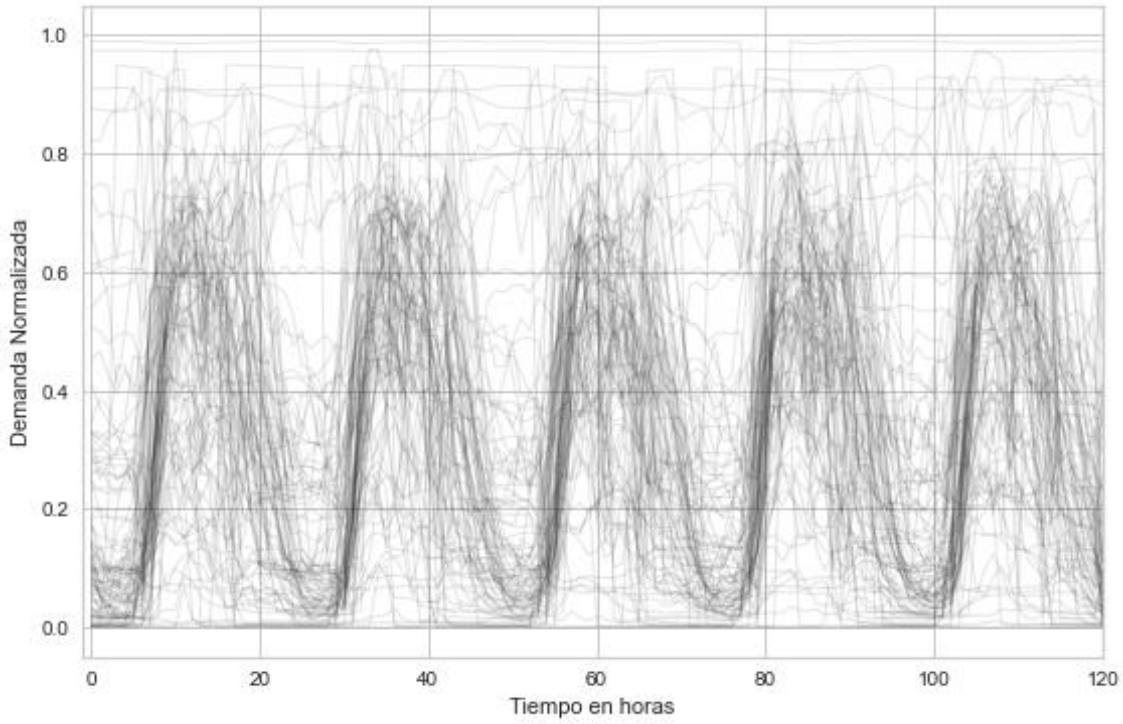


Figura 2.12: Curvas Diarias Normalizadas del grupo de lunes a viernes

2.2.6. Agrupamiento

Este subproceso consiste en conglomerar las curvas diarias en grupos compactos con propiedades distintas y significativas del resto de grupos. En este trabajo, investigamos la

capacidad de agrupamiento del algoritmo KM y sus variaciones. El algoritmo clásico de KM [11] sigue un procedimiento iterativo heurístico para agrupar los Patrones Representativos de Demanda en K grupos. Primero se eligen centroides de K grupos (elegidos al azar del conjunto de las Curvas Diarias). A continuación, cada Curva Diaria se clasifica en sus grupos más cercanos (según la función de optimización).

Finalmente, los centroides se vuelven a calcular promediando las Curvas Diarias de sus miembros. El proceso se repite hasta que los centroides del conglomerado sean estables. En el algoritmo KM clásico, la distancia entre dos puntos de la Curva Diaria, se mide por la distancia euclidiana. Sin embargo, los estudios en [12] sugieren que la distancia con la métrica de Deformación Dinámica en el Tiempo, DTW por sus siglas en inglés, es más apropiada para el agrupamiento de series temporales. Por lo tanto, en la Figura 2.13 se muestran un ejemplo de los resultados utilizando el KM clásico, y en la Figura 2.14 otro ejemplo con la métrica de DTW. La curva en rojo representa los centroides de cada clúster, y las de color negro, los clientes que conforman el grupo; el valor de “n” da el número de clientes que están en cada grupo, siendo en este caso un k = 4 y 6 respectivamente.

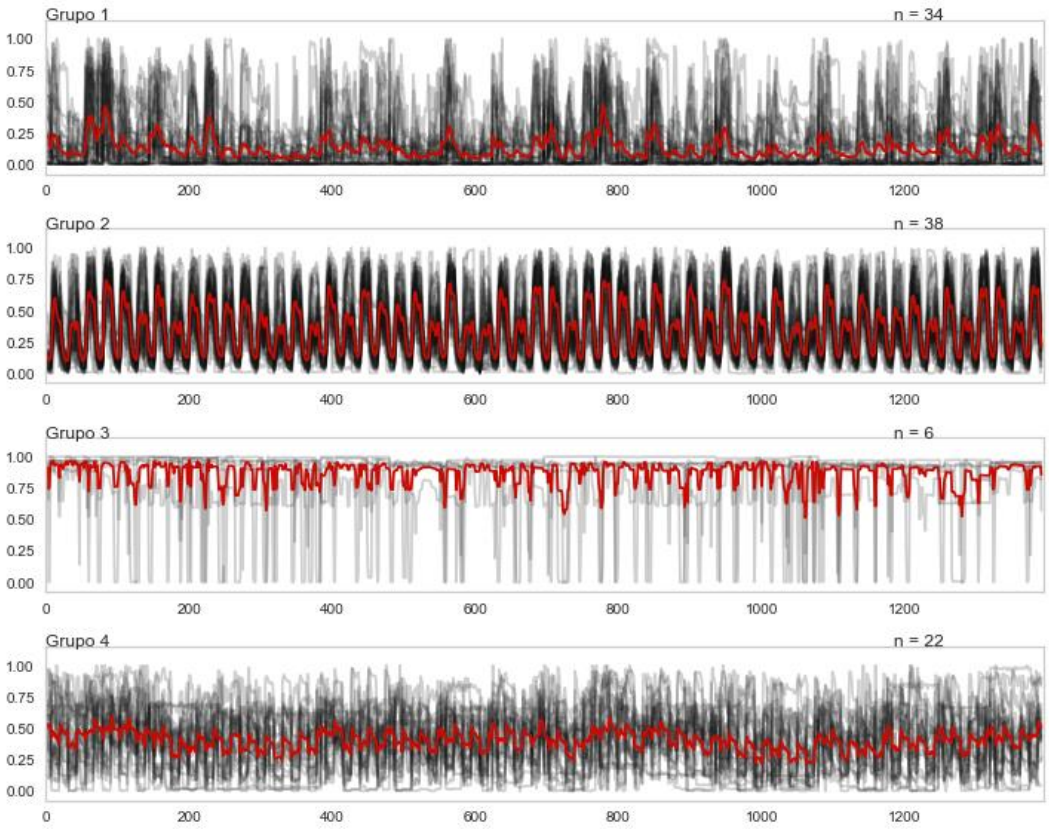


Figura 2.13: Resultado del agrupamiento con Euclidean k-Means para feriados y K=4

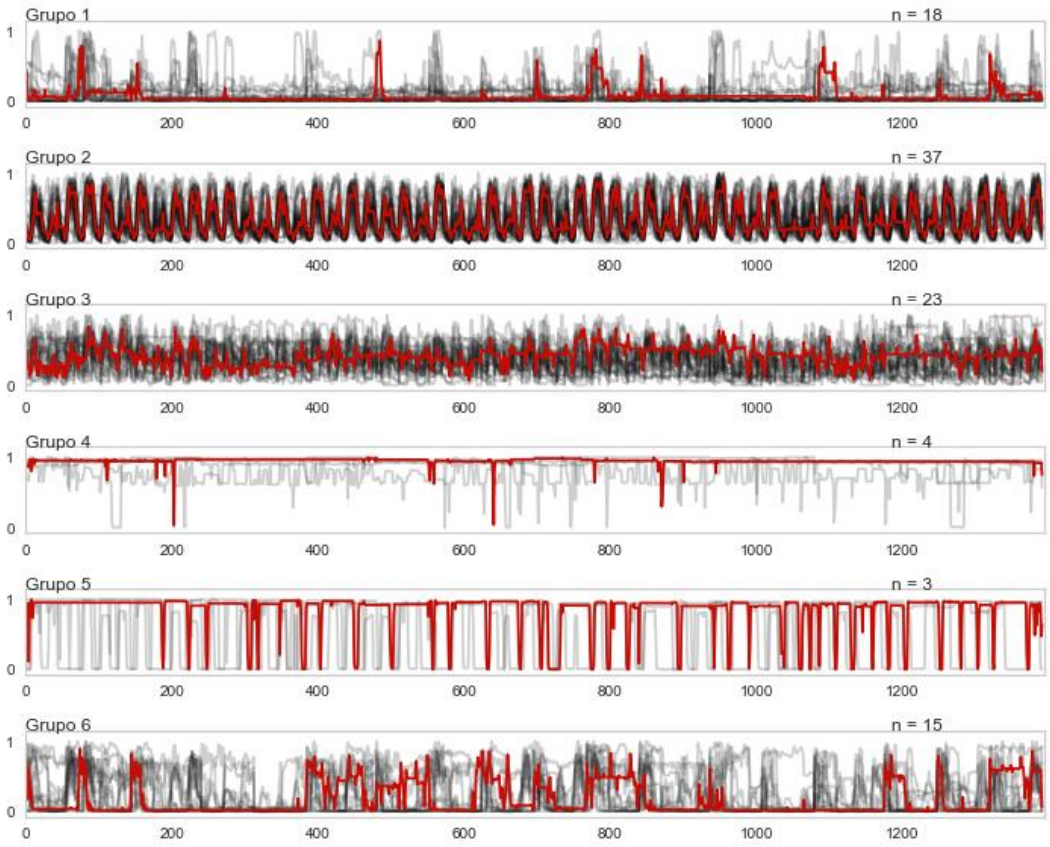


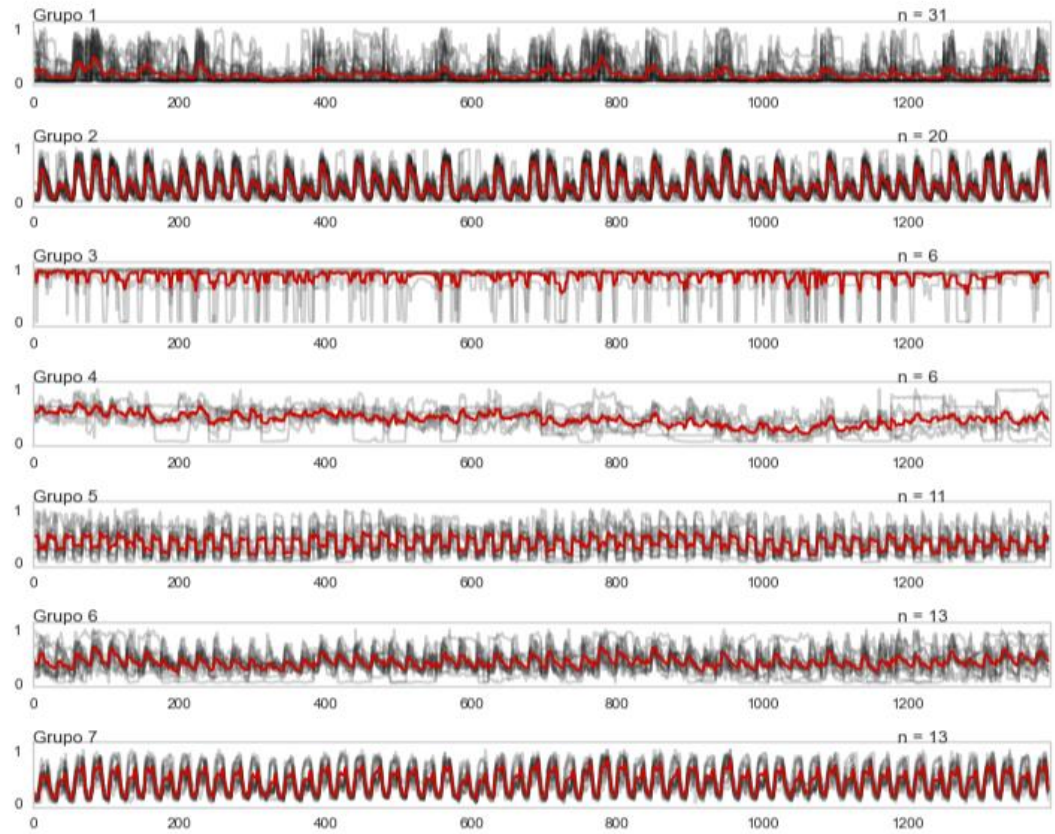
Figura 2.14: Resultado del agrupamiento con Soft-DTW k-Means para feriados y K=6

2.2.6.1. Índices de Validación

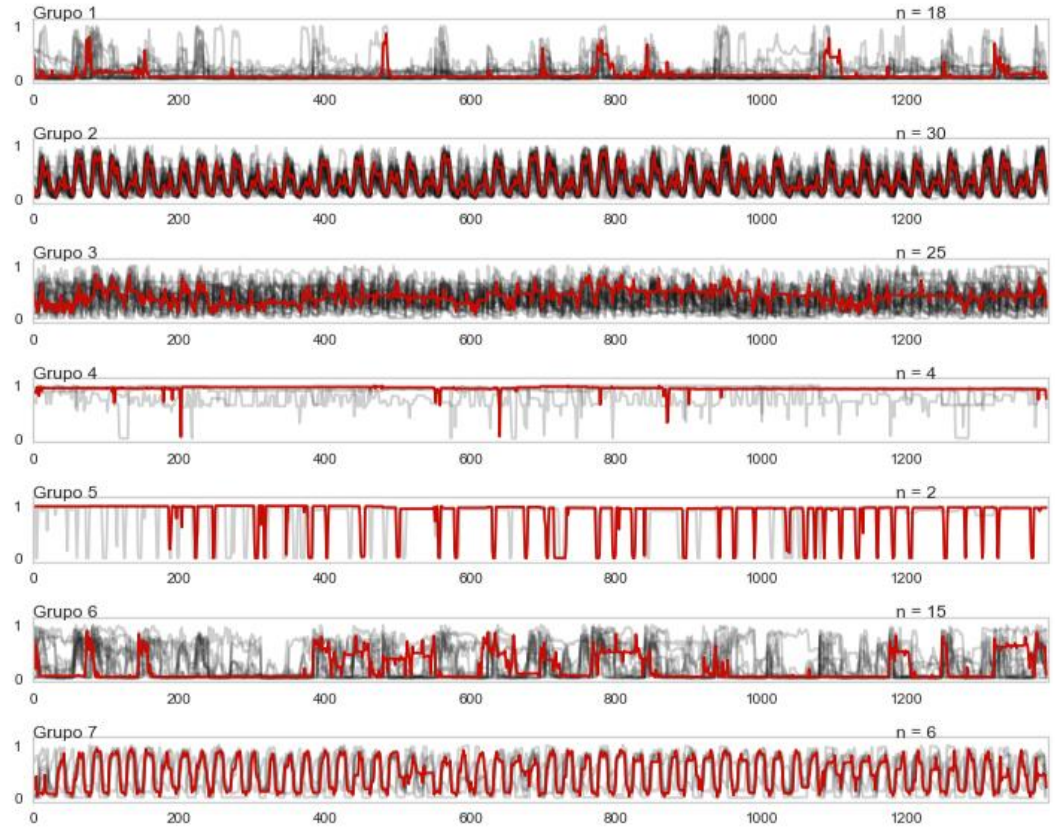
A pesar del potencial práctico del agrupamiento, la evaluación de las estructuras de clusterización sigue siendo un desafío para las aplicaciones en el sector eléctrico. Por lo que se utilizan los índices WCSS y SC, los cuales fueron estudiados en el apartado 1.5.7.2; se presentan los resultados en la Figura 2.15 con los tres métodos de agrupamiento, tomando como ejemplo un valor de k = 7, donde la diferencia de los centroides con los diferentes índices de evaluación es notoria y poco acogida para la clasificación deseada.

El análisis visual de cada agrupamiento previo con los ejemplos mostrados, la ubicación de los centroides con respecto a las curvas que conforman el clúster y la experiencia en el comportamiento de las curvas de los clientes industriales, son factores importantes para la definición del valor de K y el método de clusterización.

Euclidean k-means



DBA k-means



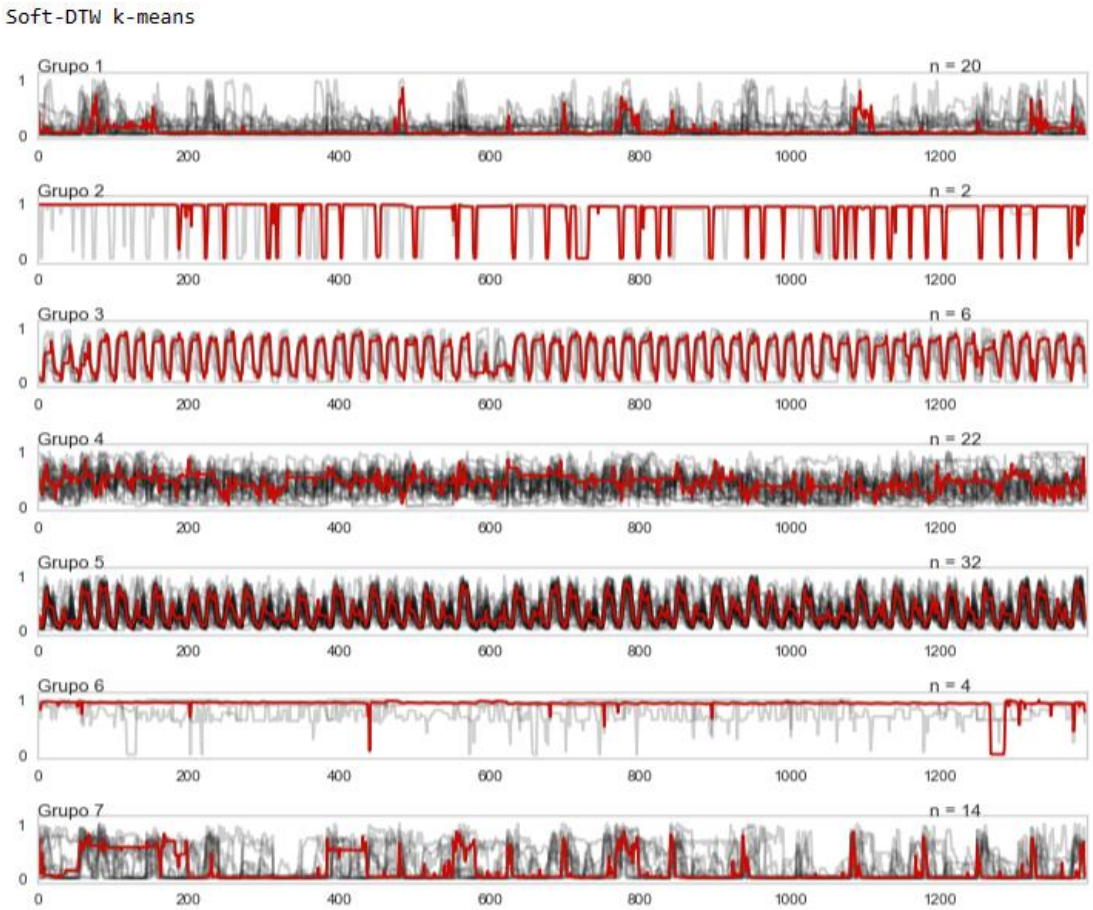


Figura 2.15: Agrupamiento con los diferentes índices de validación con K=7

Las curvas mostradas representan a uno de los agrupamientos definidos con un valor de k usando los 3 métodos; realizada a los datos clasificados como el grupo de los días feriados, exponiendo el comportamiento de las curvas diarias durante los 58 días de lecturas.

2.2.7. Definición de Grupos

Aplicando los procesos desarrollados en el agrupamiento de los clientes y los índices de validación para las tres divisiones previamente realizadas, se debe elegir el mejor método de agrupación para un valor de K definido, como se muestra en la Figura 2.16, donde se observa claramente que se define el conjunto de datos a ser estudiado y se llama a la función para realizar el estudio de agrupamiento con un valor definido de K igual a 5, valor que ha sido elegido para emplearlo en la clasificación de este estudio, y el empleo de los

tres métodos de agrupamiento, para elegir cuál de los tres métodos arroja los resultados más precisos para la segmentación de los clientes.

	0	1	2	3	4	5	6	7	8	9	...	4142	4143	
1_1	0.700195	0.705078	0.698730	0.713379	0.698730	0.712402	0.656738	0.626465	0.621094	0.639160	...	0.262939	0.269775	0.2
1_2	0.040253	0.040253	0.040710	0.040131	0.039795	0.039459	0.038879	0.038513	0.037842	0.037842	...	0.047974	0.048187	0.0
1_3	0.975098	0.975098	0.975098	0.975586	0.975586	0.975586	0.975098	0.975098	0.974609	0.974121	...	0.929199	0.929199	0.9
1_4	0.273438	0.182007	0.159424	0.151123	0.165405	0.156494	0.180054	0.250732	0.345459	0.486572	...	0.803711	0.808594	0.8
1_5	0.038177	0.038574	0.037598	0.038788	0.038574	0.037598	0.034943	0.033966	0.031769	0.031769	...	0.037384	0.039001	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
11_4	0.027359	0.026657	0.027069	0.024063	0.023956	0.024307	0.029297	0.028290	0.030533	0.029709	...	0.400879	0.394287	0.3
11_5	0.080994	0.058105	0.045990	0.047638	0.038177	0.041962	0.076721	0.129272	0.191406	0.254639	...	0.460693	0.441895	0.3
12_1	0.138428	0.134399	0.125488	0.113525	0.131836	0.130859	0.156494	0.390381	0.545410	0.453857	...	0.397217	0.396240	0.3
12_2	0.206055	0.183350	0.181152	0.176270	0.187500	0.202148	0.376953	0.448486	0.460938	0.458252	...	0.299316	0.387207	0.5
12_3	0.540039	0.537598	0.538574	0.536621	0.535645	0.491211	0.310547	0.203735	0.209473	0.212524	...	0.201416	0.160767	0.2

100 rows × 4152 columns

6.3.1 K = 5

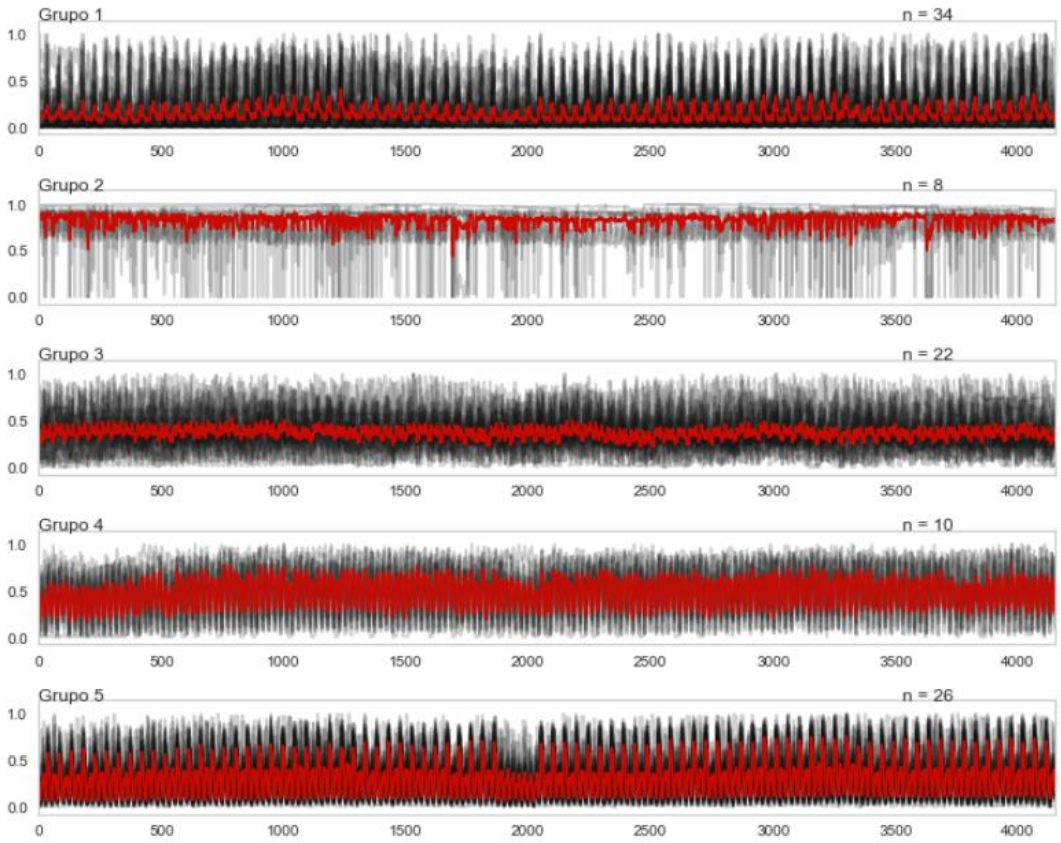
```
clusters = 5
method = 3
```

```
(100, 4152, 1)
Euclidean k-means
DTW k-means
Soft-DTW k-means
```

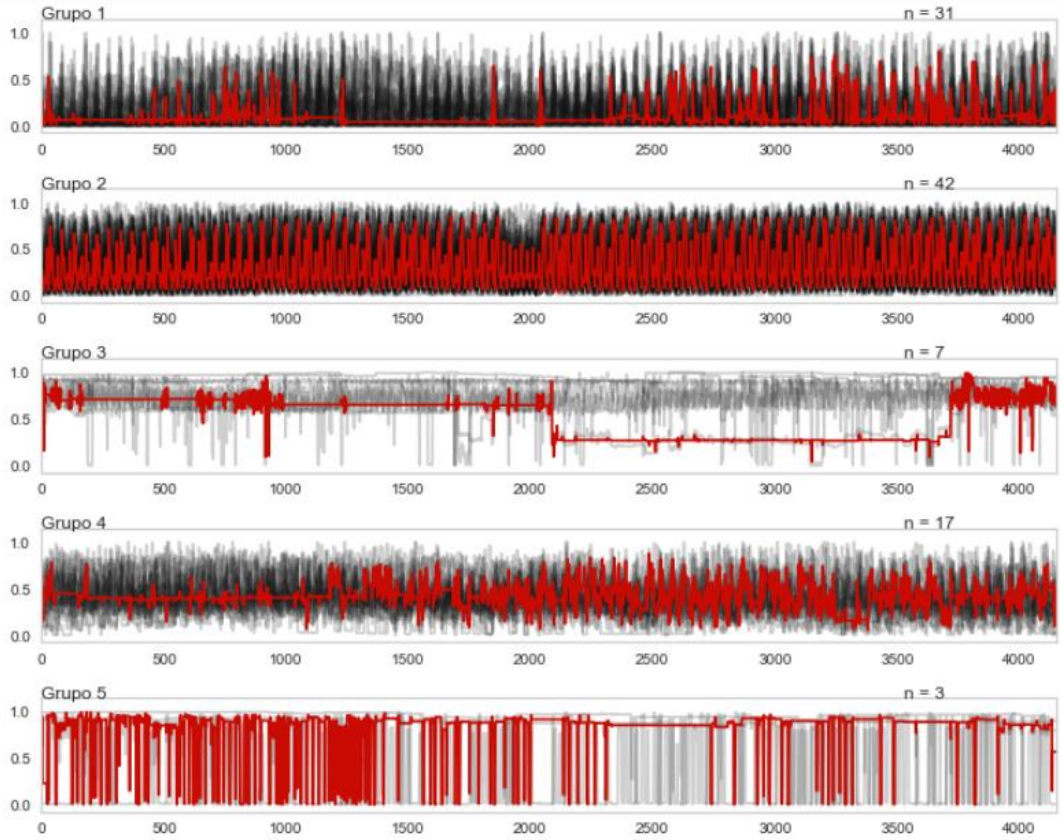
Figura 2.16: Definición de K = 5 y uso de los 3 métodos de agrupamiento

En el proceso de la evaluación visual, es importante observar minuciosamente las gráficas para determinar el método que mejor describe al grupo, tomando en cuenta la posición de los centroides calculados y plenamente definidos en color rojo, el comportamiento de las curvas de los clientes con respecto al centroide dentro de cada grupo, en la Figura 2.17, se observa cómo fueron clasificados para un k=5 y el valor de *n* con el número de clientes que se encuentra en cada grupo, es importante mencionar que el valor de k más idóneo es 5, ya que para los valores de k igual a 4, 6 y 7 no se muestra una homogeneidad en los clústeres, como se vio en las Figuras 2.13, 2.14 y 2.15, se toma como referencias a estas figuras, ya que fueron realizadas con el mismos conjunto de datos.

Euclidean k-means



DBA k-means



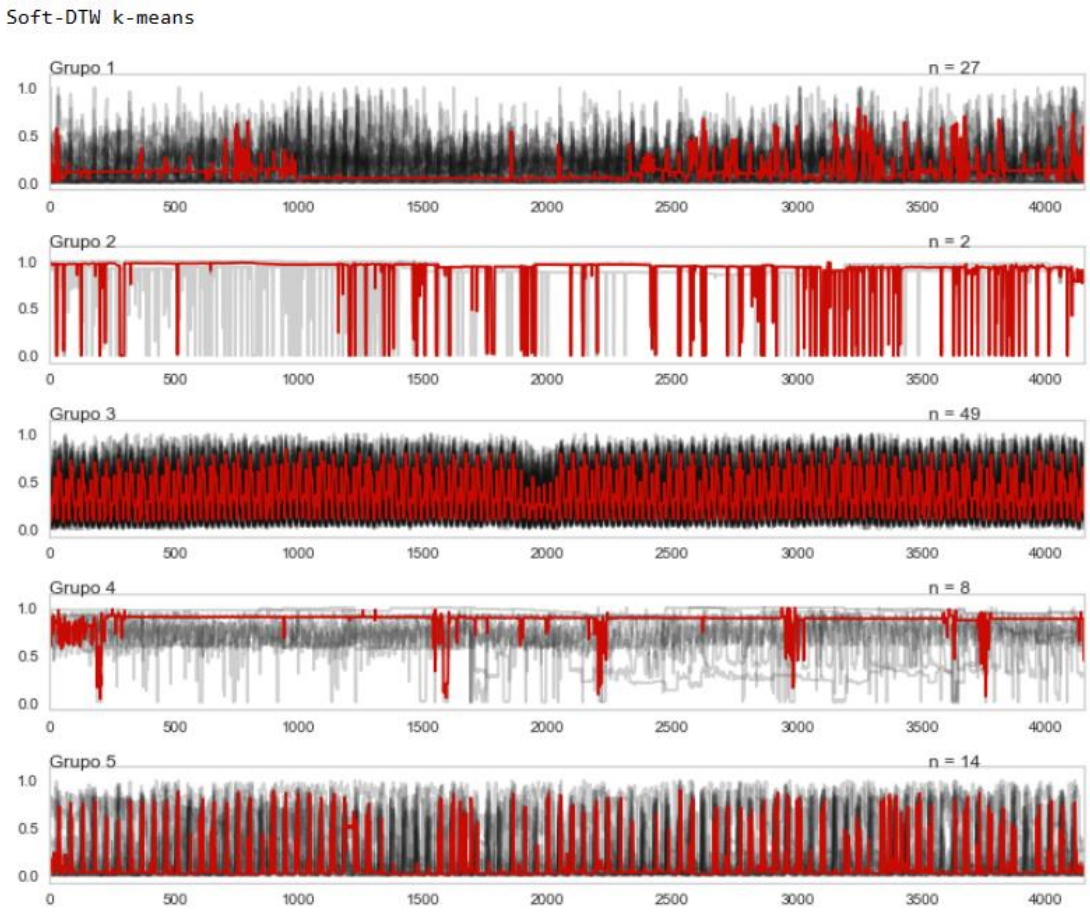


Figura 2.17: Curvas agrupadas utilizando los 3 métodos con k = 5

Con el proceso de observación detenida, podemos ver que para el método 1, que utiliza las distancias euclidianas, no define bien los centroides para el grupo 1, 2 y 3, distanciando mucho los valores del centroide con los valores máximos y mínimos de las curvas diarias de los clientes, por lo que el índice de evaluación de distancias Euclidianas no es una buena opción para la clasificación. El método 2, que emplea el promedio de baricentro del DTW, los centroides definidos del grupo 1 y 3, no describen correctamente el comportamiento de os miembros de los grupos, por lo que podría resultar una opción viable pero no con la certeza de que será la correcta. Finalmente, para el método 3, que emplea la combinación de los índices de evaluación SC y el DTW, define los centroides de las curvas de mejor manera para 4 de los 5 grupos, teniendo como valor agregado que el grupo con mayor cantidad de miembros tiene la mejor definición del grupo, lo que le hace la opción más viable de los tres métodos y de la cual se escogerá los centroides y la asignación de los grupos.

En la Figura 2.18, se puede observar que el primer segmento de líneas de código se asigna a una variable los valores de la asignación de los grupos y en el segundo segmento se asigna en otra variable los valores de los centroides para el agrupamiento realizado.

```
array([0, 3, 3, 4, 3, 4, 3, 0, 4, 0, 4, 2, 3, 0, 0, 3, 4, 4, 4, 3, 4, 3,
       2, 3, 0, 4, 0, 4, 3, 2, 0, 0, 3, 0, 3, 0, 3, 4, 4, 0, 0, 0, 4, 4,
       0, 1, 0, 4, 4, 3, 0, 1, 3, 4, 4, 4, 0, 4, 0, 0, 2, 4, 0, 4, 4, 4,
       4, 2, 4, 3, 3, 4, 0, 4, 4, 3, 4, 3, 0, 0, 0, 4, 3, 3, 0, 0, 4, 0,
       0, 2, 3, 4, 3, 0, 0, 0, 4, 4, 4, 4], dtype=int64)
```

	0	1	2	3	4	5	6	7	8	9	...
0	0.466438	0.143712	0.022801	0.055663	0.069750	0.059627	0.034052	0.054950	0.057767	0.092079	...
1	0.983352	0.983527	0.983638	0.983603	0.983419	0.983120	0.982758	0.982386	0.982050	0.981778	...
2	0.110802	0.067684	0.068538	0.066389	0.054953	0.030965	0.074894	0.255974	0.516536	0.527092	...
3	0.773451	0.470395	0.561543	0.599518	0.516557	0.267201	0.622991	0.742281	0.691559	0.755737	...
4	0.314504	0.139263	0.105178	0.060540	0.228885	0.103419	0.262314	0.313377	0.378687	0.488005	...

5 rows × 1392 columns

Figura 2.18: Asignación de Grupos y valores de los Centroides

Es importante recalcar, que para los grupos previamente clasificados como son el Grupo de lunes a viernes y el Grupo de fines de semana, se aplica la misma metodología para la agrupación de los clientes, siendo completamente diferente el comportamiento de los índices de evaluación con respecto a las curvas diarias, siendo escogidos en algunos casos los otros índices por el comportamiento descrito; ver Anexo A.

2.2.8. Conformación de Grupos y Estudio Estadístico

Con el método y el valor de k definidos, podemos observar los valores asignados para el agrupamiento y los valores de los centroides calculados, como se puede ver en la Figura 2.18, donde con un array se presentan los valores de cada grupo y en un pequeño df de (5filas x 1392columnas), los valores de los centroides calculados para cada hora de consumo de energía, estos son valores normalizados.

Con un pequeño arreglo de filas y columnas, como se puede observar en la Figura 2.19, al df normalizado del grupo de días feriados, le asignamos los valores otorgados por el índice de validación.

	0	1	2	3	...	1388	1389	1390	1391	Grupos
1_1	0.675224	0.677938	0.650804	0.716552	...	0.620121	0.617825	0.568357	0.581090	0
1_2	0.748290	0.758575	0.652717	0.680796	...	0.113015	0.107580	0.103997	0.102846	3
1_3	0.967538	0.967640	0.967640	0.967844	...	0.918538	0.918640	0.918640	0.918844	3
1_4	0.055345	0.018177	0.007325	0.008139	...	0.803310	0.664677	0.507596	0.418882	4
1_5	0.116467	0.054153	0.054153	0.054717	...	0.017030	0.017300	0.018108	0.019235	3
...	...	...	...	...	...	...	...	...	...	...
11_4	0.019700	0.025805	0.014983	0.026498	...	0.218923	0.206160	0.195339	0.208935	0
11_5	0.013710	0.007192	0.012811	0.007492	...	0.174558	0.140171	0.103012	0.074843	4
12_1	0.088134	0.101428	0.077958	0.050386	...	0.407024	0.294108	0.168554	0.155096	4
12_2	0.170968	0.178710	0.162258	0.157097	...	0.718387	0.516129	0.272258	0.208710	4
12_3	0.603521	0.601006	0.594719	0.595557	...	0.631811	0.623219	0.610855	0.611693	4

100 rows × 1393 columns

Figura 2.19: Valores asignados del agrupamiento y centroides calculados

A continuación, con los valores asignados dentro del df, se crea los grupos con los valores asignados en la columna “Grupos”, escogiendo el valor para cada cliente; como se observa en la Figura 2.20, el arreglo realizado para la asignación y creación de grupos.

```
#Grupo 1 con 32 miembros
G1_f_n = F_N_G['Grupos'] == 0
G1_F_N = F_N_G[G1_f_n] #Escojo del df principal los valores que cumplen la condicion
G1_F_N = G1_F_N.drop('Grupos',axis=1)#Elimino la columna Grupos

#Grupo 2 con 2 miembros
G2_f_n = F_N_G['Grupos'] == 1
G2_F_N = F_N_G[G2_f_n]
G2_F_N = G2_F_N.drop('Grupos',axis=1)

#Grupo 3 con 6 miembros
G3_f_n = F_N_G['Grupos'] == 2
G3_F_N = F_N_G[G3_f_n]
G3_F_N = G3_F_N.drop('Grupos',axis=1)

#Grupo 4 con 23 miembros
G4_f_n = F_N_G['Grupos'] == 3
G4_F_N = F_N_G[G4_f_n]
G4_F_N = G4_F_N.drop('Grupos',axis=1)

#Grupo 5 con 37 miembros
G5_f_n = F_N_G['Grupos'] == 4
G5_F_N = F_N_G[G5_f_n]
G5_F_N = G5_F_N.drop('Grupos',axis=1)
```

Figura 2.20: Algoritmo para la creación de Grupos

En la Tabla 2.4, se puede observar la conformación de los grupos después del proceso de agrupamiento; en el Anexo C se puede ver la conformación de los grupos para los días de lunes a viernes y fines de semana.

Tabla 2.4: Grupos Conformados de los días de feriado

Grupo 1 32 clientes	Grupo 2 2 clientes	Grupo 3 6 clientes	Grupo 4 23 clientes	Grupo 5 37 clientes
1_1	1_46	1_12	1_2	1_4
1_8	1_52	1_23	1_3	1_6
1_10		1_30	1_5	1_9
1_14		2_2	1_7	1_11
1_15		2_9	1_13	1_17
1_25		10_1	1_16	1_18
1_27			1_20	1_19
1_31			1_22	1_21
1_32			1_24	1_26
1_34			1_29	1_28
1_36			1_33	1_38
1_40			1_35	1_39
1_41			1_37	1_43
1_42			1_50	1_44
1_45			1_53	1_48
1_47			3_1	1_49
1_51			3_2	1_54
1_57			4_4	1_55
1_59			5_1	1_56
1_60			7_4	1_58
2_4			7_5	2_3
4_1			10_2	2_5
5_2			11_1	2_6
7_1				2_7
7_2				2_8
7_6				2_10
7_7				3_3
8_1				4_2
8_2				4_3
11_2				4_5
11_3				7_3
11_4				7_8
				10_3
				11_5

				12_1
				12_2
				12_3

Finalmente se realiza un estudio estadístico, para conocer cómo se muestran los valores de desviación estándar; mostrando que los valores del grupo 5 que posee 37 miembros y es el que más individuos tiene, la desviación oscila entre el 15% y 30%, como se muestra en la Figura 2.21, donde se observa los primeros y últimos resultados de la función estadística empleada.

```
0       0.237017
1       0.236771
2       0.238668
3       0.238570
4       0.231430
          ...
1387    0.305110
1388    0.289623
1389    0.267932
1390    0.247329
1391    0.239205
Length: 1392, dtype: float64
```

Figura 2.21: Cálculo de la desviación estándar del grupo 5 de los días de feriado

Para observar los resultados de la desviación estándar de los grupos de los días de lunes a viernes y fines de semana, ver Anexo A.

2.2.9. Creación de Clientes Fraudulentos

Con la falta de datos de curvas diarias que presenten consumos anormales por parte de la EEASA y el departamento de Telemedición, se crean patrones que sub registren energía para simular fraudes típicos que se pueden dar en los clientes especiales de la distribuidora.

Las lecturas de medición de la energía activa suministrada pueden verse afectadas total o parcialmente mediante ataques cibernéticos, manipulaciones mal intencionadas en los medidores inteligentes o conexiones ilegales que cambian intencionalmente las lecturas que realiza la AMI mediante el uso de un puente (byPass) en la bornera del medidor o una

doble acometida, pudiendo utilizar un interruptor de activación automática que permite el subregistro de energía por ventanas de tiempo permanentes o aleatorias, obteniendo un porcentaje de energía no registrada por el cliente que lleve a cabo el fraude [31]. Por lo que para la creación de estos patrones se busca simular este tipo de comportamientos con un conjunto de variables aleatorias, aparentando los diferentes tipos de subregistro que se pueden crear al emplear estas actividades mal intencionadas y afectando directamente a la distribuidora y aumentando las pérdidas no técnicas de la misma.

- ***Fraude tipo 1:*** Disminución proporcional constante en el tiempo.

  $$\hat{p}_{t_i,n} = \nu\, p_{t_i,n};\ \nu \in [\min, \max] \qquad \textbf{Ecuación 2.2}$$

  donde $\hat{p}_{t_i,n}$ es el valor de consumo de energía del cliente n en el instante t_i y $\hat{p}$ el consumo modificado con fraude. Para cada cliente n se asigna un valor fijo de ν con distribución uniforme $\nu \sim \mathrm{U}$ y una fecha de comienzo de fraude $t_{f,n}$ aleatoria dentro de: los 500 días de datos de los lunes a viernes, los 173 días de datos de fines de semana y los 58 días de datos de los feriados.

- ***Fraude tipo 2:*** Disminución proporcional en ventanas de tiempo con franja horaria en el máximo consumo (horas pico).

  Este fraude modela el reporte nulo de consumo y es un caso particular del fraude tipo 3 con $\alpha > 0$.

  $$\hat{p}_{t_i,n} = \frac{\delta_{t_i}\ p_{t_i,n}}{\delta_{t_i}} = \begin{cases} > 0 \ \ si\ t_{start} \leq t_i \leq t_{start} + l \\ 1\ en\ otro\ caso \end{cases} \qquad \textbf{Ecuación 2.3}$$

  En este modelo se utiliza $t_{min} = 6\,pm\ \ y\ \ t_{max} = 9\,pm$, para que la ventana de fraude sea dentro del horario de cresta de la demanda energética.

- ***Fraude tipo 3:*** Disminución proporcional en ventanas de tiempo con franja horaria indistinta y aleatoria.

  $$\hat{p}_{t_i,n} = \frac{\delta_{t_i}\ p_{t_i,n}}{\delta_{t_i}} = \begin{cases} \alpha\ si\ t_{start} \leq t_i \leq t_{start} + l \\ 1\ en\ otro\ caso \end{cases} \qquad \textbf{Ecuación 2.4}$$

donde α modela el olvido en el accionamiento del fraude de algunos días, tomando valores $[\nu, 1]$ con distribución de Bernoulli.

$$\alpha = (\mathrm{F} - 1)\nu + \mathrm{F} \qquad \textbf{Ecuación 2.5}$$

El fraude se comete diariamente en una ventana de tiempo. $t_{start} + l$ modelan el ruido en comienzo y duración de la ventana. $t_{start} \sim N(\mu_{ini},\ \sigma_{ini})$, se asume el valor medio de la distribución fijo $(\mu_{ini}[n])$ para cada cliente y asignado en forma aleatoria con probabilidad uniforme dentro de un rango horario $\mu_s \sim U(t_{min}, t_{max})$. La duración de accionamiento también es una variable aleatoria con distribución gaussiana $l \sim N(\mu_d\,[i], \sigma_d)$. El valor medio $\mu_d\,[i]$ es fijo a lo largo de los días y es asignado a cada cliente utilizando una distribución de probabilidad uniforme $\mu_d \sim U(l_{min}, l_{max})$.

Para establecer este tipo de curvas se utilizaron los parámetros detallados en la tabla 2.5, en la cual se detalla los valores para crear la base sintética que contiene las curvas con subregistro, donde el cálculo de la desviación estándar de cada hora en todo el conjunto de datos previamente clasificado, brinda el valor de desviación estándar máxima, siendo este el margen mínimo de variación para crear las curvas diarias fraudulentas para cada grupo y conformar así la base sintética con subregistros de energía.

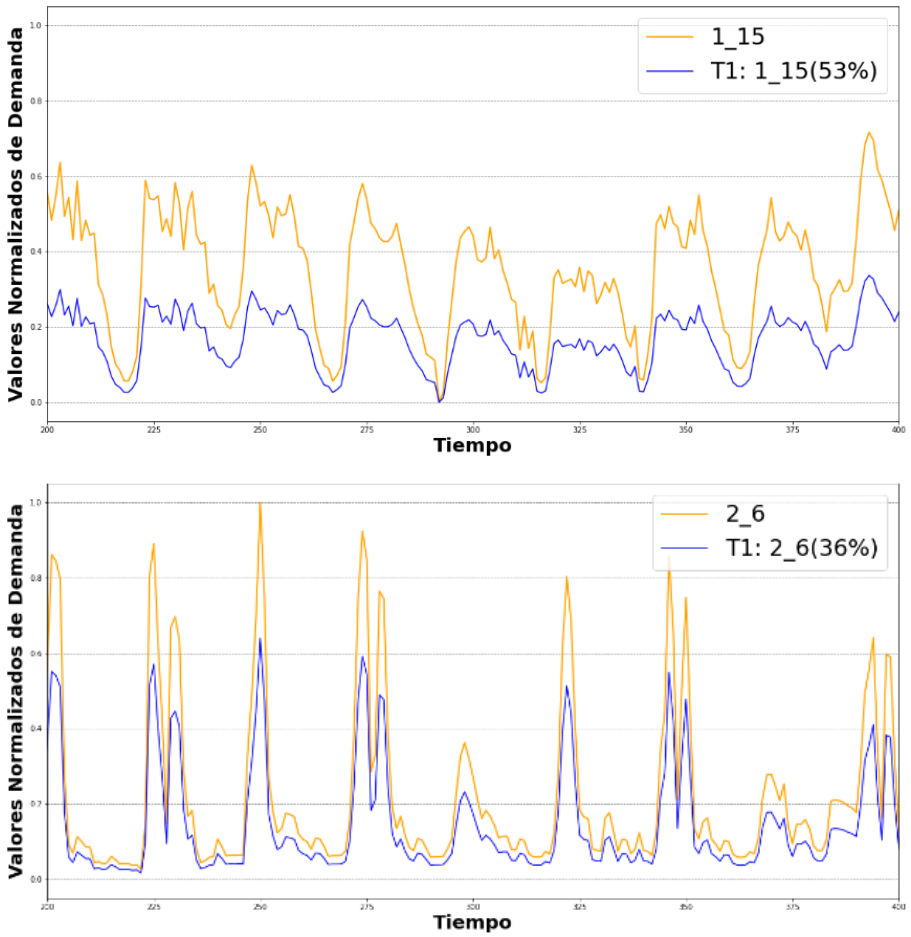
Tabla 2.5: Creación de las Curvas Fraudulentas para cada grupo de días

	Desviación Estándar		**Curvas Fraudulentas**		
	MIN	MAX	% mínimo de variación permitido	% mínimo de variación elegido	% máximo de variación elegido
Grupo de días de lunes a viernes	0.172888139	0.303566006	30.35%	35.00%	85.00%
Grupo de los días de fin de semana	0.228457341	0.32703843	32.70%	35.00%	85.00%
Grupo de los días de feriado	0.210337241	0.344457664	34.45%	35.00%	85.00%

Los valores de porcentaje de variación para las curvas fraudulentas serán escogidos de forma aleatoria, formando diferentes patrones de curvas anormales para cada tipo de fraude; éstas tendrán el detalle de mostrar el porcentaje de variación realizada junto al

nombre asignado para cada cliente y el tipo de fraude que está simulando, con el fin de identificar el porcentaje de variación de la curva anormal de la curva real.

Las curvas fraudulentas del Tipo 1, para el grupo de clientes de los días de feriado se muestran en la Figura 2.22, en la cual se puede observar 6 clientes diferentes con curvas engañosas y el porcentaje de variación que tienen con respecto a las curvas reales. Las curvas diarias con subregistro son las mostradas en color azul, mientras que las curvas reales se muestran en color naranja.



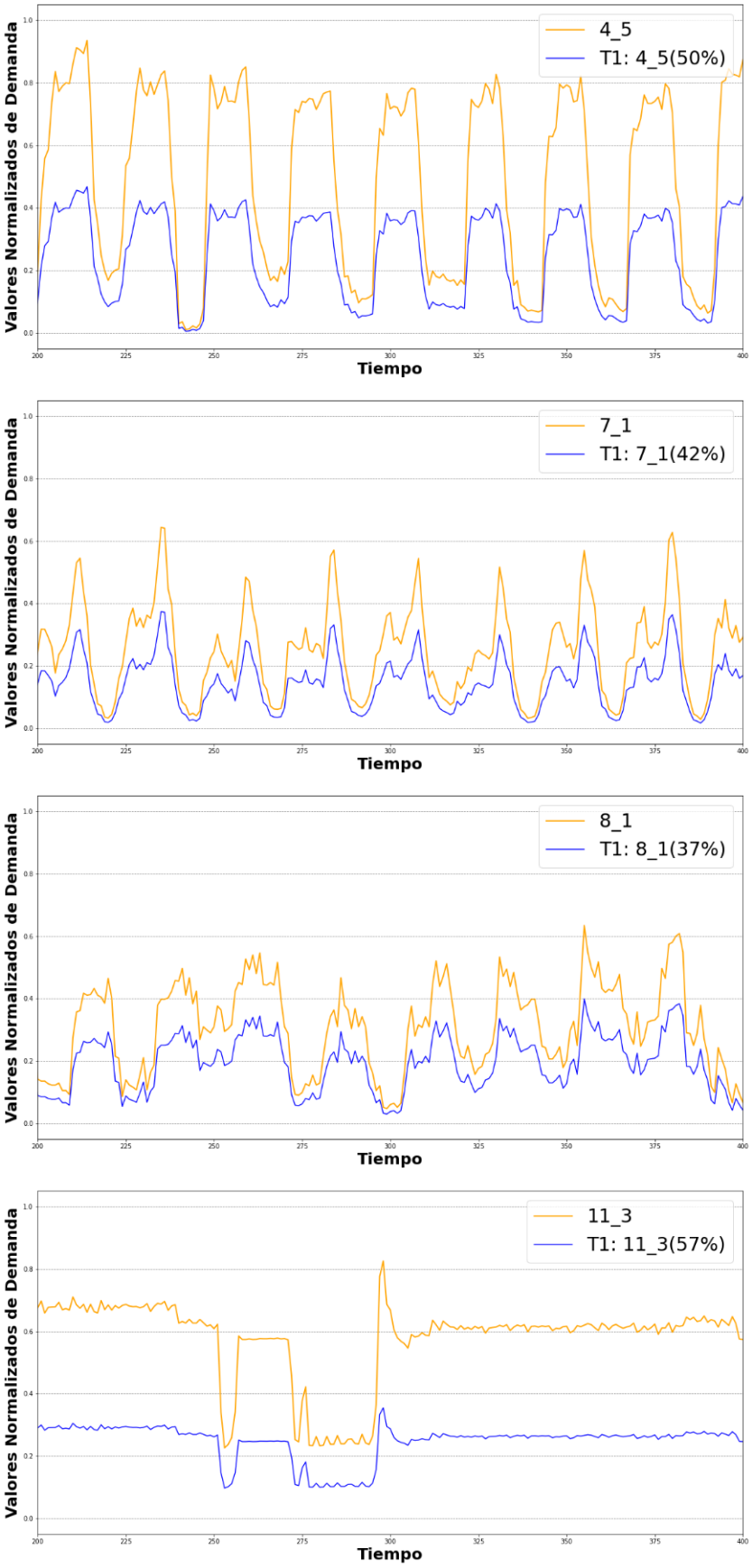
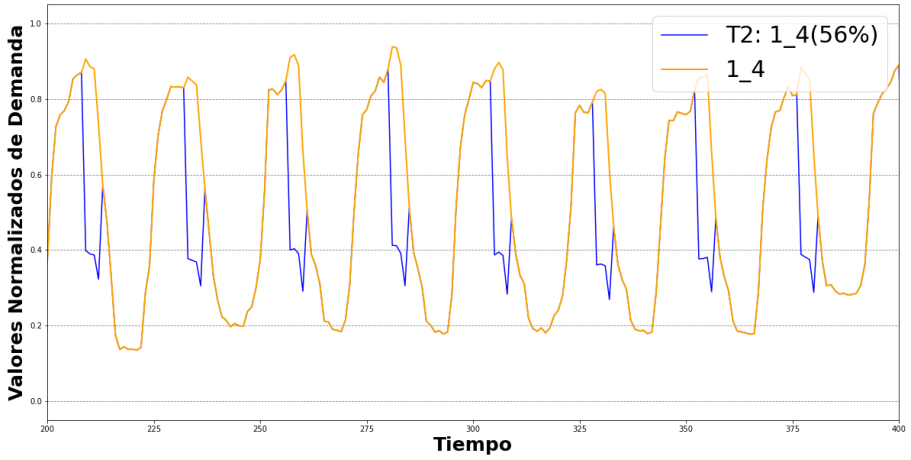
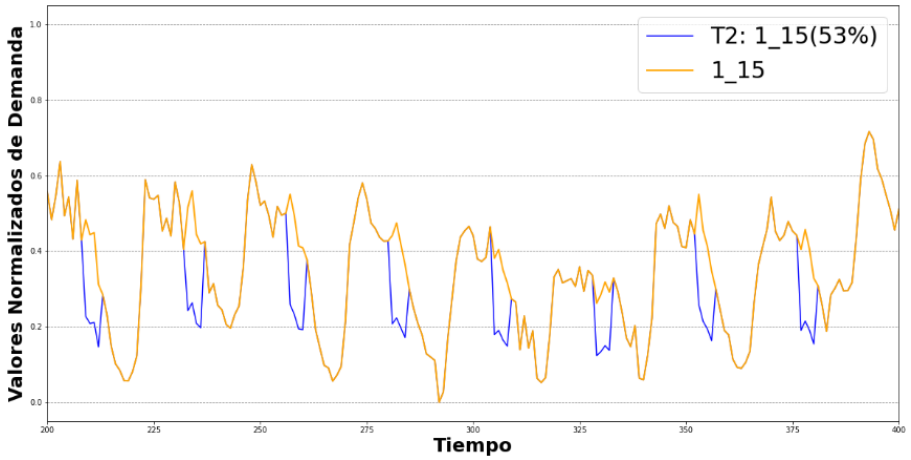
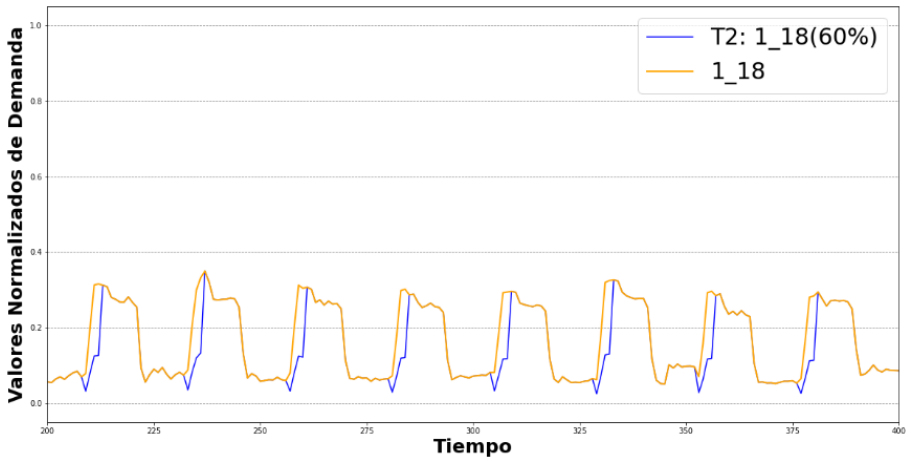


Figura 2.22: Curvas Fraudulentas Tipo 1 del grupo de feriado

Las curvas fraudulentas del Tipo 2, para el grupo de clientes de los días de feriado se muestran en la Figura 2.23, en la cual se puede observar 6 clientes diferentes con curvas engañosas y el porcentaje de variación que tienen durante ventanas de tiempo fijas entre las 6 pm y las 9 pm de cada día con respecto a las curvas reales. Las curvas diarias con subregistro son las mostradas en color azul, mientras que las curvas reales se muestran en color naranja.







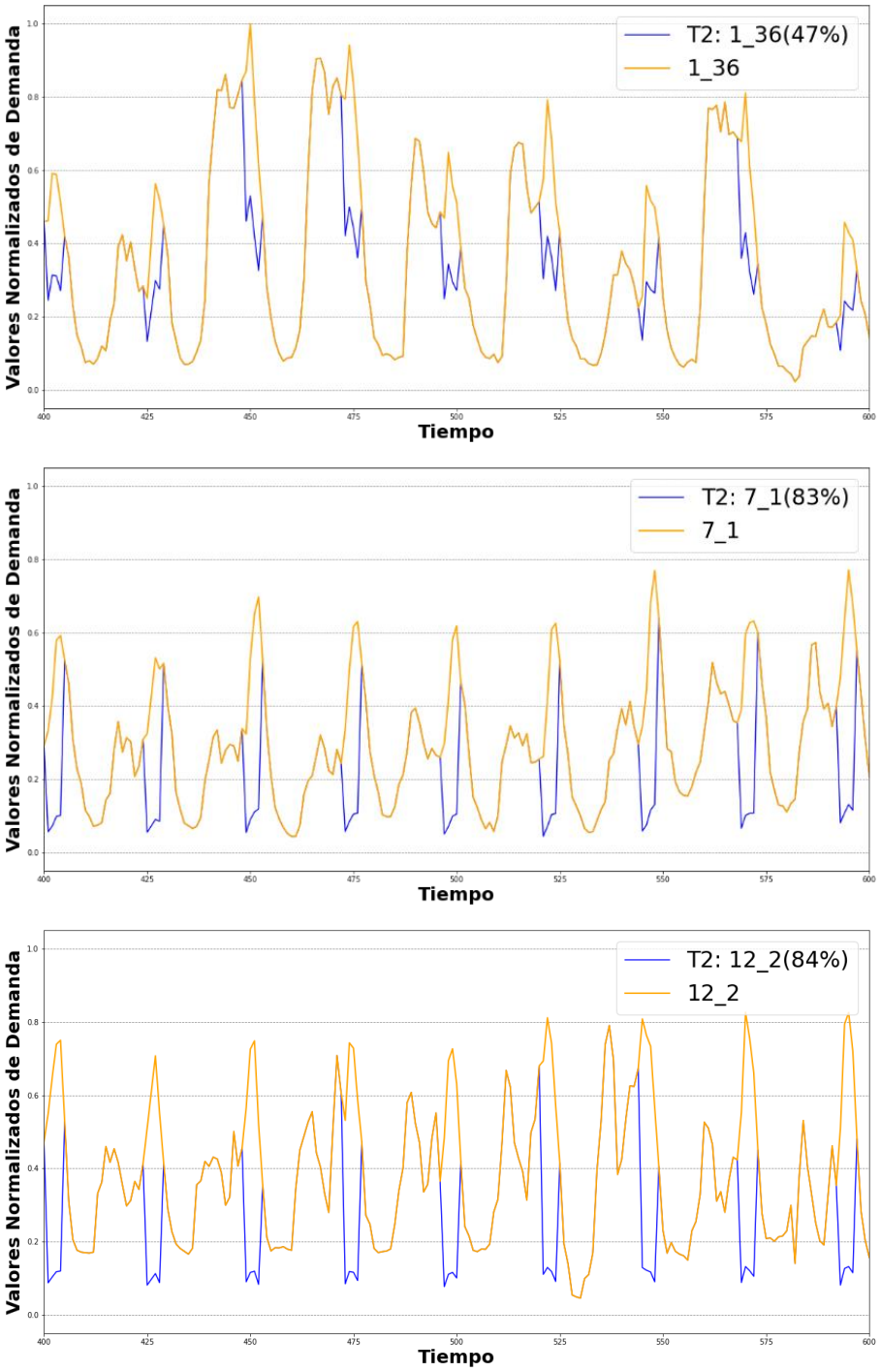
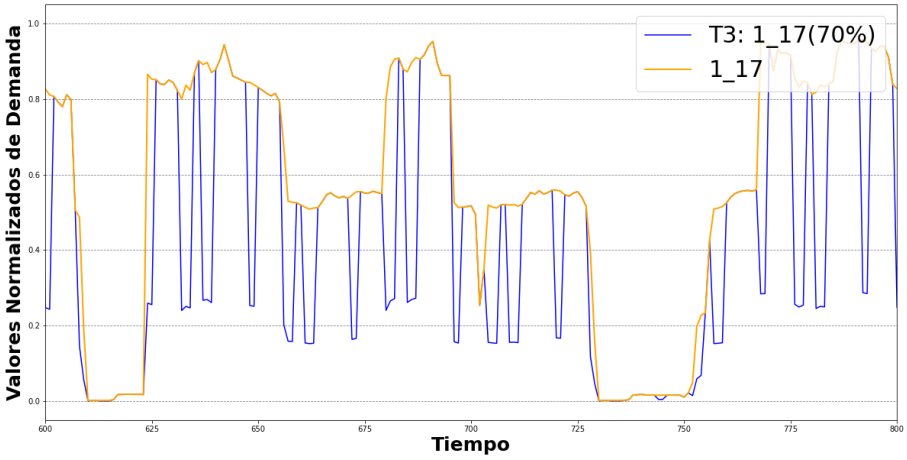
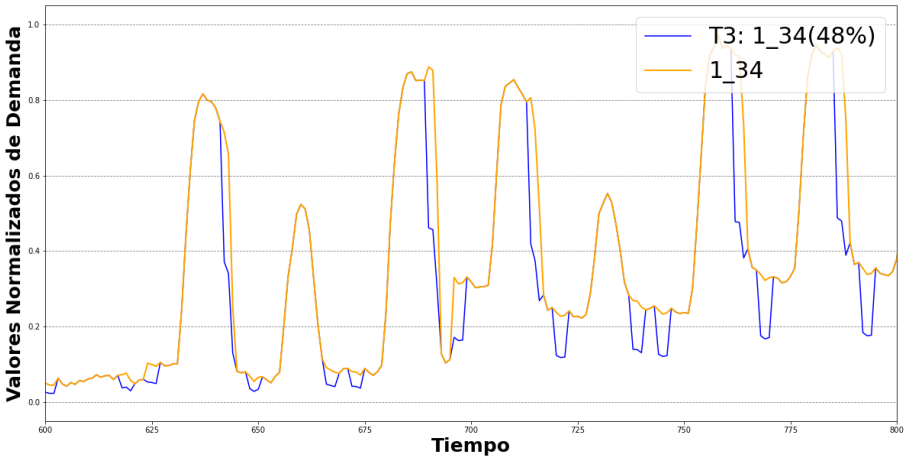
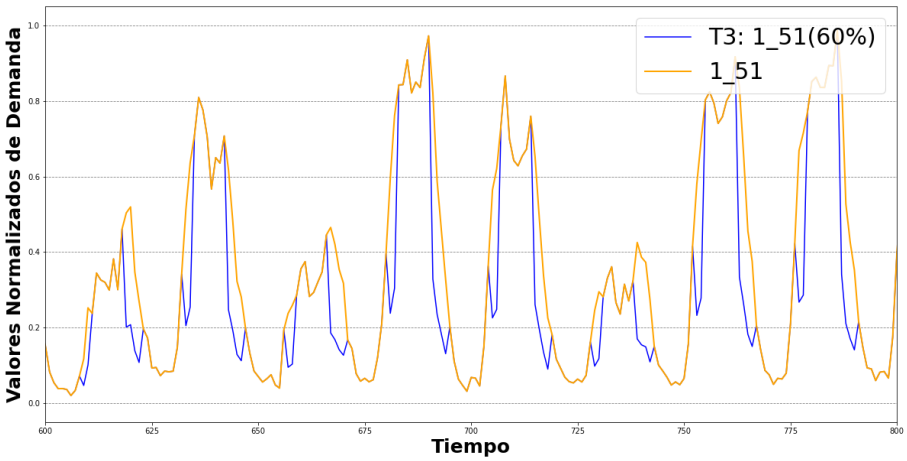


Figura 2.24: Curvas Fraudulentas Tipo 2 del grupo de feriado

Las curvas fraudulentas del Tipo 3, para el grupo de clientes de los días de feriado se muestran en la Figura 2.24, en la cual se puede observar 6 clientes diferentes con curvas engañosas y el porcentaje de variación que tienen durante ventanas de tiempo aleatorias en el día con respecto a las curvas reales. Las curvas diarias con subregistro son las mostradas en color azul, mientras que las curvas reales se muestran en color naranja.







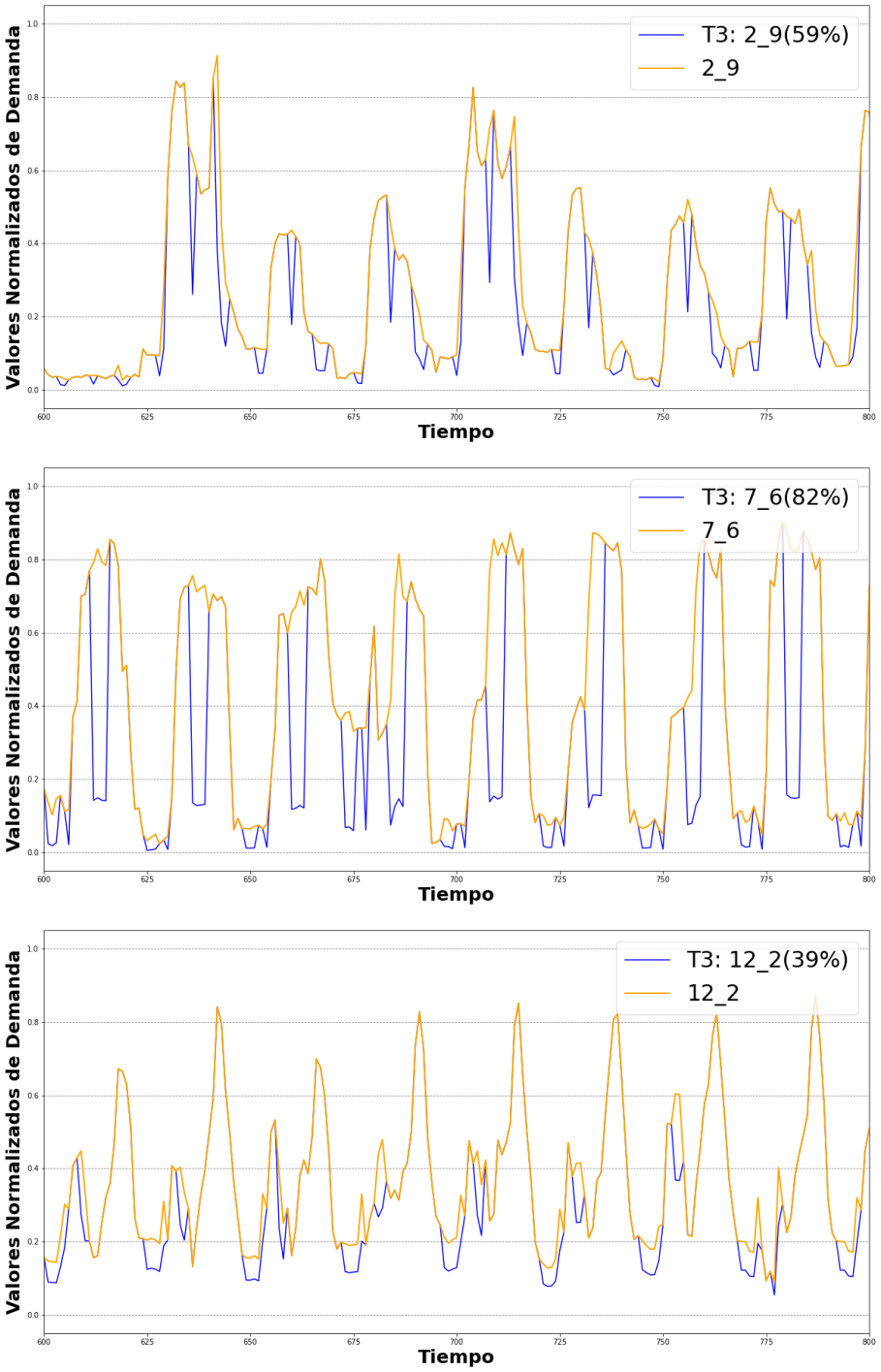


Figura 2.24: Curvas Fraudulentas Tipo 3 del grupo de feriado

En el Anexo D, se puede observar las curvas generadas para los días de lunes a viernes y fines de semana.

2.2.10. Creación de la Red Neuronal en KNIME Analytics Platform

Para el uso de este software en la creación de la red neuronal es importante realizar la instalación del paquete “Python Deep Learning” con las librerías Keras y Tensor Flow 2, que sean compatibles para el Python 3.6.13, que viene por default en la instalación de Conda en su última versión (Conda 4.12.0). Ya enlazados estos programas, como muestra la figura 2.25, se puede utilizar las librerías para la construcción de la Red neuronal.

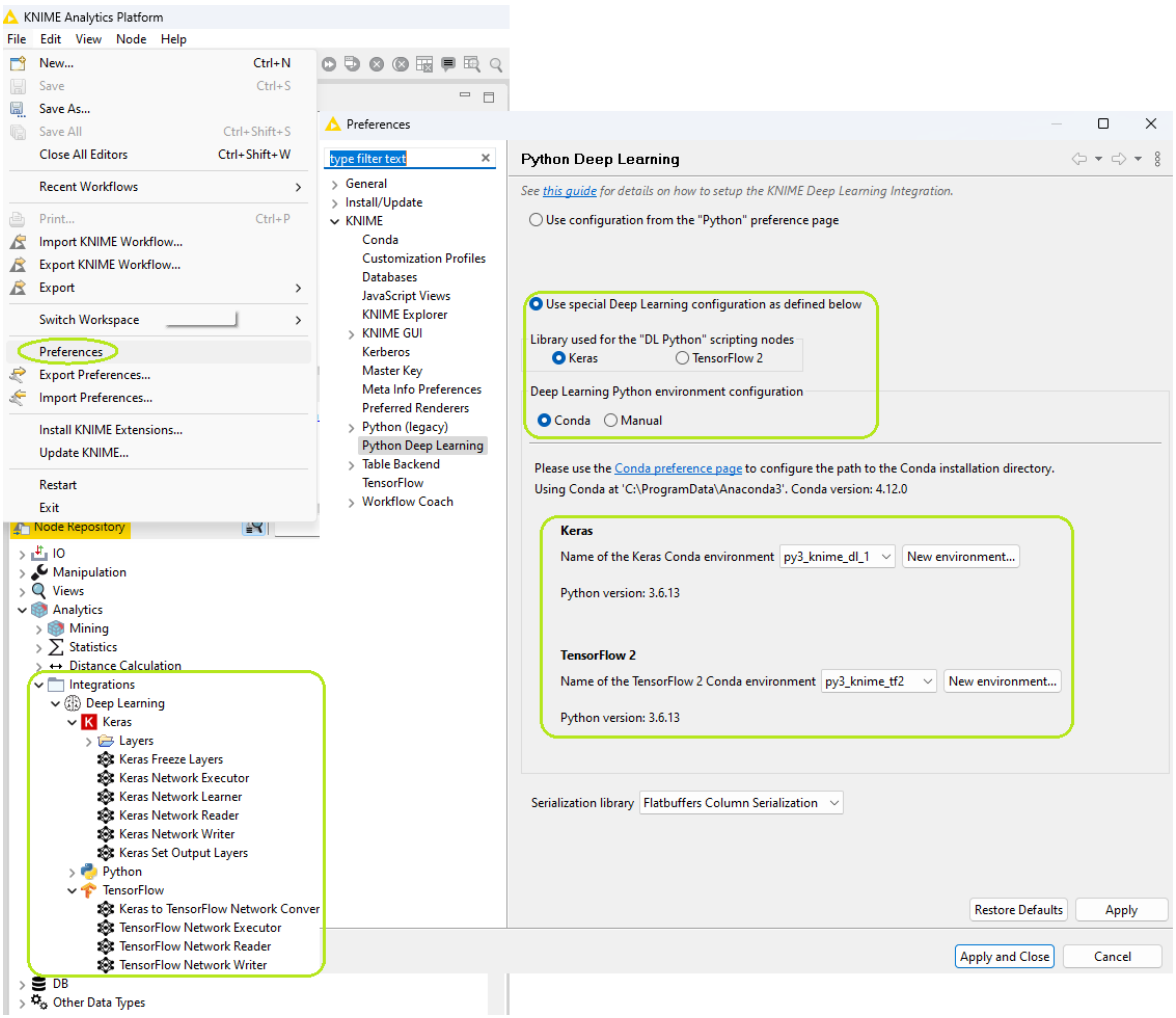


Figura 2.25: Enlace KNIME – Python y librerías Deep Learning.

2.2.10.1. Lectura de Archivos en KNIME

Para llevar a cabo la lectura de la data tratada con formato .xlsx, del *Repositorio de Nodos* se extrae el nodo ***Excel Reader*** y se lo lleva al entorno de trabajo, el mismo que se lo configura como muestra la Figura 2.26, al dar doble clic sobre el nodo, en el cual se busca al archivo deseado y con la ayuda de la pestaña *Preview* se observa los datos que serán leídos, las otras pestañas son las configuradas por defecto.

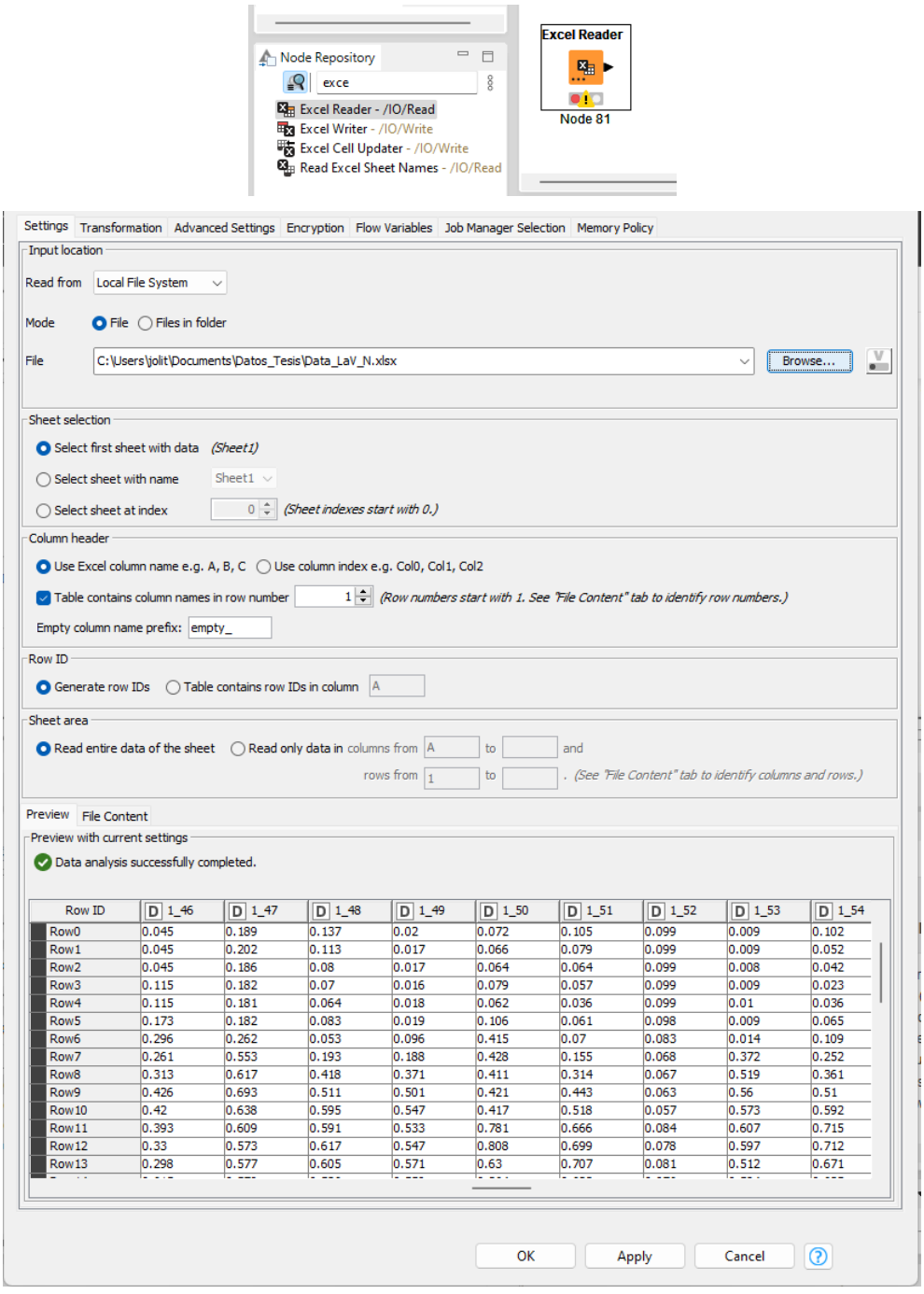


Figura 2.26: Configuración Nodo Excel Reader

2.2.10.2. Selección, Filtrado y Arreglo de Datos para la Red Neuronal

Del *Repositorio de Nodos*, extraer los nodos de Filtro de Columna, Transpuesta, Manejo de Colores y Partición, como muestra la figura 2.27, los cuales se configuran de acuerdo con los grupos conformados en la clusterización de los grupos de los días de lunes a viernes, fines de semana y feriados.

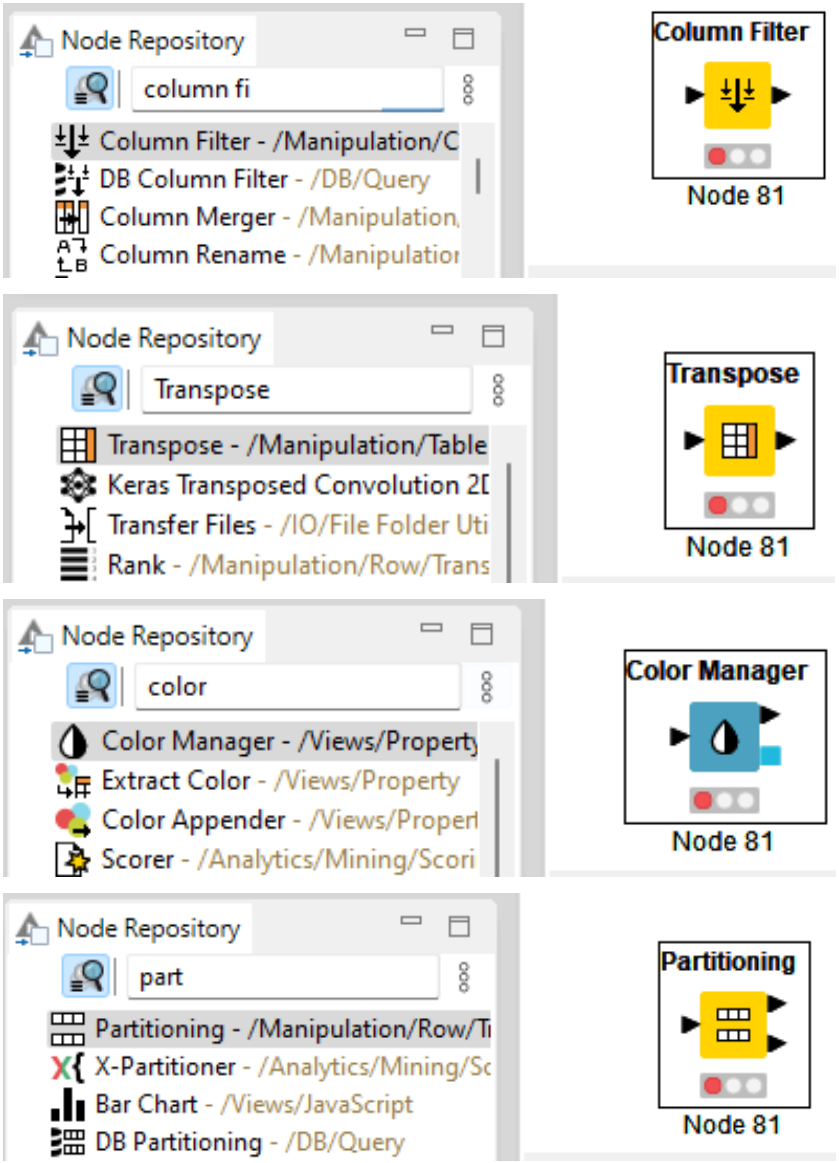


Figura 2.27: Nodos Filtro de Columna, Transpuesta, Manejo de Colores y Partición

El Nodo *Column Filter*, nos permitirá elegir las filas donde se encuentran los clientes que conforman cada grupo, con ayuda de la Tabla 2.4, se va seleccionado los clientes que

conforman cada grupo, como se muestra en la Figura 2.28, donde se observa la conformación del Grupo 1 de los días de feriado.

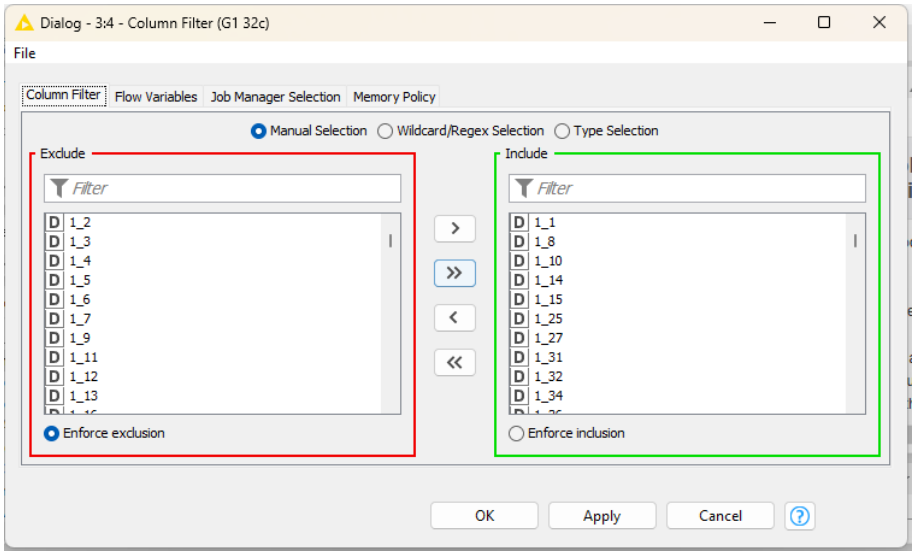


Figura 2.28: Configuración del Nodo Column Filter

El Nodo *Transpose*, como su nombre lo indica es el siguiente proceso por seguir para transponer los datos filtrados, dejando la configuración por defecto, como muestra la Figura 2.29, en la cual se especifica que se realice la acción en base a todos los datos filtrados.

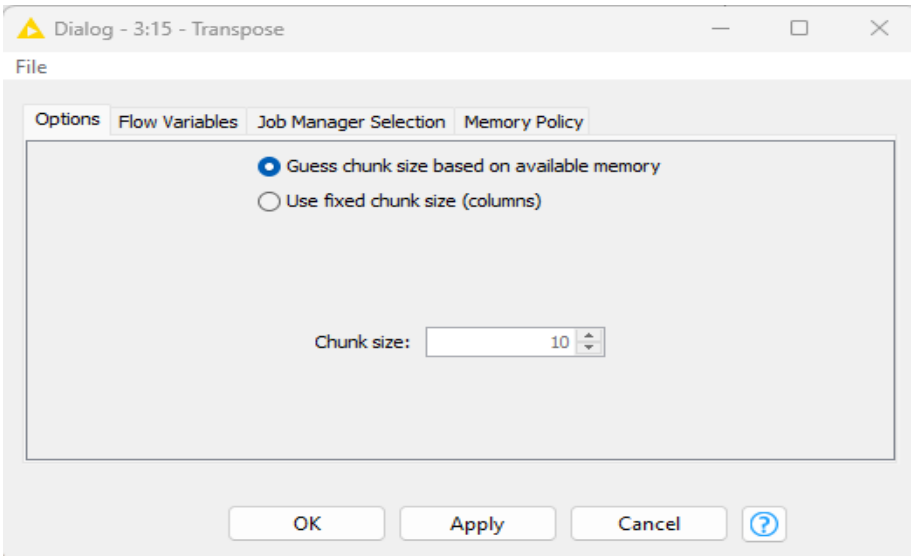


Figura 2.29: Configuración del Nodo Transpose

El Nodo *Color Manager*, como su nombre describe, permite colorear e identificar los datos de las curvas reales, de las curvas creadas como fraudulentas de los diferentes clientes, como se muestra la Figura 2.30, es importante mencionar que la base de datos sintética y la base de datos real, se encuentran mezcladas en el mismo archivo cargado en el nodo de lectura de Excel, además previamente a la extracción del archivo, existe una columna que identifica con valores booleanos de 0 (datos reales) y 1 (data). La columna 1392 es la etiqueta para el aprendizaje supervisado que será direccionado en la Red Neuronal. Los datos de la columna 0 a 1391 los datos para entrenar la red

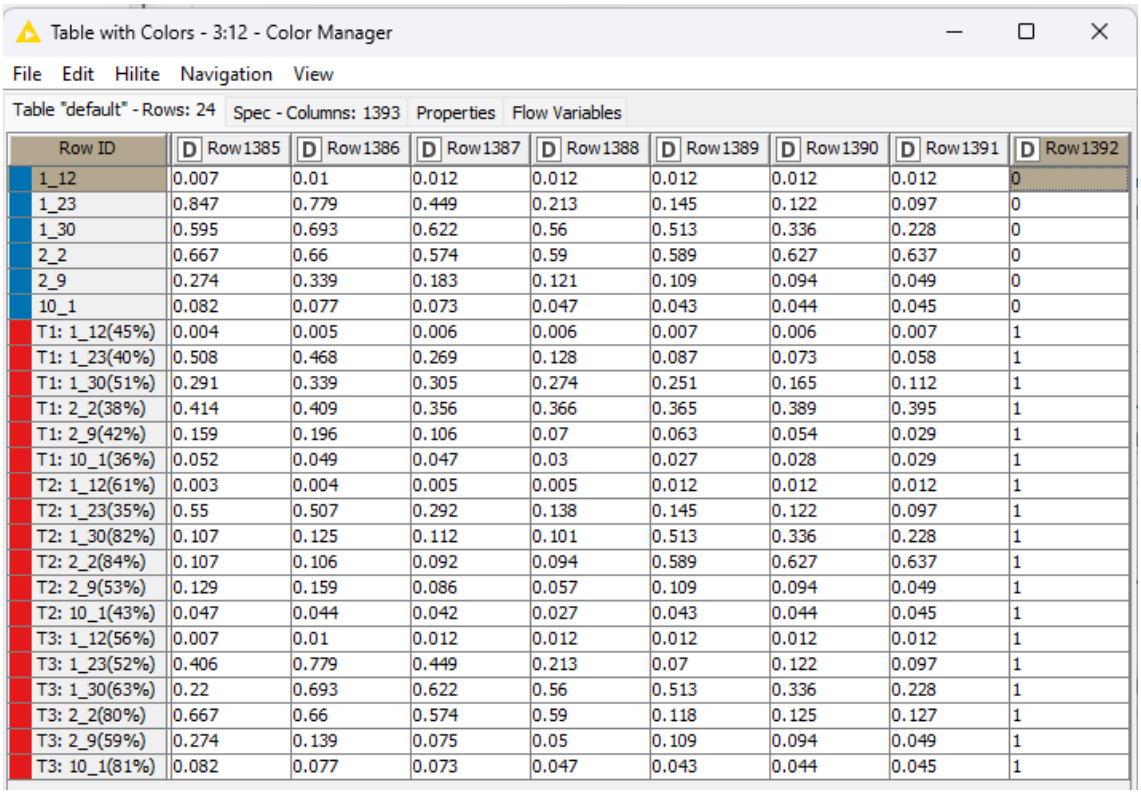


Table with Colors - 3:12 - Color Manager

File Edit Hilite Navigation View

Table "default" - Rows: 24 Spec - Columns: 1393 Properties Flow Variables

Row ID	Row1385	Row1386	Row1387	Row1388	Row1389	Row1390	Row1391	Row1392
1_12	0.007	0.01	0.012	0.012	0.012	0.012	0.012	0
1_23	0.847	0.779	0.449	0.213	0.145	0.122	0.097	0
1_30	0.595	0.693	0.622	0.56	0.513	0.336	0.228	0
2_2	0.667	0.66	0.574	0.59	0.589	0.627	0.637	0
2_9	0.274	0.339	0.183	0.121	0.109	0.094	0.049	0
10_1	0.082	0.077	0.073	0.047	0.043	0.044	0.045	0
T1: 1_12(45%)	0.004	0.005	0.006	0.006	0.007	0.006	0.007	1
T1: 1_23(40%)	0.508	0.468	0.269	0.128	0.087	0.073	0.058	1
T1: 1_30(51%)	0.291	0.339	0.305	0.274	0.251	0.165	0.112	1
T1: 2_2(38%)	0.414	0.409	0.356	0.366	0.365	0.389	0.395	1
T1: 2_9(42%)	0.159	0.196	0.106	0.07	0.063	0.054	0.029	1
T1: 10_1(36%)	0.052	0.049	0.047	0.03	0.027	0.028	0.029	1
T2: 1_12(61%)	0.003	0.004	0.005	0.005	0.012	0.012	0.012	1
T2: 1_23(35%)	0.55	0.507	0.292	0.138	0.145	0.122	0.097	1
T2: 1_30(82%)	0.107	0.125	0.112	0.101	0.513	0.336	0.228	1
T2: 2_2(84%)	0.107	0.106	0.092	0.094	0.589	0.627	0.637	1
T2: 2_9(53%)	0.129	0.159	0.086	0.057	0.109	0.094	0.049	1
T2: 10_1(43%)	0.047	0.044	0.042	0.027	0.043	0.044	0.045	1
T3: 1_12(56%)	0.007	0.01	0.012	0.012	0.012	0.012	0.012	1
T3: 1_23(52%)	0.406	0.779	0.449	0.213	0.07	0.122	0.097	1
T3: 1_30(63%)	0.22	0.693	0.622	0.56	0.513	0.336	0.228	1
T3: 2_2(80%)	0.667	0.66	0.574	0.59	0.118	0.125	0.127	1
T3: 2_9(59%)	0.274	0.139	0.075	0.05	0.109	0.094	0.049	1
T3: 10_1(81%)	0.082	0.077	0.073	0.047	0.043	0.044	0.045	1

Figura 2.30: Uso del Nodo Color Manager en la data del Grupo 3 de los días de feriado

El Nodo *Partitioning*, es el encargado de segmentar los datos filtrados y coloreados en el grupo; este nodo es el indicado para seleccionar los datos de entrenamiento (70%), validación (15%) y prueba (15%) para la red neuronal. Para realizar este proceso se necesitarán dos nodos de partición, en el cual el primero será configurado con la partición 70 – 30 y el segundo unido al primer nodo de partición con la configuración del 50 – 50, para obtener los 15% de datos para la validación y prueba antes descrita, en la Figura 2.31, se muestra la configuración descrita anteriormente.

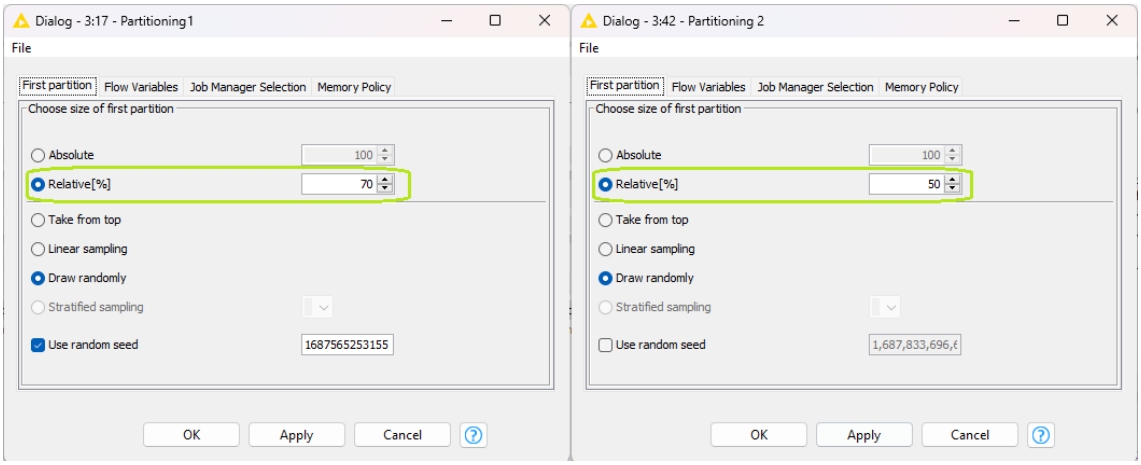


Figura 2.31: Configuración de los Nodos Partiotioning

Es importante mencionar que en la parte inferior de cada nodo, existe un pequeño segmento de colores, que muestra los colores rojo (Nodo no configurado), amarillo (Nodo Configurado) y Verde (Nodo Ejecutado sin problemas); los cuales indican el estado de ejecución en el que se encuentra el Nodo, también existe la figura de alarma que se puede presentar en el estado Rojo o Amarillo, que indica la especificación errónea, faltante o si se debe realizar otra configuración extra a las descritas para ejecutarlo. Finalmente, el último estado es la de la barra de procesos, en donde se puede observar el porcentaje actual en la que el nodo va ejecutándose, como se muestra en la Figura 2.32, en la que se indica los estados comunes de ejecución de un Nodo. La ejecución del Nodo se lo realiza con la tecla F7 o dando clic derecho sobre él y escoger la opción "*Execute*".

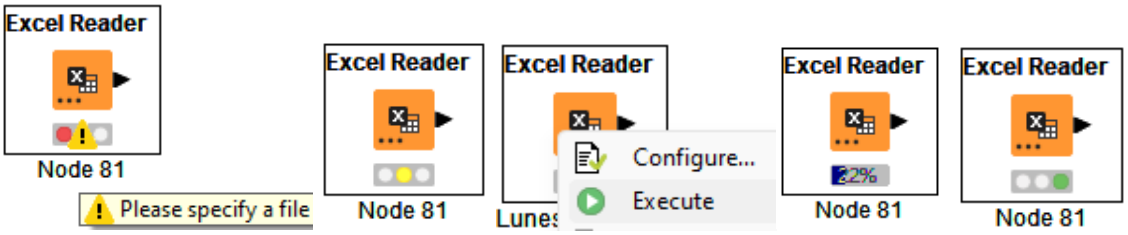


Figura 2.32: Ejecución de un Nodo

El proceso de configuración de los nodos descritos, se realiza para los cinco grupos de los días de feriado, fines de semana y de los lunes a viernes, exceptuando el nodo de partición, ya que este nodo se lo configura solo para el grupo que contenga uno de los mayores números de clientes como se indica en la Figura 2.33, donde se muestra la configuración

completa del entorno de trabajo con los semáforos en verde de todos los nodos que forman el entorno de trabajo, previo a la construcción de la Red Neuronal.

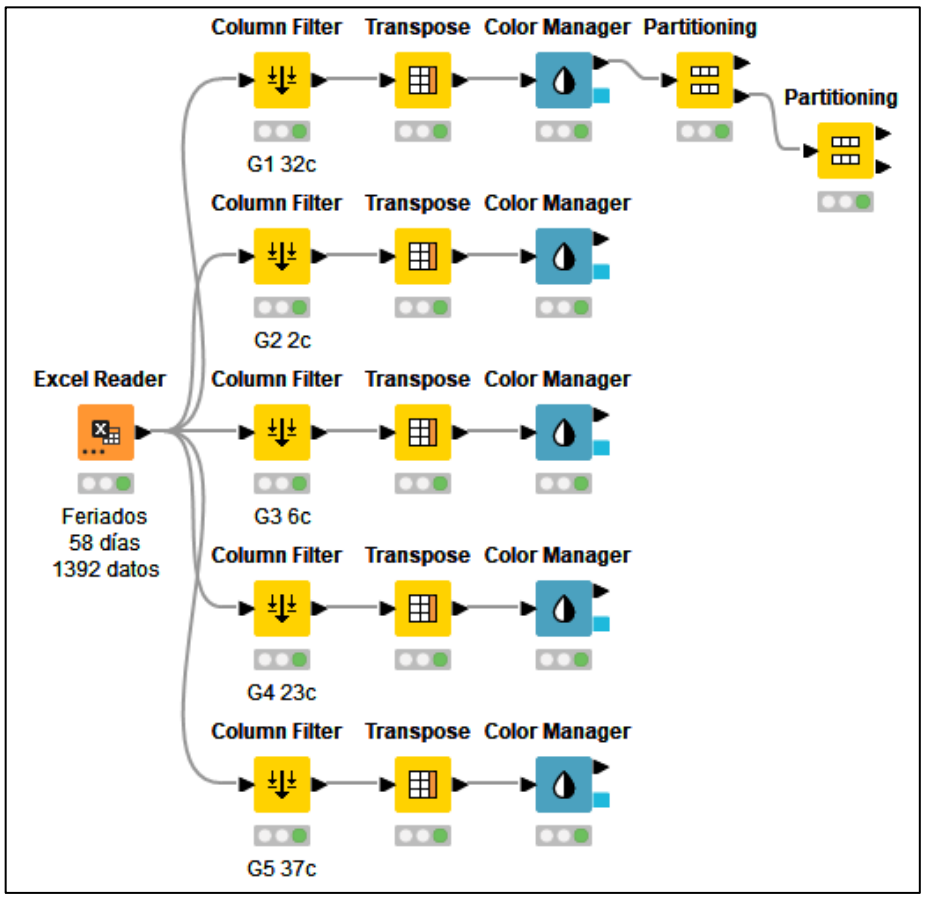


Figura 2.33: Arreglo de Nodos para la Red Neuronal

2.2.10.3. Construcción y Configuración del Modelo de la Red Neuronal

Para la construcción de las capas de la red, se debe dirigir al *Node Repository* a la librería *Analytics/Integrations/DeepLearning/Keras/Layers*, como se indicó en la Figura 2.25 y buscar los Nodos: *Keras Input Layers*, *Keras Dense Layers*, *Keras Network Learner*, *Keras Network Executor* y un Nodo de visualización de resultados que será *Line Plot (local)*. Para los cuales se realizará las siguientes configuraciones, para formar la red neuronal densamente conectada mostrada en la Figura 2.34.

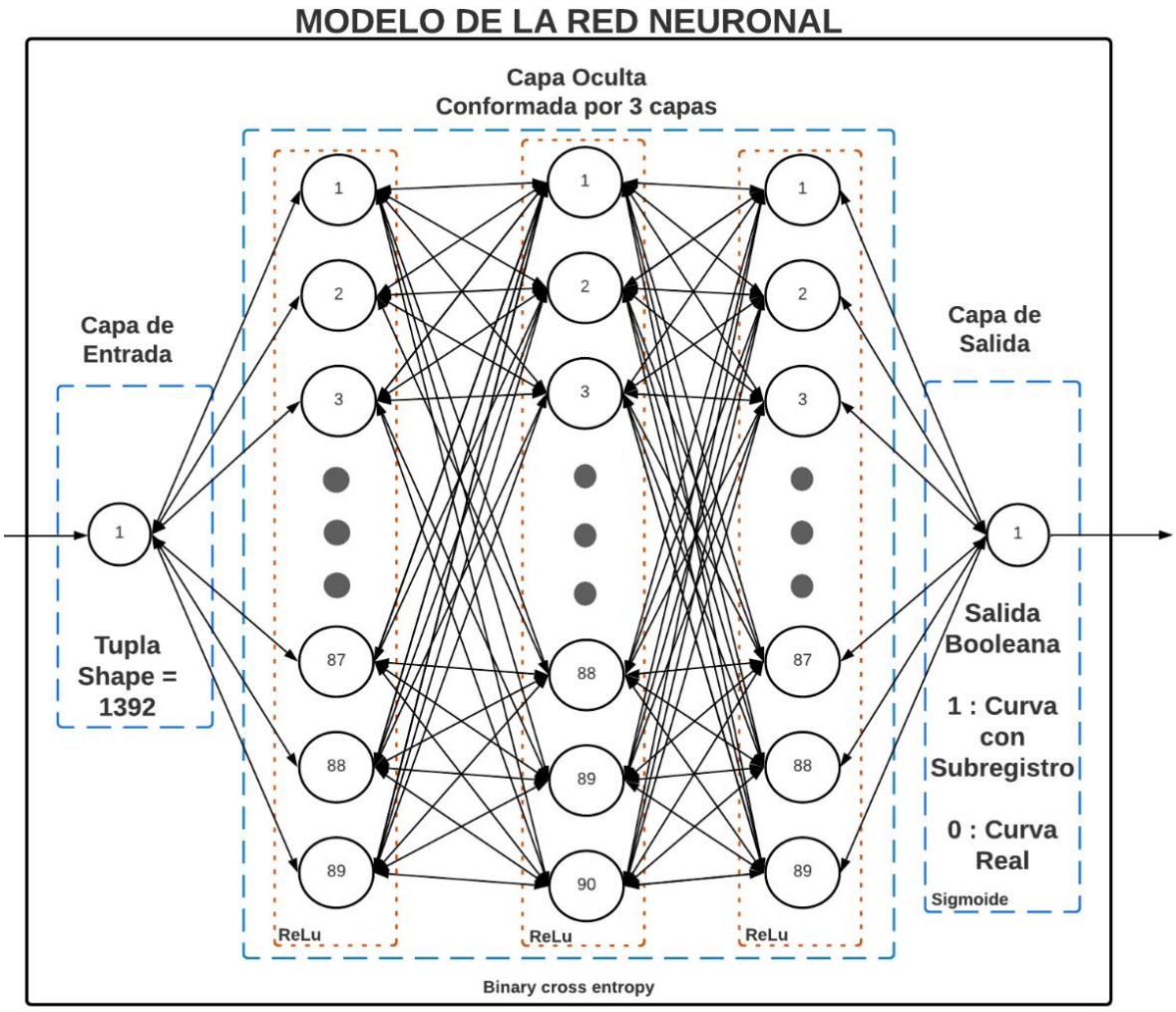


Figura 2.34: Modelo de Red Diseñada para el Grupo de los días Feriados

2.2.10.3.1. Nodo Keras Input Layers

Es el Nodo para especificar la capa de Entrada de la Red Neuronal, donde se define el *Shape* de Entrada, el Batch de la Red Neuronal, el tipo de datos que ingresaran a la Red y el canal de ingreso de la data, como se muestra en la Figura 2.35, la cual revela el Shape de 1392 atributos que son las tuplas contenidas en cada cliente (58 días x 24 horas de lecturas diarias por cliente), de los días de feriado; el tipo de data, los cuales son del tipo “Float 32” y en el Data Format el último canal habilitado para el entrenamiento, que representa la conexión que se realiza en el Nodo *Keras Network Executor*; la pestaña Name Prefix se la habilita si se desea asignar un nombre específicos a la Capa de Entrada y la pestaña de Batch Size se lo configurará en el *Keras Network Layer*.

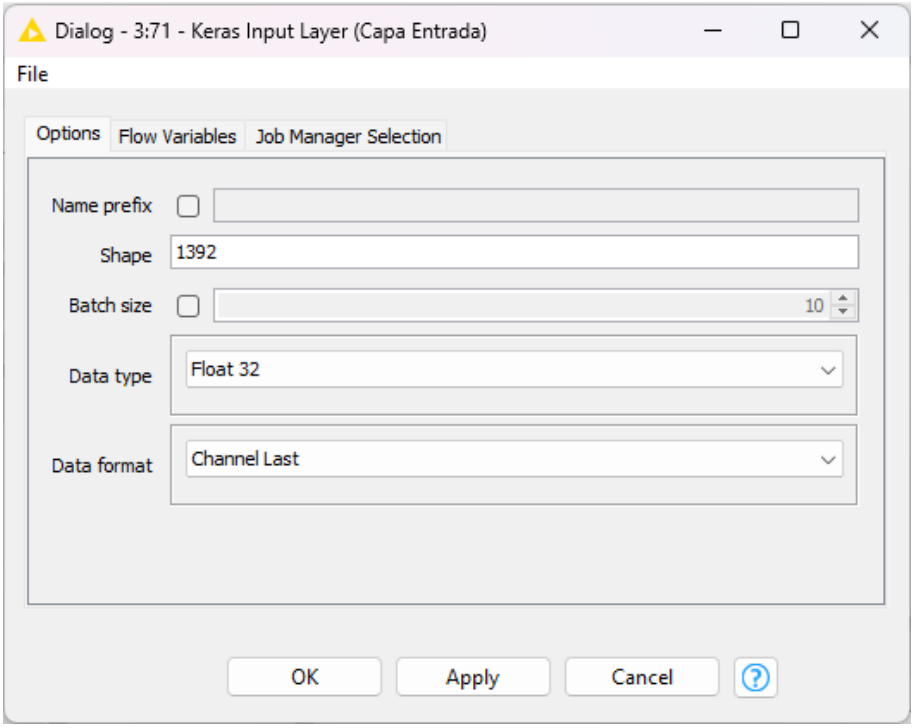


Figura 2.35: Configuración Nodo Keras Input Layers

2.2.10.3.2. Nodo Keras Dense Layers

Un nodo de estos representa una de las tres capas ocultas, por lo que se debe parametrizar en la pestaña *Options*, de la siguiente forma, en *Input tensor* se debe verificar que sea la salida de la capa preliminar de la red para utilizarla como entrada de la capa en configuración; en *Units* especificar el número de neuronas que conforman la capa, para las que se pondrán 89 para la capa oculta 1 y 3, y 90 para la capa oculta 2, ya que ingresarán 89 patrones de curvas reales y fraudulentas; finalmente en la opción *Activation Function* se definirá la Función ReLU, la cual es la elegida para realizar redes neuronales de clasificación; como se muestra en la Figura 2.36. Las otras pestañas del Nodo quedan definidas por defecto.

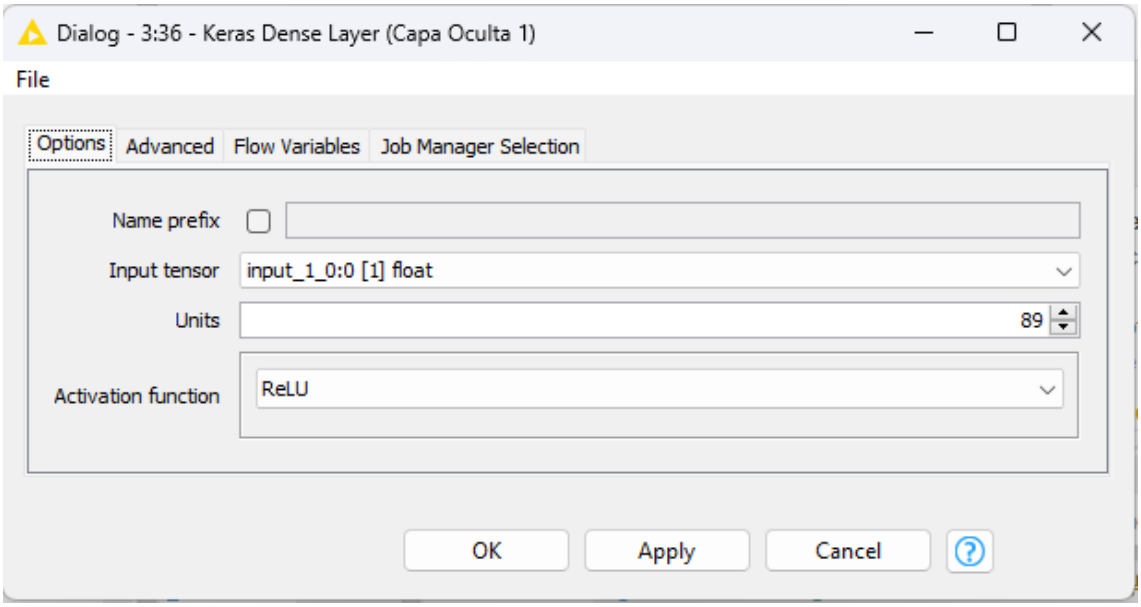


Figura 2.36: Configuración de Keras Dense Layer como capas ocultas

Este proceso se realiza de la misma forma con las 3 capas, con la modificación en el número de neuronas que conforman cada capa.

Para la configuración de la capa de salida, se utiliza el mismo Nodo, pero con las configuraciones diferentes; en la opción *Units* escribir el valor de 1, ya que el modelo tendrá una neurona en su capa de salida, con valores de salida de 0 (para curvas normales) o 1 (para curvas que subregistran); y para forzar esta salida, en la opción *Activation Function* se elige la *Sigmoid*, como se observa en la Figura 2.37, terminando de definir la capa de salida del Modelo de la Red Neuronal.

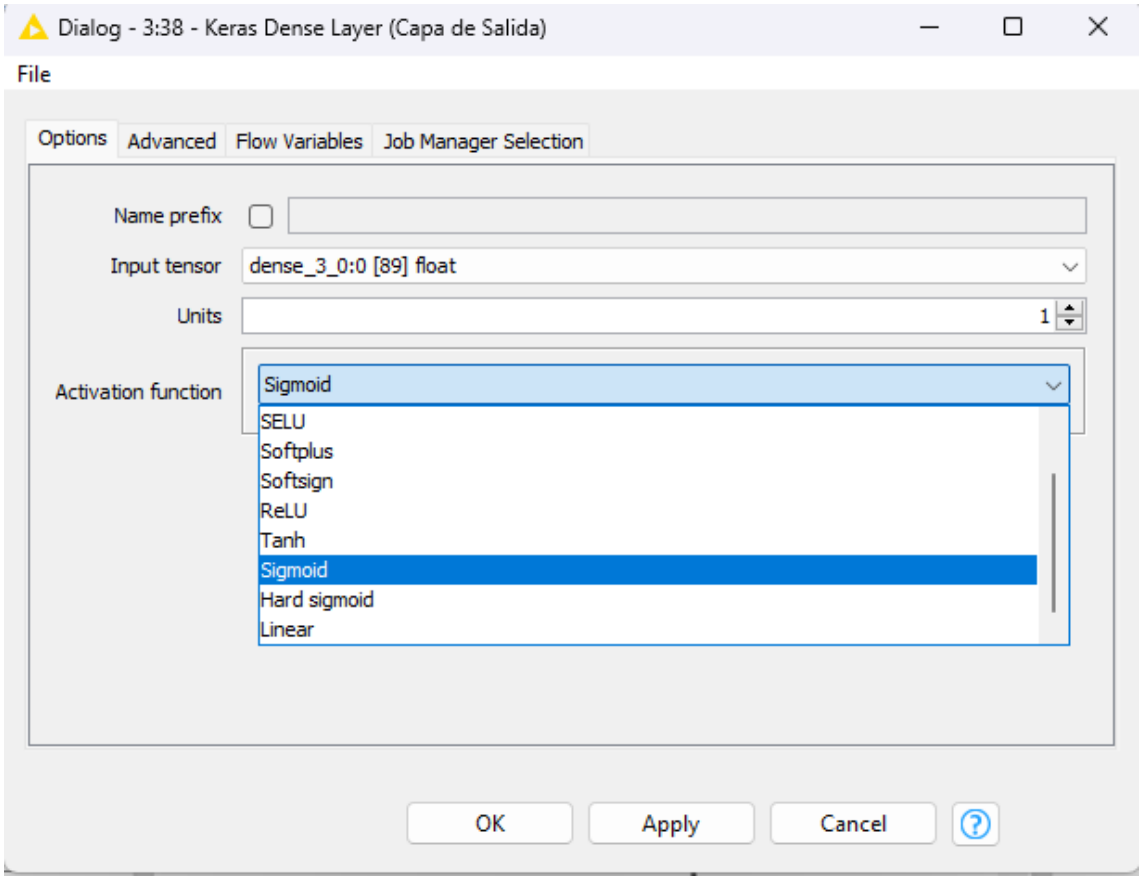


Figura 2.37: Configuración de Keras Dense Layer como capa de salida

2.2.10.3.3. Nodo Keras Network Learner

Es el Nodo donde se define como se va a realizar el entrenamiento de la Red Neuronal con los parámetros definidos en las capas, por lo que se debe realizar las siguientes configuraciones. En la pestaña *Input Data* se seleccionan todas las filas desde la 0 hasta la 1391 (los 1392 datos de lectura), como muestra la Figura 2.38.

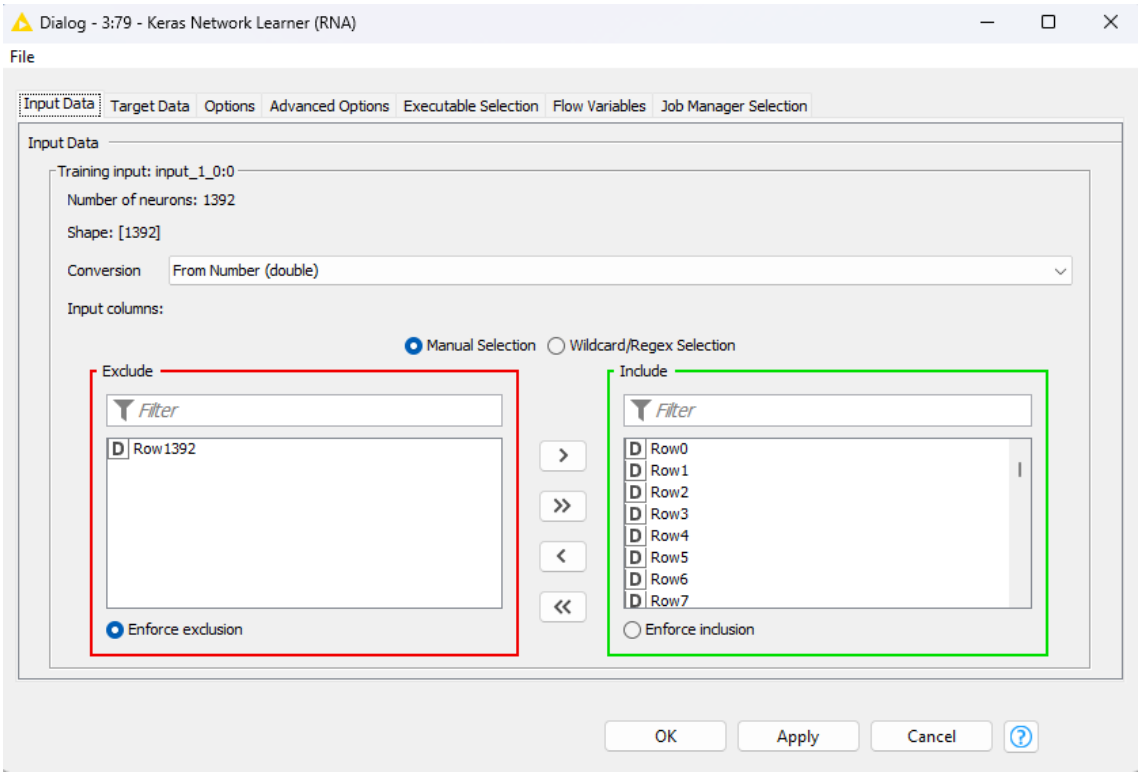


Figura 2.38: Configuración de Keras Network Learner pestaña Input Data

En la pestaña *Target Data* se selecciona la salida esperada de la Red Neuronal, que en este caso es la Fila 1392, donde está la calificación de la red de 0 y 1, como se indicó en la Figura 2.30 al cargar la data, además se escoge la función de error que se utilizará en la Red Neuronal, siendo la *Binary cross entropy*, la elegida debido a que es la función de coste utilizada cuando solo hay 2 posibles resultados de clasificación, en la Figura 2.39 se muestra la configuración.

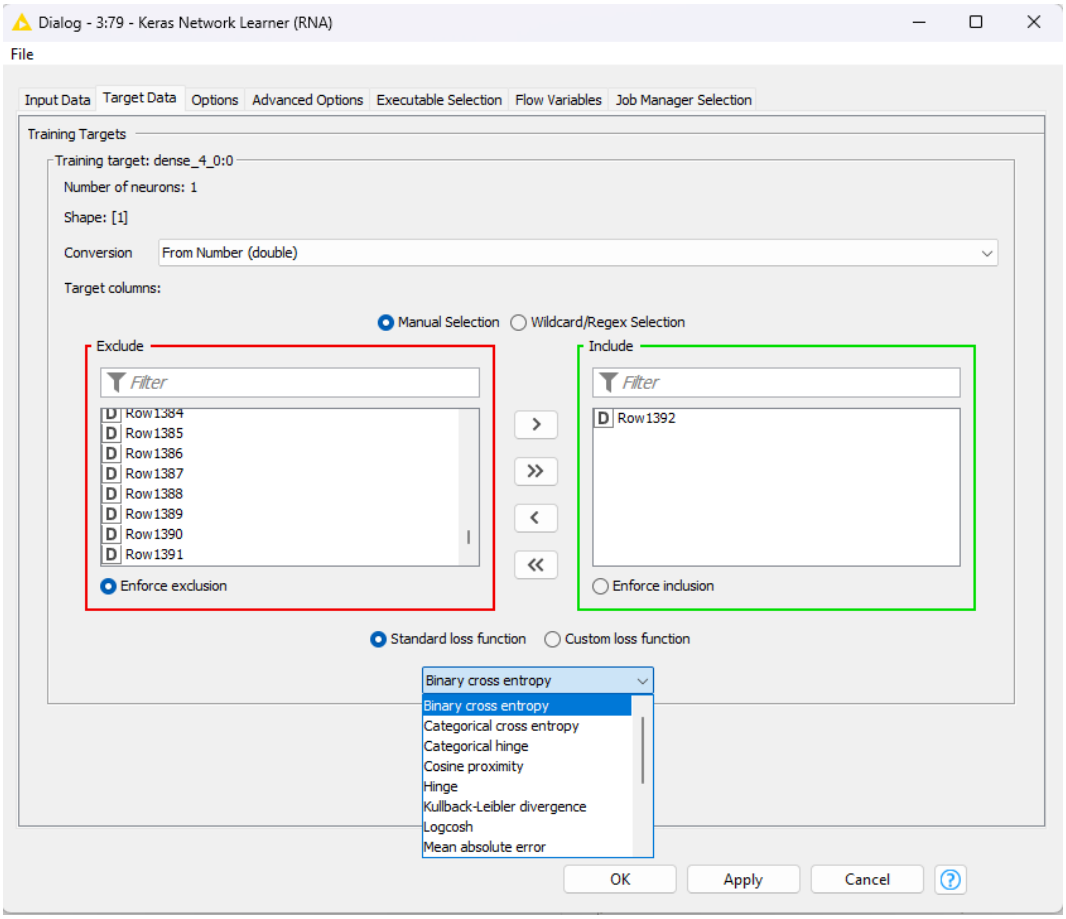


Figura 2.39: Configuración de Keras Network Learner pestaña Target Data

En la pestaña *Options*, se configura los Hiperparámetros de la Red Neuronal, iniciando por el *Back* end que es el paquete de *Keras (TesnorFlow)* el único elegible; luego se define los valores de *Epoch* colocados en 250, el *Training Batch size colocado en 100*, el *Validation Batch size* en 100; además es importante seleccionar el *Shuffle training data before each Epoch*, porque deseamos que, en cada época de entrenamiento, los datos sean tomados en desorden de filas. Finalmente se escoge el método de optimización, que en este caso se eligió Adam, que es un método de optimización estocástica basado en gradiente descendiente con estimaciones adaptativas en momentos de primer y segundo orden [32] con parámetros definidos en la taza de aprendizaje, Beta 1 y 2, Epsilon y Decaimiento de la Tasa de Aprendizaje, que si se requiere pueden ser modificables. La Figura 2.40 muestra la configuración detallada. Las otras pestañas quedan con la configuración por defecto del Keras.

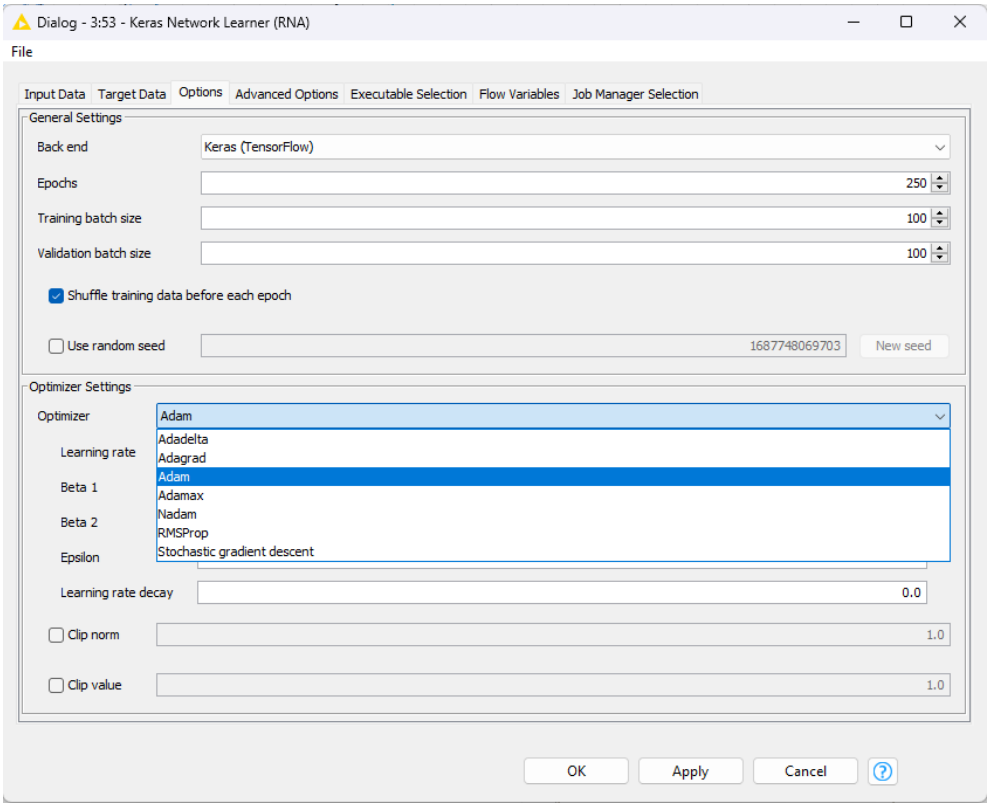


Figura 2.40: Configuración de Keras Network Learner pestaña Options

Quedando la red diseñada, configurada y lista para ser entrenada y validada. Si todos los Nodos se configuran correctamente la red presentará la forma de la Figura 2.41; en la cual se puede observar cómo fueron realizadas las conexiones entre capas, la conexión del Nodo *Partitioning* con el Nodo *Keras Network Learner*, enlazando los datos de entrenamiento y validación previamente realizados, listos para ser ejecutados.

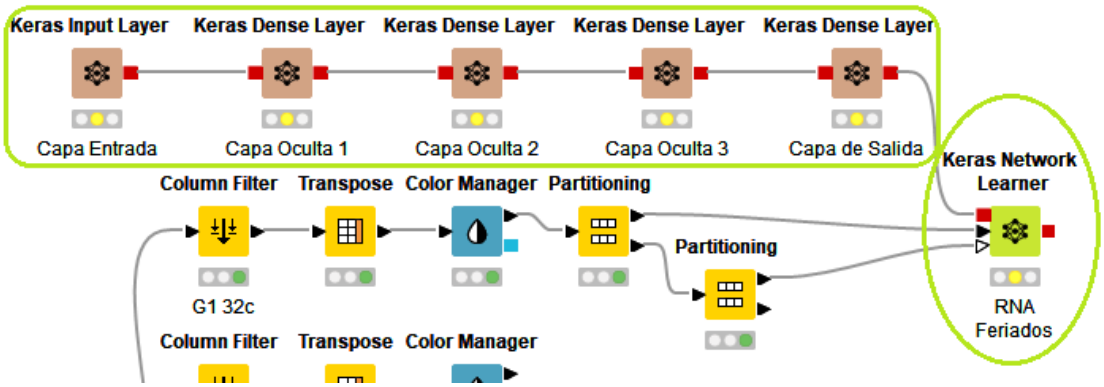


Figura 2.41: Red Neuronal Configurada para Ejecutarse

2.2.10.3.4. Nodo Keras Network Executor

Este es el Nodo dónde se podrá probar a la Red Neuronal, pues tiene como entradas el entrenamiento directo del aprendizaje de la Red y los datos particionados previamente para realizar las pruebas. La configuración del Nodo empieza en la pestaña *Options*, el *Back end* es el mismo del Nodo *Keras Network Learner*, siendo la única opción la librería *Keras (TensorFlow)*; el *Input Batch size* es el colocado previamente; en el *Inputs*, se incluye la fila de salida 1392, que es la salida que será evaluada con respecto a todas las filas anteriores; además en *Outputs*, dando clic en el botón *add output* se selecciona la última capa de salida con la función de activación sigmoide y se da en aceptar, quedando agregada la salida a ser ejecutada, como se muestra en la Figura 2.42. en la que se muestra las configuraciones descritas antes de elegir la capa de salida y luego de aceptar.

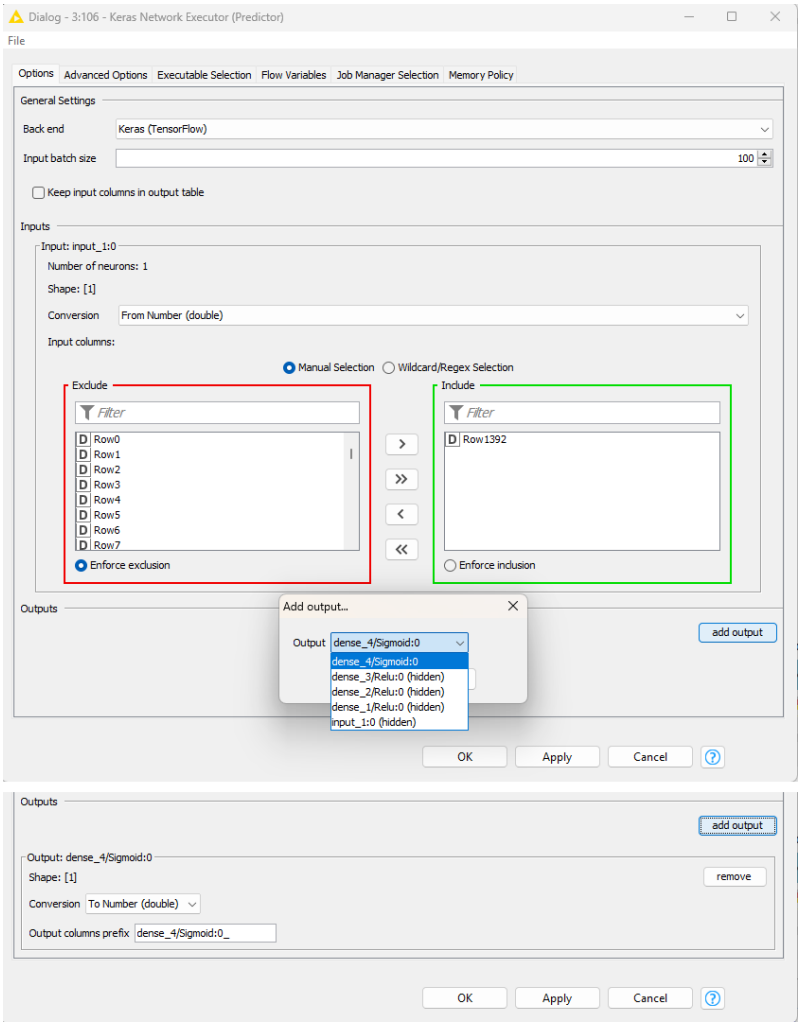


Figura 2.42: Configuración del Nodo Keras Network Executor

2.2.10.3.5. Nodo Line Plot (local)

El Nodo es parte de la librería por defecto y es empleado para poder observar los resultados arrojados por el modelo de la Red Neuronal, para lo cual se conecta la entrada de este Nodo con la salida del Nodo *Keras* Network Executor, como se muestra en la Figura 2.43.

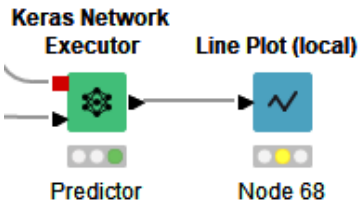


Figura 2.43: Red Neuronal completada en el entorno de trabajo.

Después de todos los parámetros descritos, el Modelo de la Red Neuronal y los Nodos de ejecución de datos quedará configurado en el área de trabajo como se muestra en la Figura 2.44, donde todos los Nodos están ejecutados y en color verde, mostrando el correcto procedimiento en la carga y filtrado de datos, la configuración, entrenamiento, validación y prueba del Modelo de la Red Neuronal.

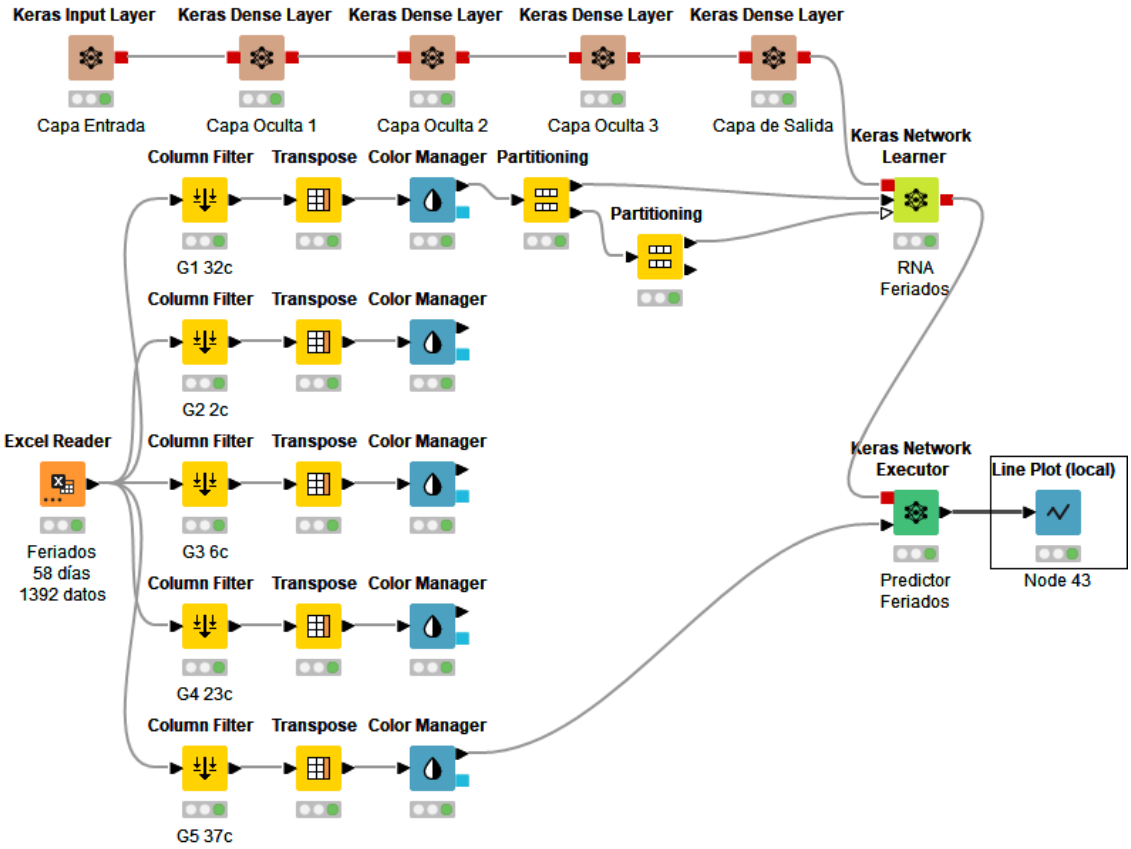


Figura 2.44: Red Neuronal completada en el entorno de trabajo.

Este proceso de construcción del Modelo de la Red neuronal se debe realizar para cada clasificación previamente hecha (lunes a viernes y fines de semana), por lo que al final se tendrá que entrenar tres modelos de redes neuronales. Ver Anexo E.

3. RESULTADOS Y DISCUSIÓN

3.1. Resultados

Los resultados están enfocados en la predicción que realiza la red neuronal para clasificar las curvas reales y las fraudulentas, por lo que se enfocará directamente en los márgenes de precisión y pérdidas de la Red Neuronal entrenada.

Las curvas descritas en los resultados del entrenamiento de la Red Neuronal se lo hicieron en uno de los grupos más numerosos.

La data completa leída en la plataforma está conformada por 128 curvas, de la cuales 32 pertenecen a las curvas diarias reales, 32 a curvas fraudulentas del Tipo 1, 32 a curvas fraudulentas del Tipo 2 y 32 a curvas fraudulentas del Tipo 3, siendo estas seleccionadas arbitrariamente al particionarlas y dividirlas en datos para entrenamiento (70% = 89 curvas), validación (15% = 19 curvas) y pruebas (15% = 20 curvas). Obteniendo los resultados descritos.

3.1.1. Precisión

Durante la Ejecución de los módulos, el entrenamiento se va realizando y con la opción *View: Learning Monitor* del *Keras Network Layer*, se puede observar el *Accuracy*, del entrenamiento, que indica la precisión del modelo y como se va entrenando con los hiperparámetros seteados durante la construcción de la red, como se muestra en la Figura 3.1; donde se puede visualizar como la curva suavizada va alcanzando los valores esperados con márgenes de error pequeños y aceptables dentro del entrenamiento y constatando que la función de optimización Adam y las Funciones de Activación en las capas para las neuronas trabajan conjuntamente para el aprendizaje de la Red Neuronal.

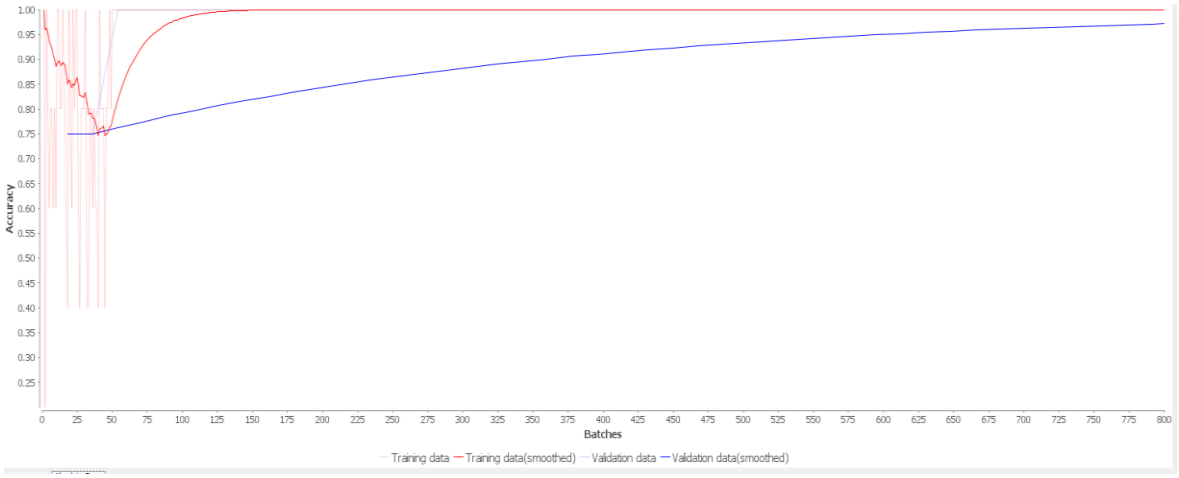


Figura 3.1: Curvas del Accuracy de la Red Neuronal

Los valores de los datos entrenados se van generando en la pestaña *Training data* del mismo View: Learning Monitor, en la Figura 3.2, se detallan los valores probabilísticos entregados por la red en cada ciclo de entrenamiento.

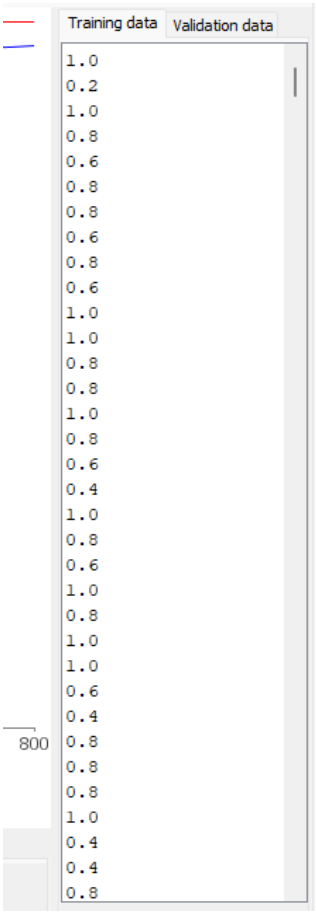


Figura 3.2: Valores de la Data Entrenada del Accuracy de la Red Neuronal

3.1.2. Pérdidas

Cambiando de pestaña del *Accuracy a Loss*, dentro del mismo *View: Learning Monitor* podemos observar cómo los valores del error calculado por la función *Binary cross entropy*, utilizada para la clasificación de los patrones de curvas, va haciendo que los valores sean los esperados, ya que el error va tendiendo a cero y las curvas van convergiendo en el tiempo que los Batch y Epoch van avanzando, como se muestra en la Figura 3.3, donde se visualiza que la curva en rojo va generando los valores entrenados y la curva en azul los valores de validación, llegando a tener un distanciamiento mínimo con respecto a los valores pronosticados y los valores deseados.

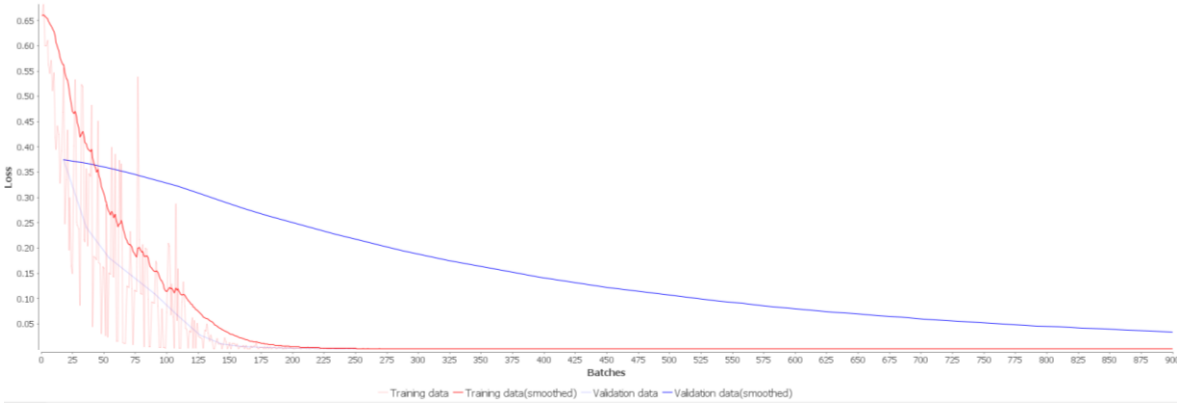


Figura 3.3: Curvas del Loss de la Red Neuronal

En la Figura 3.4 se muestra los valores probabilísticos del modelo de la Red Neuronal de la data entrenada.

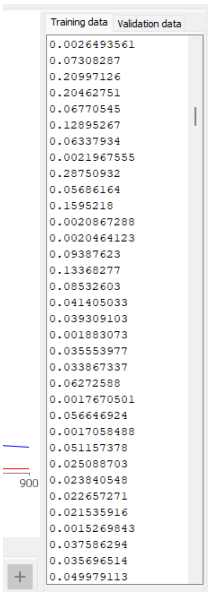


Figura 3.4: Valores de la Data Entrenada del Loss de la Red Neuronal

3.1.3. Pruebas

Con el 15% dejado para pruebas y con el Nodo *Keras Network Executor,* se realizan las pruebas de la Red neuronal entrenada, donde previamente se tiene 20 curvas mezcladas entre patrones reales y patrones fraudulentos; con el Nodo Line Plot, podemos observar cuál es el resultado de la predicción con respecto a esta pequeña data de curvas. En la Figura 3.5 se puede observar cómo se obtuvo resultados favorables para la Data descrita, donde se tenía 5 Curvas reales que se esperaba los calificara con cero, por tratarse de un modelo de red neuronal probabilístico el cálculo obtenido fue de 0.136, por lo cual se puede considerar que estas curvas tienen una etiqueta de cero y en el caso de las curvas fraudulentas se obtuvo un valor probabilístico de 0.984, lo cual se puede considerar que obtuvieron la etiqueta de 1.

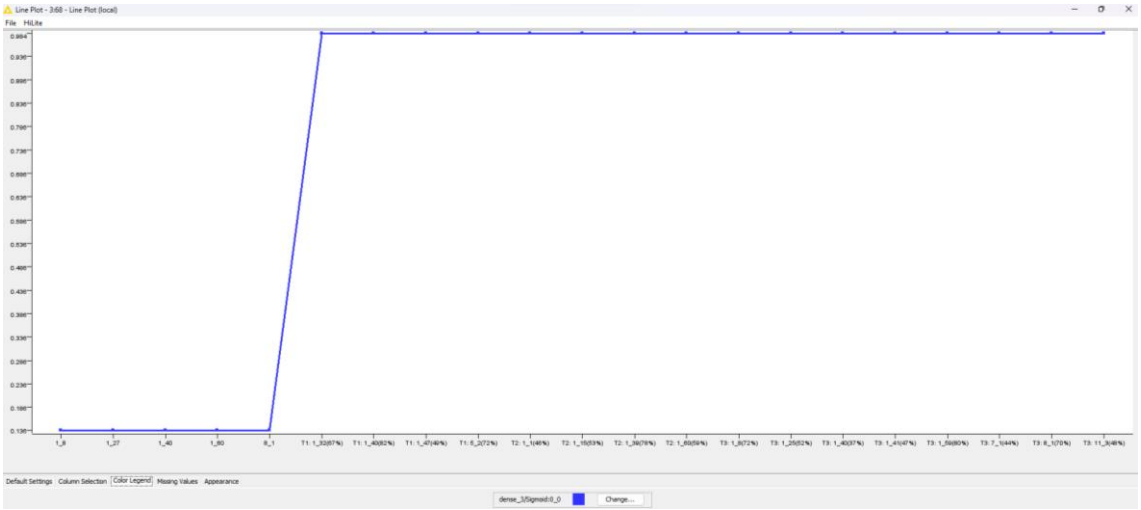


Figura 3.5: Resultados de la Red Neuronal en el Grupo 1

Como la Red Neuronal ya se encuentra entrenada, ahora se realizarán las pruebas en dos grupos con datos que son completamente desconocidos para la red entrenada, para lo cual realizaremos el mismo procedimiento con el Nodo *Keras Network Executor,* pero ahora con los datos del Grupo 4 que está conformado por 23 clientes con su data real y sintética para todos los tipos de fraudes. La Figura 3.6, muestra los resultados de la predicción para las curvas del grupo, siendo estos aceptables y favorables para la clasificación e identificación de patrones fraudulentos en el consumo de energía.

Figura 3.6: Resultados de la Red Neuronal en el Grupo 4

El Grupo 5 es el que más miembros tiene, al estar conformado por 37 clientes; por lo que se realiza el mismo proceso que en los grupos anteriores. La Figura 3.7 muestra los resultados predichos por la red neuronal, siendo estos también aceptables para la clasificación y reconocimiento de los clientes que subregistran energía.

Figura 3.7: Resultados de la Red Neuronal en el Grupo 5

3.1.4. Evaluación del Impacto de Energía y Económico

Con los resultados obtenidos y con la ausencia de patrones anómalos por parte de la distribuidora de clientes que subregistran energía, se realiza este análisis en base a los datos obtenidos en este trabajo.

3.1.4.1. Energía

Para calcular el valor de energía no facturado por los posibles clientes fraudulentos, se toma como ejemplo los posibles casos de fraude que se muestran en la Figura 2.22, de los que se tiene consumos promedios mensuales de energía y de demanda máxima, Ver Anexo C, que son valores entregados por la distribuidora; asumiendo que los clientes se encuentra perjudicando con los porcentajes de energía descritos a la distribuidora, se puede generalizar los valores de la Tabla 3.1, donde se muestra cómo afectan los fraudes de estos clientes en una proyección anual.

Tabla 3.1: Clientes Fraudulentos Detectados

Clientes	Porcentaje de Energía No Registrada (%)	Energía Promedio Consumida Mensual (kWh)	Energía No Facturada Mensual (kWh)	Energía Facturada Anualmente Real (kWh)	Energía No Facturada Anualmente (kWh)
1_15	53	97200	51516	1166400	618192
2_6	36	31920	11491.2	383040	137894.4
4_5	50	10845	5422.5	130140	65070
7_1	42	70560	29635.2	846720	355622.4
8_1	37	4936	1826.32	59232	21915.84
11_3	57	10920	6224.4	131040	74692.8
			TOTAL	**2716572**	**1273387.44**

Con estos valores se puede determinar que el valor de energía no facturado está en los 1273,4 MWh, pudiendo ser este una posibilidad de reducción de pérdidas con un valor porcentual del 0,18% de los 720,63 GWh de Energía Facturada y 41,21 GWh (5,41%) de pérdidas registrados por la distribuidora en el 2022 [34], al aplicar las técnicas desarrolladas en el presente proyecto. Es importante mencionar que los valores y datos son hipotéticos, porque la distribuidora no posee una base de datos con perfiles de carga de los clientes que han sub registrado energía.

3.1.4.2. Económico

Con los valores de la Tabla 3.1, y conociendo que el valor del Pliego Tarifario del Servicio Público de Energía Eléctrica está en $0,095 por KWh, para estos clientes específicamente, como valor máximo para los clientes especiales de la distribuidora, se puede aseverar que los clientes están perjudicando a la distribuidora con un valor de $120971,8068 anual, como se muestra en la Tabla 3.2.

Tabla 3.2: Valores No Facturados por Clientes Fraudulentos Detectados

Cliente	Energía Facturada Anualmente (kWh)	Energía No Facturada Anualmente (kWh)	Valor Facturado Anualmente Real (S)	Valor No Facturado Anualmente ($)
1_15	1166400	618192	110808	58728.24
2_6	383040	137894.4	36388.8	13099.968
4_5	130140	65070	12363.3	6181.65
7_1	846720	355622.4	80438.4	33784.128
8_1	59232	21915.84	5627.04	2082.0048
11_3	131040	74692.8	12448.8	7095.816
		TOTAL	258074.34	**120971.8068**

El valor calculado en dólares corresponde al valor de energía que se estableció como posible fraude en este caso de estudio.

4. CONCLUSIONES

- En el desarrollo del modelo de la red neuronal, es importante el proceso de entrenamiento, ya que se deben elegir y configurar correctamente los hiperparámetros (Estructura y Topología de la Red Neuronal: El número de capas, El número de neuronas de cada capa, las funciones de activación, etc.; A nivel del Algoritmo de Aprendizaje: Las Epoch, Los Batch Size, El Learning Rate, El Momentum, etc.), porque estos hiperparámetros cumplen con su función en cada etapa, deviniendo un paso esencial para conseguir un buen diseño del modelo de red neuronal, para que cumpla con el objetivo para el que fue diseñada.

- Durante el proceso de entrenamiento, si no se alcanza los resultados esperados, se debe empezar a probar cambiando los hiperparámetros, como por ejemplo aumentando el número de neuronas en las capas ocultas en concordancia con la cantidad de variables que se ingresa, al igual que escoger las funciones de activación adecuadas, ya que son decisiones transcendentales, pues serán las encargadas de activar o no dichas neuronas en cada ciclo de entrenamiento. En este trabajo justamente se aumentaron el número de neuronas y se pusieron las funciones de activación apropiados y se lograron buenos resultados.

- Las funciones de pérdida (*Loss*) y precisión (*Accuracy*) son importantes ir analizándolas y monitoreándolas durante el proceso de entrenamiento, porque su forma y tendencia nos permite conocer si el modelo está bien diseñado. En este trabajo se obtuvieron los comportamientos apropiados.

- El *overfitting* (sobreajuste o sobre entrenamiento) y el *underfitting* (desajuste o bajo entrenamiento), son referencias que hacen que el modelo de la red neuronal no pueda capturar las tendencias subyacentes de los datos, es decir que el modelo al generalizar no encaje con el conocimiento que pretendemos que adquiera, estableciendo que las curvas *Accuracy* y *Loss* diverjan y el modelo de la red neuronal no cumpla con el propósito para la que fue creada. En el entrenamiento y

validación de este modelo se tuvo la precaución de que no se produzca el overfitting y underfitting.

- Es importante analizar el comportamiento de las curvas *Loss* y *Accuracy*, tanto para el entrenamiento como para la validación, porque ese monitoreo nos va a permitir observar si en algún instante empieza a ocurrir el *overfitting*. En el modelo de la red neuronal desarrollada en este trabajo se evitó cometer este error, disminuyendo las *Epoch* y los *Batch* de forma liviana y ajustada, para evitar caer en el *underfitting* u *overfitting* y mantenernos en las franjas del training.

- El entrenamiento del modelo de la red neuronal con respecto a los patrones fraudulentos, no se lo realizo con curvas fraudulentas reales, ya que la distribuidora no contaba con esta base de datos histórica, por lo que se propuso en esta metodología el crear diferentes tipos de patrones de curvas fraudulentas.

- Realizar las pruebas del modelo de la red neuronal con los registros de energía suministrada actual a los clientes de la distribuidora, arrojará resultados interesantes para el ajuste y precisión en la predicción de patrones con subregistro o consumos anormales en la red de la empresa.

- Con la metodología descrita en este trabajo se puede replicar el modelo y probarlo con los datos obtenidos de las AMI instaladas en los clientes que poseen las unidades de negocio de la CNEL.

- La clasificación en base a los días de la semana, fines de semana y feriados, por sus consumos similares de energía, hacen que el estudio y análisis de los clientes se enfoque en todos los aspectos para la detección de posibles fraudes o registros de consumos anómalos.

- La clusterización de los datos creadas por los días de la semana, representó uno de los puntos clave, ya que en el primer método, utilizando el k-medias con sus

diagramas de codo, resultaron ser no idóneos para este tipo de datos, por lo que la inclusión de índices de validación que sean empleados en datos con variaciones en el tiempo, hacen que la data sea mejor agrupada con los índices DTW, SC y el empleo del WCSS, que basados en su metodología caracterizaban y agrupaban mejor las curvas de los clientes y de sus centroides, eligiendo en cada caso el mejor agrupamiento de los tres métodos empleados.

- La consolidación de la Base de Datos ha sido un arduo trabajo que ha conllevado un tiempo demasiado extenso, debido a la descarga lenta, individual (cliente por cliente) y la generación de varios archivos para la consolidación de un archivo completo de datos de un solo cliente, por lo que se sugiere a la distribuidora, se mejore en este aspecto.

- Para futuros trabajos en el diseño de modelos de redes neuronales se sugiere utilizar modelos convolucionales y recurrentes (RNN) y comparar con las redes densamente conectadas con la finalidad de ver el desempeño de cada una de estas arquitecturas.

5. REFERENCIAS BIBLIOGRÁFICAS

[1] J. J. Estupiñan, D. A. Giral y F. Martínez, Implementación de algoritmos basados en máquinas de soporte vectorial (SVM) para sistemas eléctricos: revision de tema, Universidad Distrital Francisco José de Caldas, feb 2016.

[2] J. C. Bermeo, G. A. Arguello y J. C. Cepeda, Estadística Anual y Multianual del sector Eléctrico Ecuatoriano, Agencia de Regulación y Control de Energía y recursos Naturales No Renovables, mar 2022, pp. 82 - 88.

[3] M. C. Giraldo, C. Ríos, A. Alarcón, V. Snyder, C. Echeverría, A. Riobo, M. Hallack y J. L. Irigoyen, Energizados: los beneficios de una herramienta basada en la maetodologiía de machine learning para facilitar la detección de robo eléctrico, Banco Interamericano de Desarrollo: División de Energía, ene 2022.

[4] M. L. Álvarez, J. C. Olivares, N. E. Rodríguez, E. R. Archundia, J. E. Alcaraz y J. A. Gutierrez, Modelo de predicción de lecturas de consumo/producción de energía eléctrica para detectar fraudes de energía, 8° Encuentro de Jóvenes Investigadores del estado de Michoacán: Area 2: Ingeniería y Tecnología, 2022.

[5] P. C., Combate, prevenção e otimização das perdas comerciais de energia elétrica, Sao Paulo: Ph. D. thesis Universidade de São Paulo, 2008.

[6] H. J., Processo não invasivo de baixo custo para otimizaão da rotina de inspeão na detecão de furto de energía eléctrica, evista Pesquisa e Desenvolvimento da ANEEL - P&D, [S.l.], n.5,, Ago 2013, pp. 47-51.

[7] R. Cruz y F. Pérez, Detecting Non-Technical Losses in Radial Distribution System Transformation Point through the Real Time State Estimation Method, EEE/PES Transmission & Distribution Conference and Exposition: Latin america, 2006.

[8] A. M. K. Y. S. T. a. S. A. J. Nagi, Non-technical loss analysis for detection of electricity theft using support vector machines, Proc. 2nd IEEE Int. Power and Energy Conf., 200, 2007.

[9] F. B. C. L. J. B. a. R. M. I. Monedero, MIDAS: Detection of non-technical losses in electrical consumption using neural networks and statistical techniques, Berlin/Heidelberg Germany: Conf. Computational Science and Applications, Springer, 2006.

[10] G. C. y. otros, Comparisons Among Clustering Techniques for Electricity Custome Classification, IEEE Transactions on Power Systems , vol. 21, nº 2, Mayo 2016, pp. 933-940.

[11] E. A. y. otros, Detection and Identification of Abnormalities in Customer Consumptions in Power Distribution Systems, EEE Transactions on Power Delivery, 2011.

[12] R. A., Estimação de perdas técnicas e comerciais: métodos baseados em fluxo de carga e estimador de estados, Porto Alegre: M.S. thesis, Universidade Federal do Rio Grande do Sul, 2014.

[13] A. Rodrigues, A. Costa y D. Issicaba, Identification of Non-Technical Losses in Distribution Systems via State Estimation and Geometric Tests, Brasil: Universidade Federal de Santa Catarina Florianópolis, 2018.

[14] F. C. L. T. W. F. F. J. C. M. V. J. T. S. D. Ferreira, Detecção de Perdas Não Técnicas na Presença dos Medidores Inteligentes, Anais do V Simpósio Brasileiro de Sistemas Elétricos - SBSE, 2014.

[15] M. I., Detection of frauds and other non-technical losses in a power utility using Pearson coefficient, Bayesian networks and decision trees, International Journal of Electrical Power & Energy Systems, 2012.

[16] A. Abur y A. G. Exposito, Power System State Estimation: Theory and Implementation. New York: Marcel Dekker, 2004.

[17] W. S. R., Detección de pérdidas no técnicas en redes de distribución radiales usando estimación de estado, Rio de Janeiro , Diciembre 2016.

[18] M. M. B. David, IDENTIFICACIÓN DE PÉRDIDAS NO TÉCNICAS DE ENERGÍA ELÉCTRICA MEDIANTE LA COMBINACIÓN DE UN CLASIFICADOR DE SVM (SUPPORT VECTOR MACHINE) Y UN ESTIMADOR DE ESTADO, Quito: Escuela Politécnica Nacional, Noviembre 2022.

[19] J. Han, M. Kamber y J. Pei, Data Mining: Concepts and Techniques, M. K. Inc., 2011.

[20] A. L. L. Z. J. Z. J. G. S. &. J. G. M. Rajabi, A Review on Clustering of Residential Electricity Customers and Its Applications, 20th International Conference on Electrical Machines and Systems (ICEMS), 2017, pp. 1-6..

[21] J. &. G. L. Paparrizos, K-Shape: Efficient and Accurate Clustering of Time Series. SIGMOD, 2016, p. 69–76..

[22] R. F. J. V. G. D. F. A. G. H. C. .. .. .. W. E. Tavenard, Tslearn, A Machine Learning Toolkit for Time Series Data, Journal of Machine Learning Research, 21(118),, 2020, pp. 1-6.

[23] M. M. B. Jonathan, Segmentación de clientes usando datos de medidores inteligentes de electricidad, La Rioja, España, 2022.

[24] T.-H. O. R. &. W. H. Dang-Ha, Clustering Methods for Electricity Consumers: An Empirical Study in Hvaler - Norway., Norway: Norsk Informatikkonferanse (NIK), 2016.

[25] EduPristine, Unsupervised Learning: Evaluating Clusters, 2018.

[26] W. S. McCulloch y W. Pitts, A logical calculus of the ideas immanent in nervous activity., The Bulletin of Mathematical Biophysics, 1943, p. 115–133.

[27] S. S., Artificial neural network modelling: An introduction, Springer International Publishing., 2016.

[28] W. S. C., Artificial neural network. En Interdisciplinary computing in java programming, Boston: Springer US, 2003, pp. 81-100.

[29] S. S., Neural Networks: All You Need to Know - Towards Data Science., 2020.

[30] F. E. Á. Francisco, Motor selector de técnicas de machine learning para la detección de anomalías de consumo eléctrico en clientes residenciales, Valaparaíso: Universidad de Valparaiso Chile, Dic. 2021, pp. 21 - 22.

[31] G. C. Pau, Diseño e Implementación de un Clasificador Mediante Redes Neuronales para un Sistema de Inspección Indutrial 3D, Univesitat Politécnica de Valencia, 2019.

[32] D. Agustina, Aprendizaje y Análisis de Redes Neuronales Artificiales Profundas, Universidad Nacional de Cuyo, Junio 2018.

[33] KeepCoding, Optimización de hiperparámetros en Deep Learning, Tech School, Agosto 2022.

[34] EEASA, «Informe de Rendición de Cuentas,» Socialización Interna, Ministerio de Energía y Minas, 2022.

[35] ARCERNNR, «Regulación Nro. ARCERNNR,» Quito, 2020.

[36] M. E, Chronology of Interpolation: From Ancient Astronomy to Modern Signal and Image Processing, IEEE, 90, 319-342. doi:10.1109/5.993400, March 2022.

[37] J. Nagi, A. Mohammad, K. Yap, S. Tiong, and S. Ahm, Non-technical loss analysis for detection of electricity theft using support vector machines, Proc. 2nd IEEE Int. Power and Energy Conf., 200., 2006.

[38] J. T. Tou y R. C. Gonzalez, Pollern Recognilion Principles, En Addison-Wesley, 1974.

[39] P. Massaferro Saquieres, Detección de pérdidas no técnicas en redes eléctricas en un contexto de migración tecnológica y maximizando el retorno económico, Montevideo: Facultad de Ingeniería de la Universidad de la República, marzo 2022, pp. 71-73.

[40] D. P. Kingma y J. Ba, Adam: A Method for Stochastic Optimization, Cornell University, Jan 2017.

6. ANEXOS

ANEXO A. Programación desarrollada en JupyterLab de la Data Completa: Carga de Datos, Reducción, Clasificación, Normalización, Clusterización, Creación de Grupos y Estudio Estadístico.

ANEXO B. Curvas de clientes generadas en JupyterLab.

ANEXO C. Grupos Conformados de la clasificación de los días.

ANEXO D. Programación desarrollada en JupyterLab de los Clientes Fraudulentos.

ANEXO E. Modelos de Red Neuronal realizada para los días de lunes a viernes y los Fines de Semana.

ANEXO A

PROGRAMACIÓN DESARROLLADA EN JUPYTERLAB DE LA DATA COMPLETA: CARGA DE DATOS, REDUCCIÓN, CLASIFICACIÓN, NORMALIZACIÓN, CLUSTERIZACIÓN, CREACIÓN DE GRUPOS Y ESTUDIO ESTADÍSTICO.

- En el formato pdf con nombre Anexo A.

ANEXO B:

CURVAS DE CLIENTES GENERADAS EN JUPYTERLAB

Figura B.1: Verificación que no haya lagunas de datos temporales

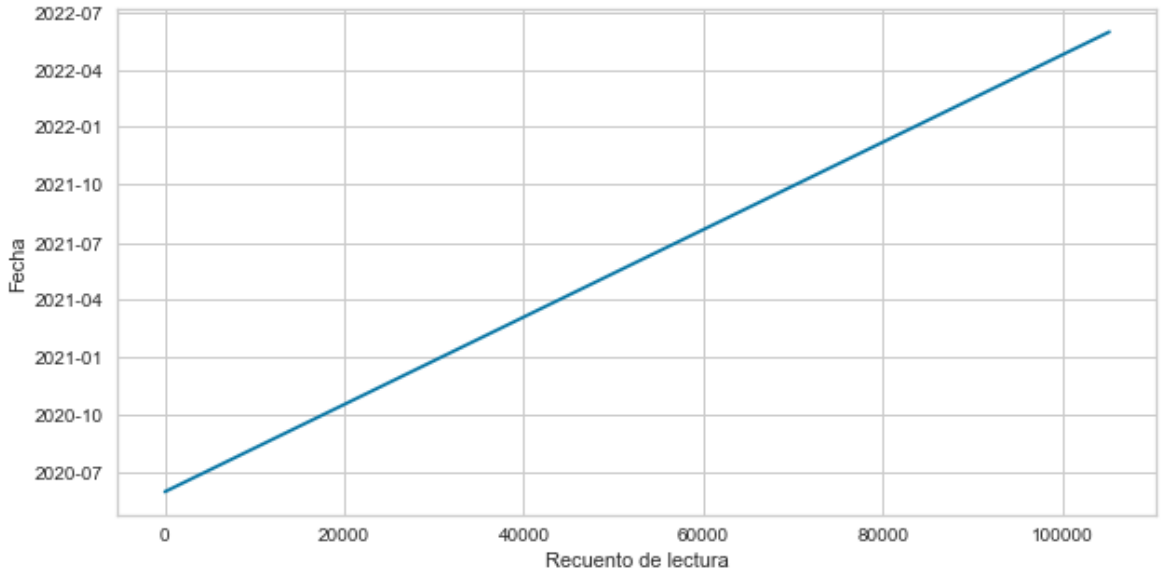


Figura B.2: Gráfica de las Curvas Diarias de los Clientes 1_1, 2_2 y 12_3

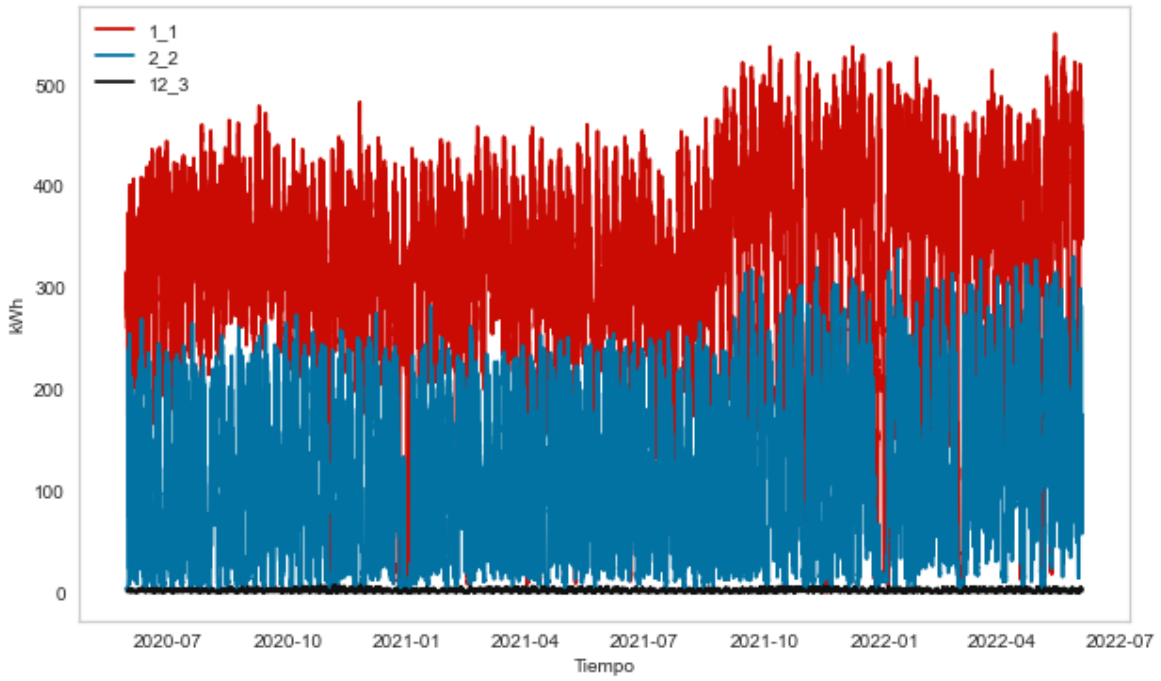


Figura B.3: Curvas de los Clientes de los días de lunes a viernes

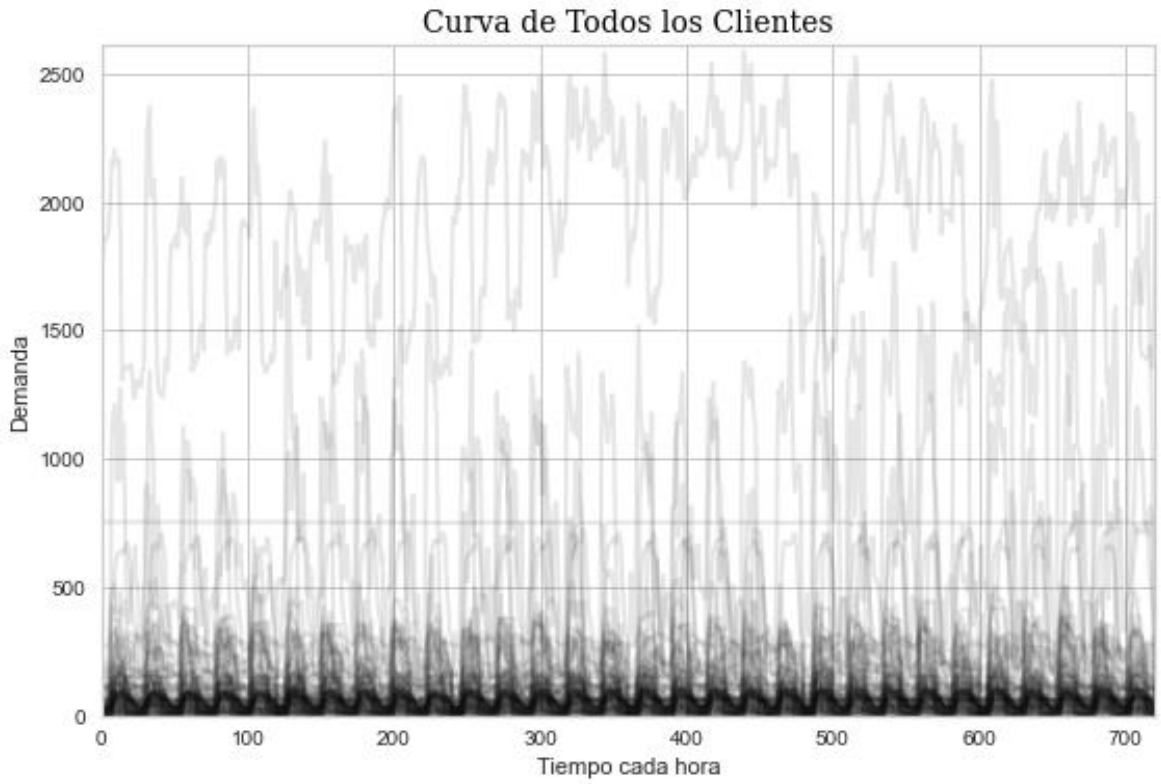


Figura B.4: Curvas de los Clientes de los días de Fin de Semana

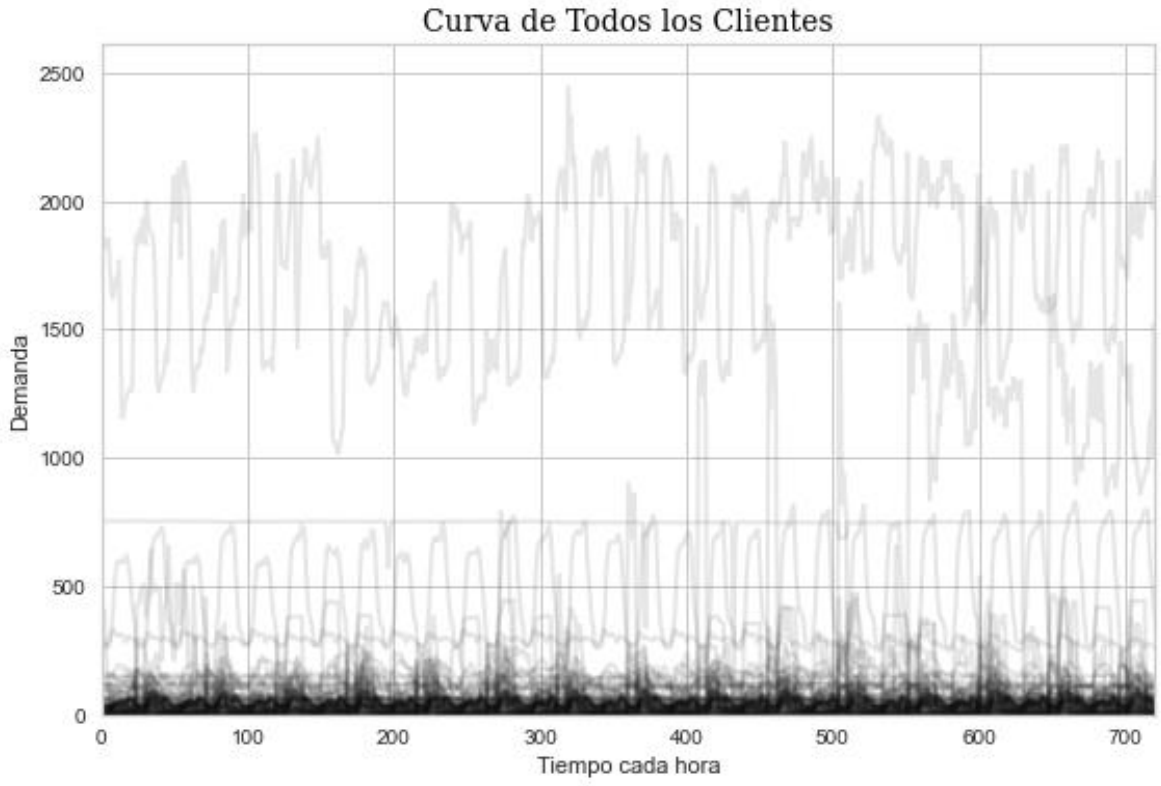


Figura B.5: Curvas de los Clientes de los días de Feriado

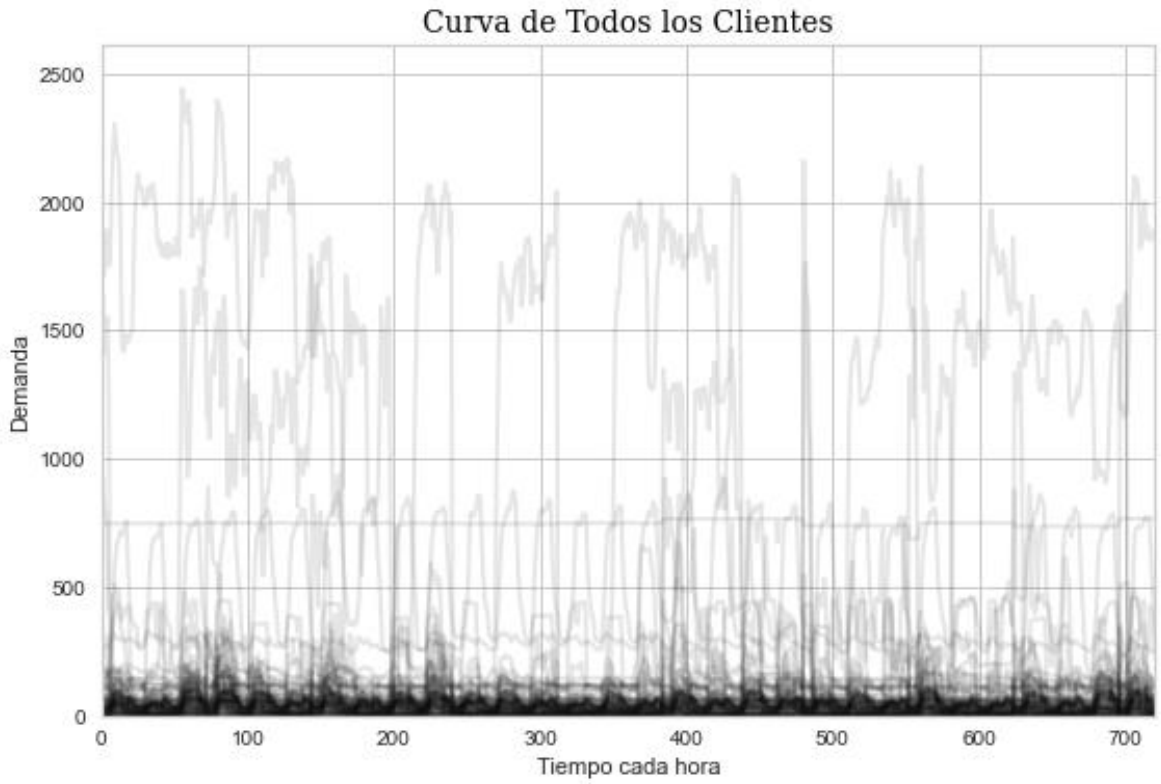


Figura B.6: Curvas de los Clientes de los días de lunes a viernes normalizada

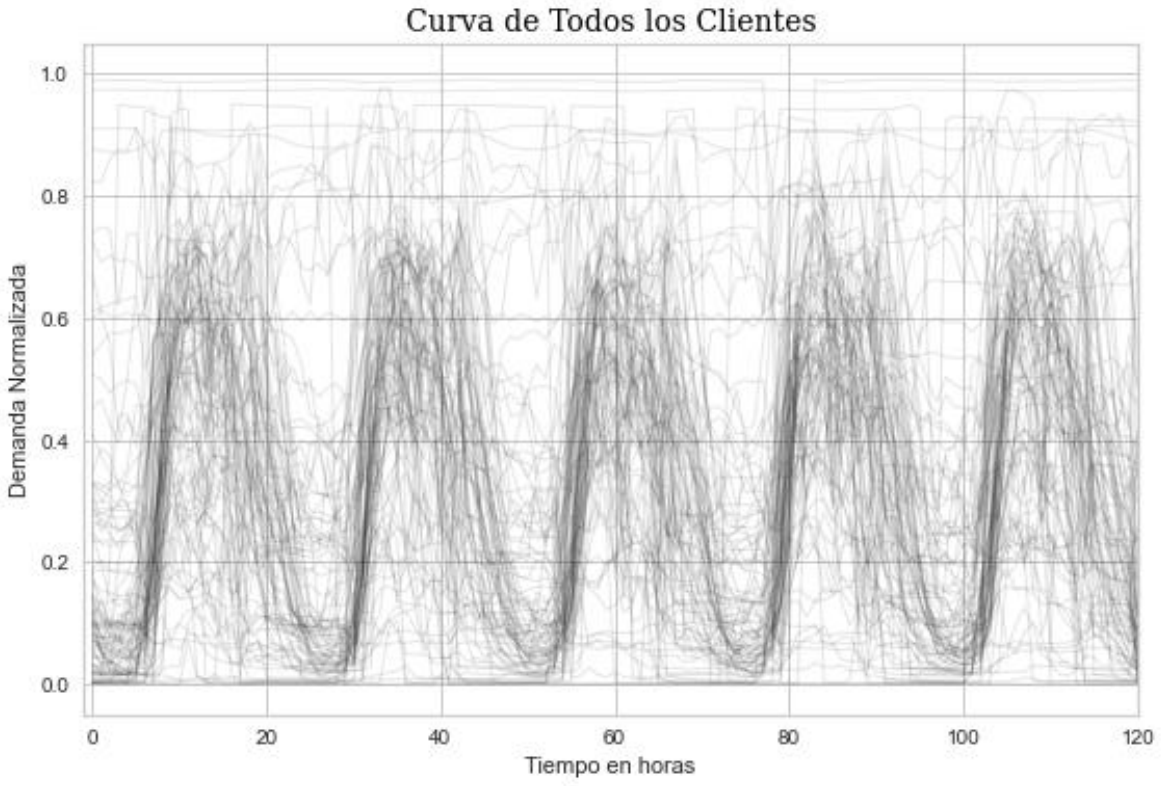


Figura B.7: Curvas de los Clientes de los días de fines de Semana normalizada

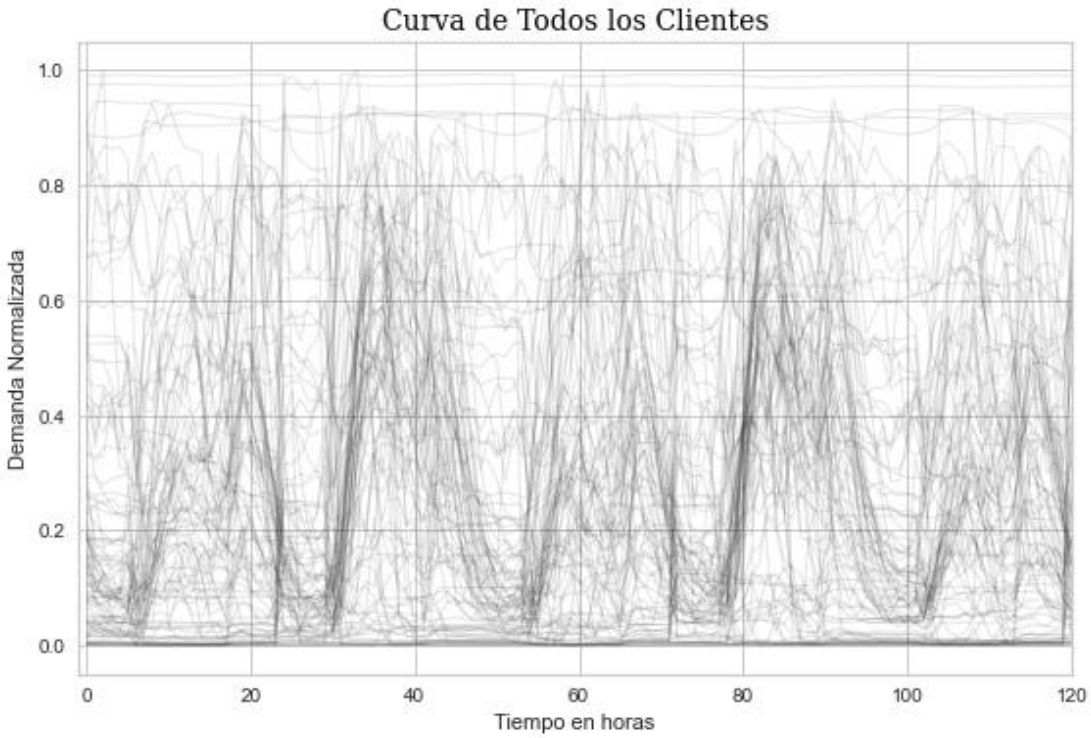


Figura B.8: Curvas de los Clientes de los días de Feriado normalizada

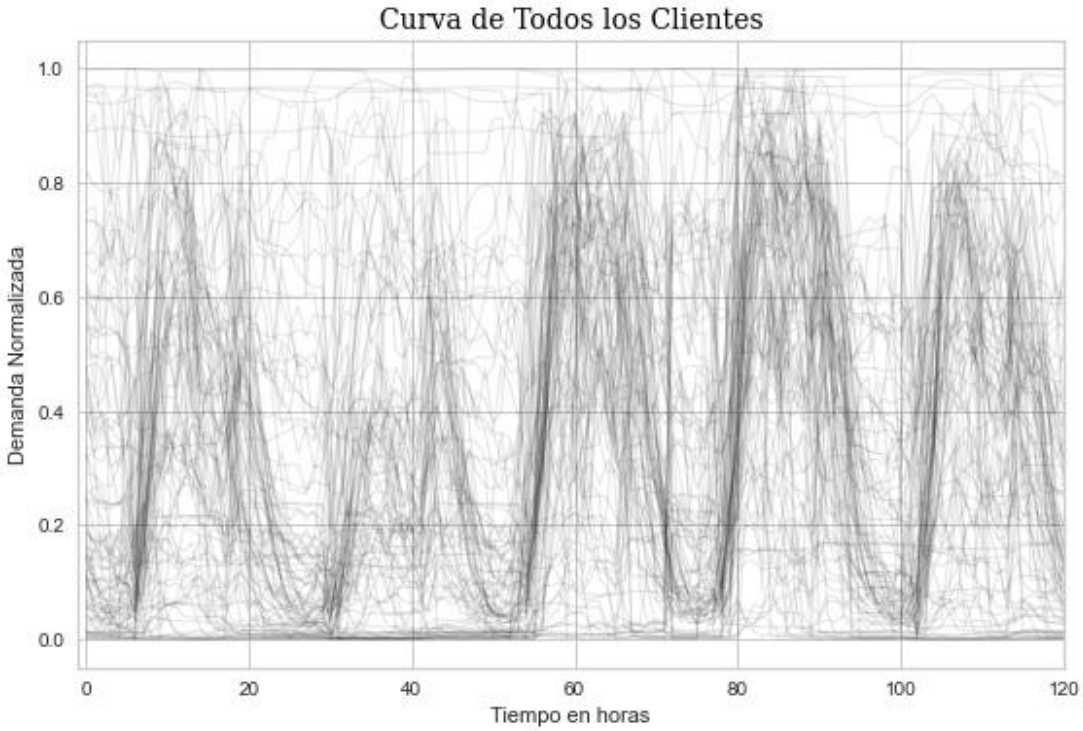
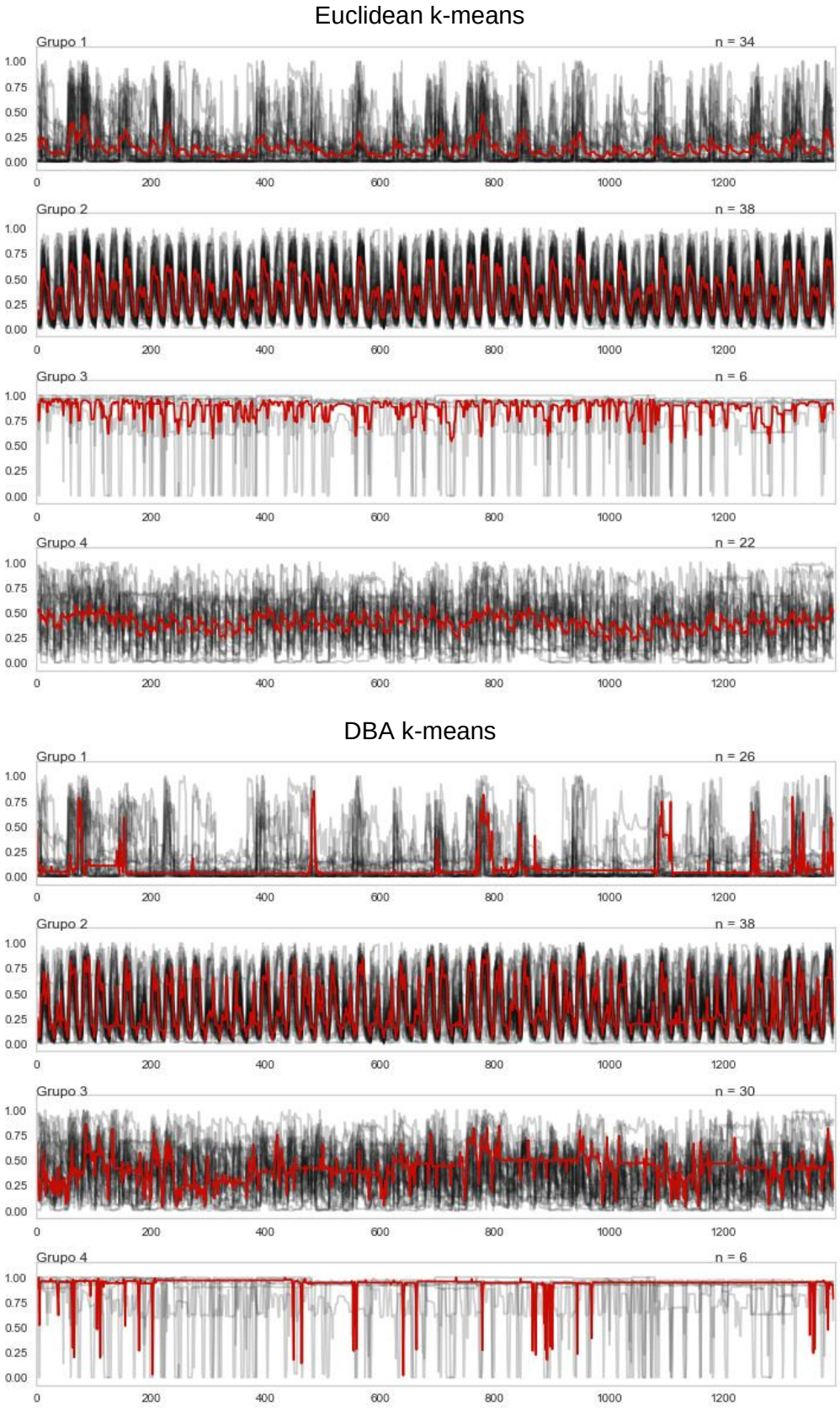


Figura B.9: Agrupamiento del Grupo de los días de feriado con K = 4 y los 3 métodos de evaluación.



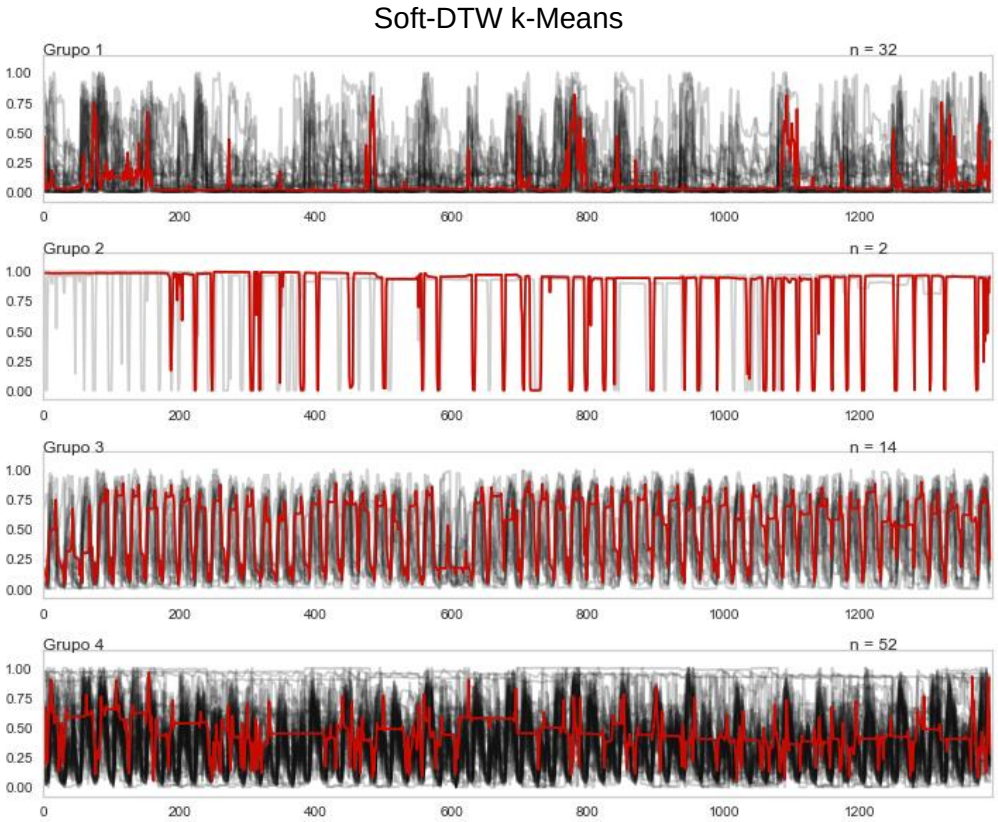
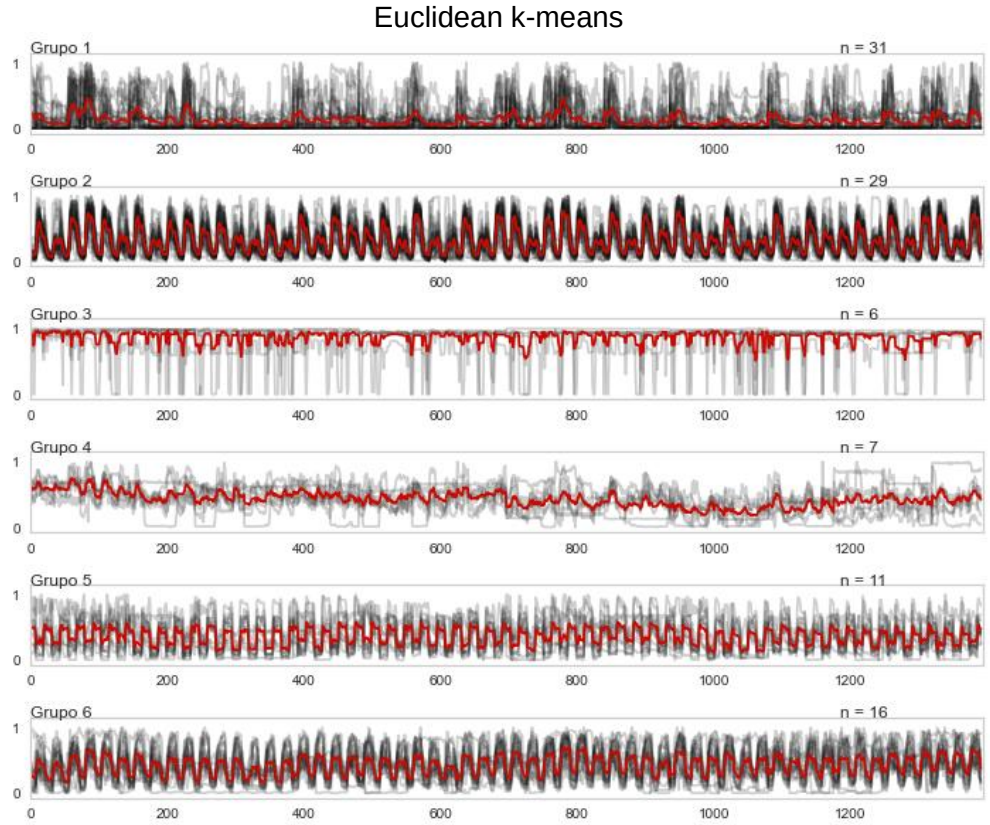
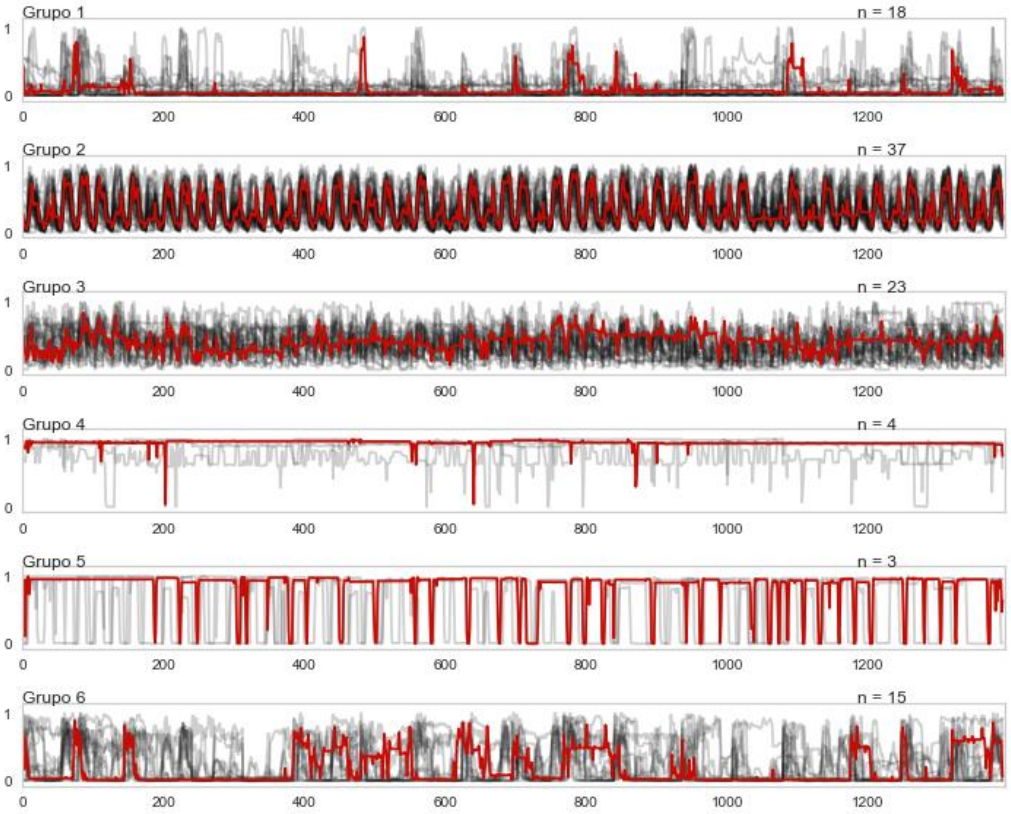


Figura B.10: Agrupamiento del Grupo de los días de feriado con K = 6 y los 3 métodos de evaluación.



DBA k-means



Soft-DTW k-Means

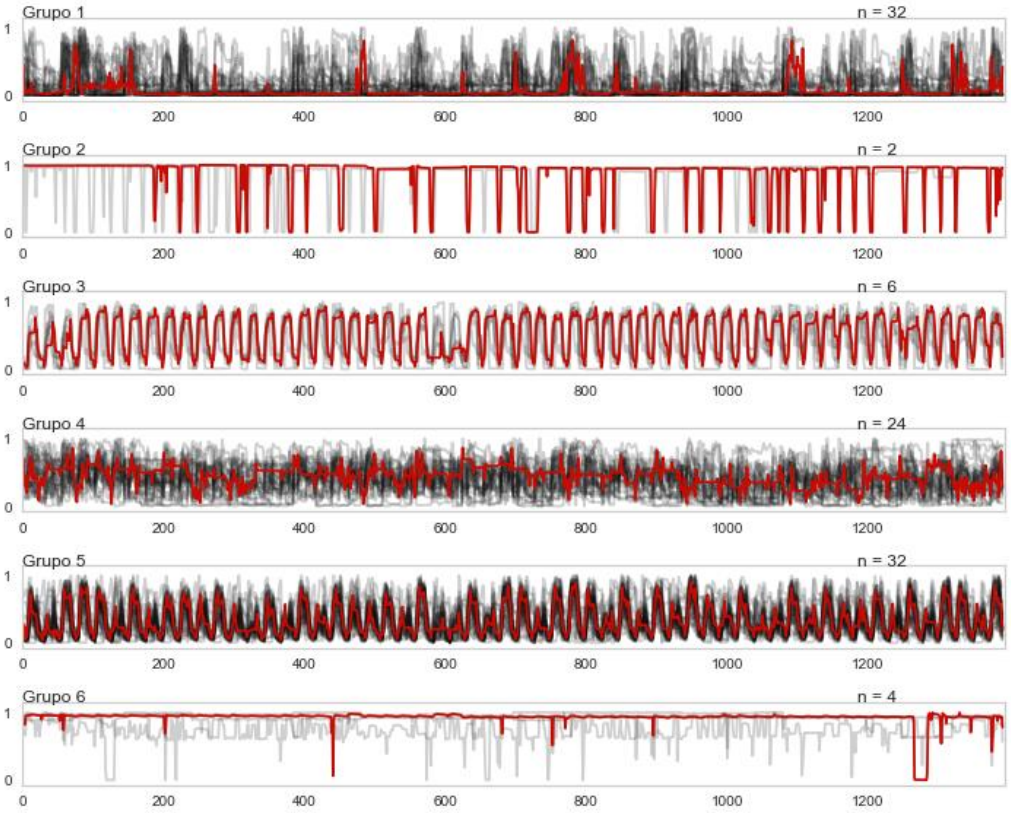
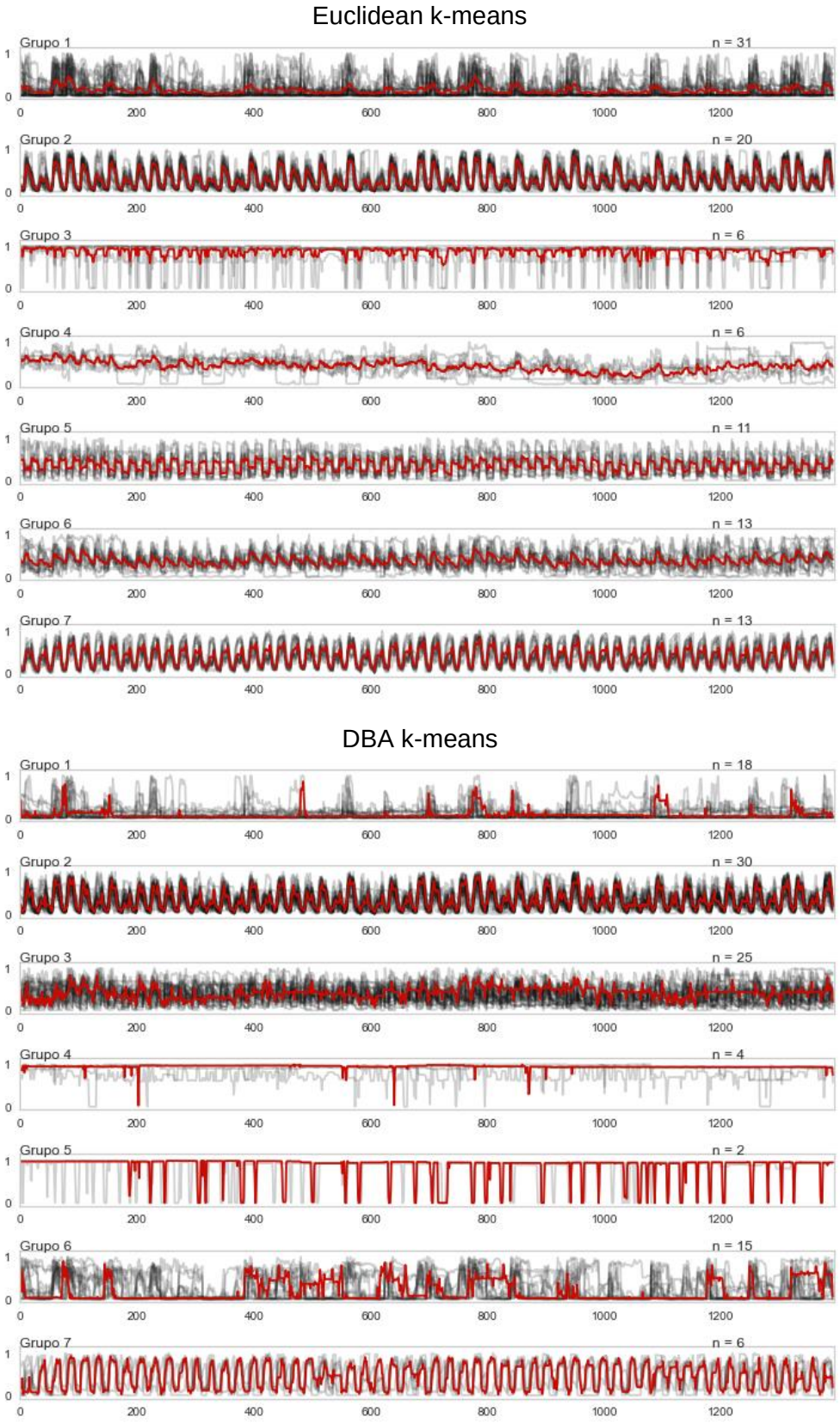


Figura B.11: Agrupamiento del Grupo de los días de feriado con K = 7 y los 3 métodos de evaluación.



Soft-DTW k-Means

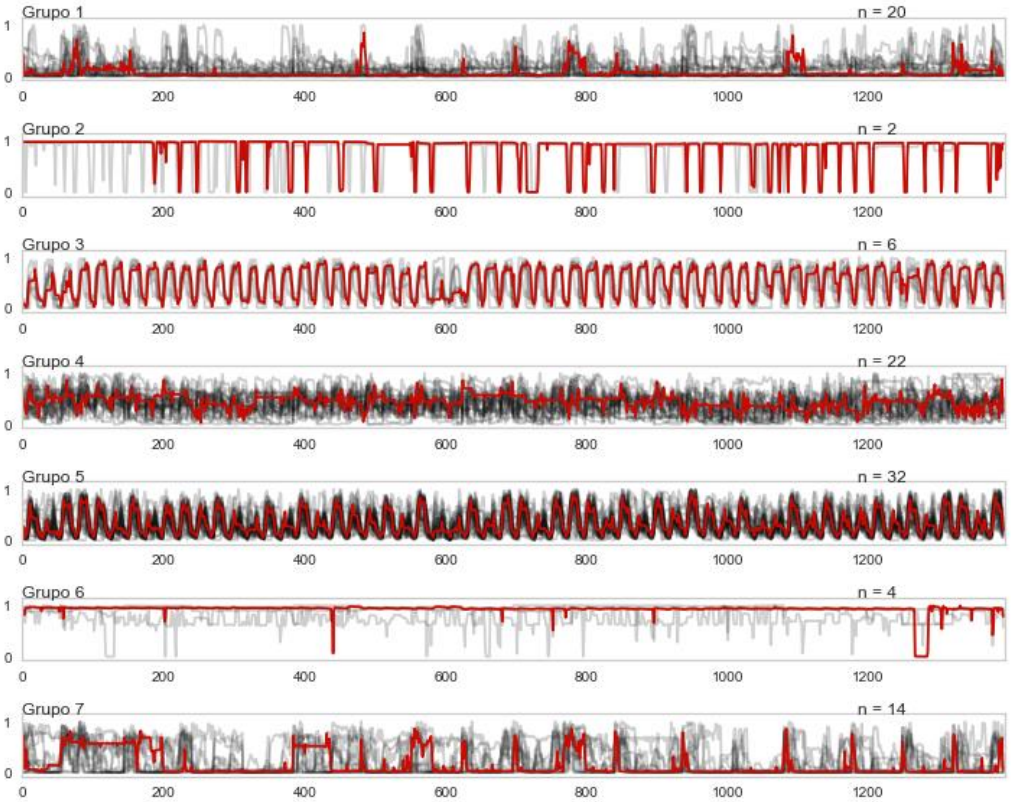
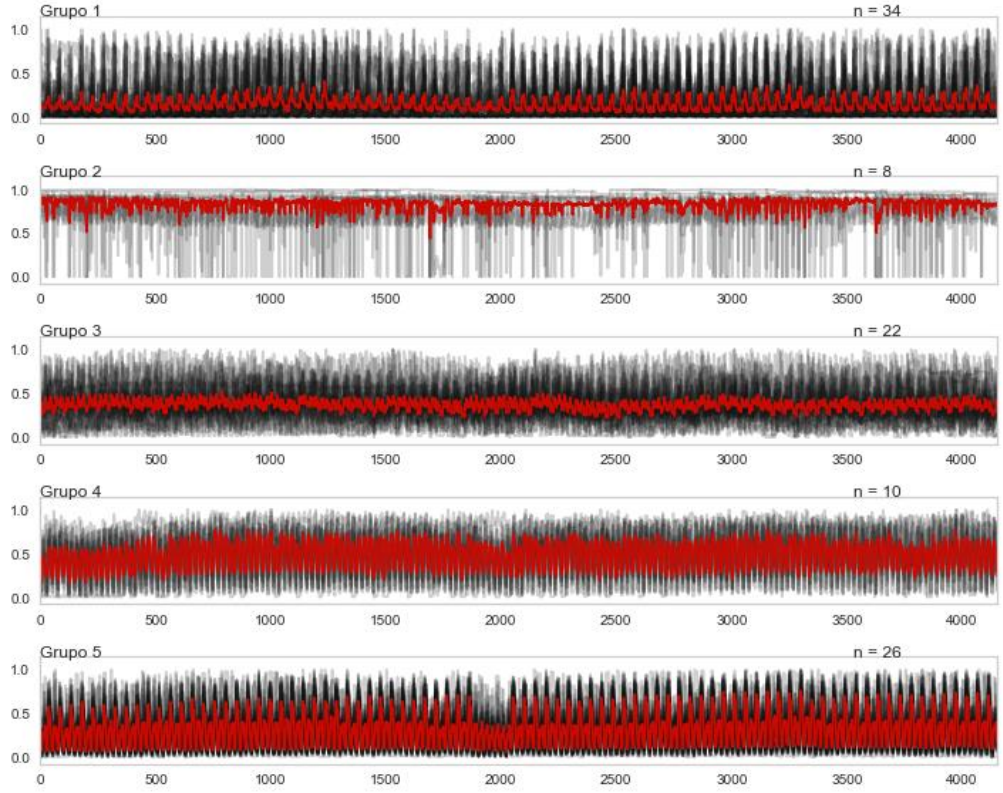
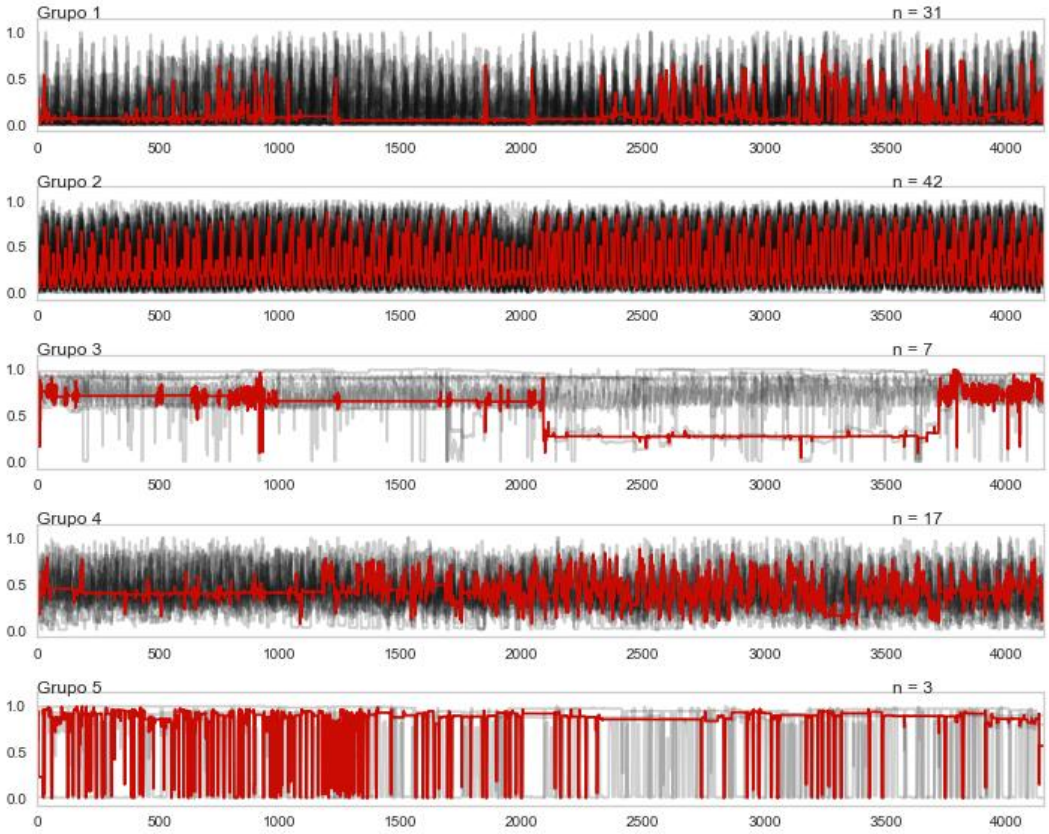


Figura B.12: Agrupamiento del Grupo de los días de fin de semana con K = 5 y los 3 métodos de evaluación.

Euclidean k-means



DBA k-means



Soft-DTW k-Means

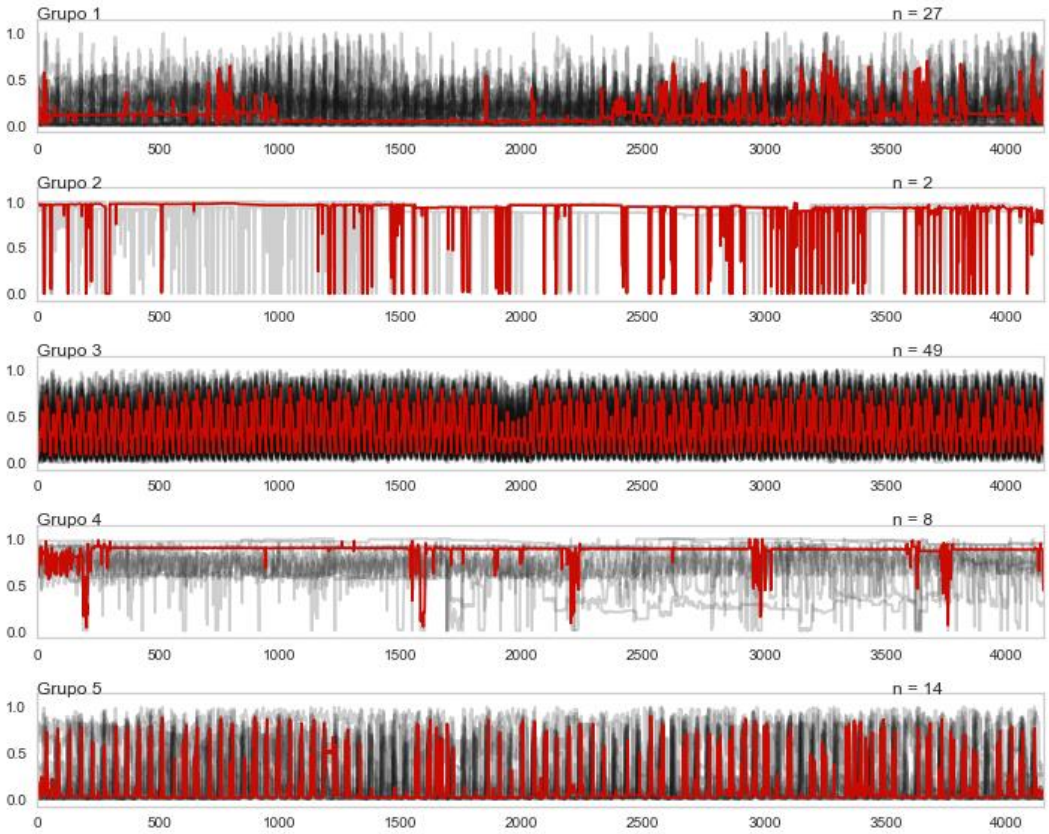


Figura B.13: Seleccionado el índice de validación: Euclidean k-Means, siendo los valores de los grupos para cada cliente y los valores de los centroides los mostrados a continuación:

```
y_FdS_km_5=y_pred_km
y_FdS_km_5
```

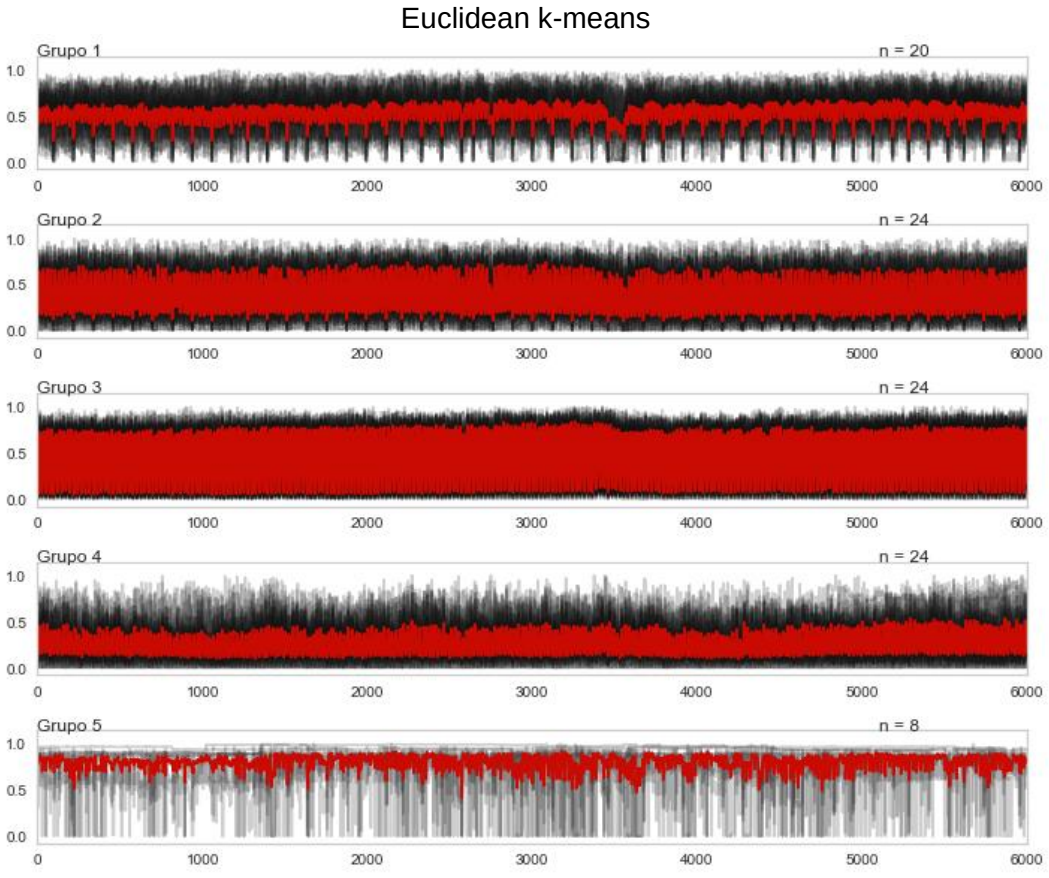
```
array([0, 0, 1, 4, 2, 0, 3, 0, 4, 0, 4, 3, 2, 0, 0, 1, 4, 2, 4, 2, 4, 2,
       3, 3, 0, 4, 0, 2, 2, 3, 0, 0, 2, 4, 2, 0, 1, 2, 0, 0, 0, 0, 4, 4,
       0, 1, 0, 4, 1, 1, 0, 1, 3, 2, 4, 4, 0, 4, 0, 0, 4, 4, 0, 3, 4, 4,
       4, 3, 3, 2, 1, 4, 0, 4, 4, 2, 4, 2, 0, 0, 0, 0, 2, 2, 0, 0, 2, 0,
       2, 3, 2, 2, 0, 2, 0, 4, 4, 0, 4, 2], dtype=int64)
```

```
meth1_FdS_k5=np.reshape(km.cluster_centers_,(clusters,int(df.size/100)))
Cent_FdS_Eucl_k5=pd.DataFrame(np.array(meth1_FdS_k5))
Cent_FdS_Eucl_k5
```

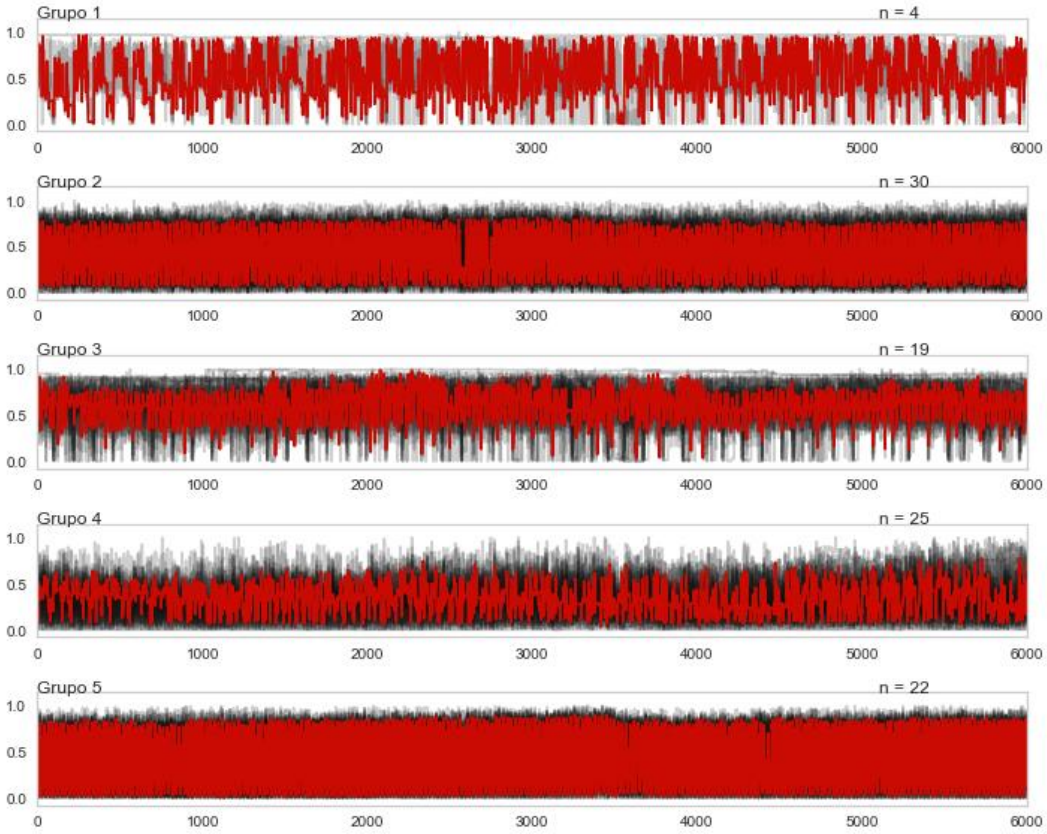
	0	1	2	3	4	5	6	7	8	9	...
0	0.111719	0.125691	0.126323	0.117310	0.107481	0.105462	0.099574	0.096821	0.101290	0.112197	...
1	0.863708	0.849304	0.833801	0.837646	0.840820	0.843262	0.849731	0.880127	0.894043	0.883118	...
2	0.393402	0.371144	0.370755	0.360008	0.357894	0.368644	0.299197	0.259885	0.255925	0.278180	...
3	0.212843	0.184946	0.179140	0.176485	0.168884	0.170422	0.220216	0.340917	0.394489	0.425171	...
4	0.120919	0.094885	0.085983	0.079107	0.077045	0.088577	0.083728	0.129666	0.181973	0.230762	...

5 rows × 4152 columns

Figura B.14: Agrupamiento del Grupo de los días de lunes a viernes con K = 5 y los 3 métodos de evaluación.



DBA k-means



Soft-DTW k-Means

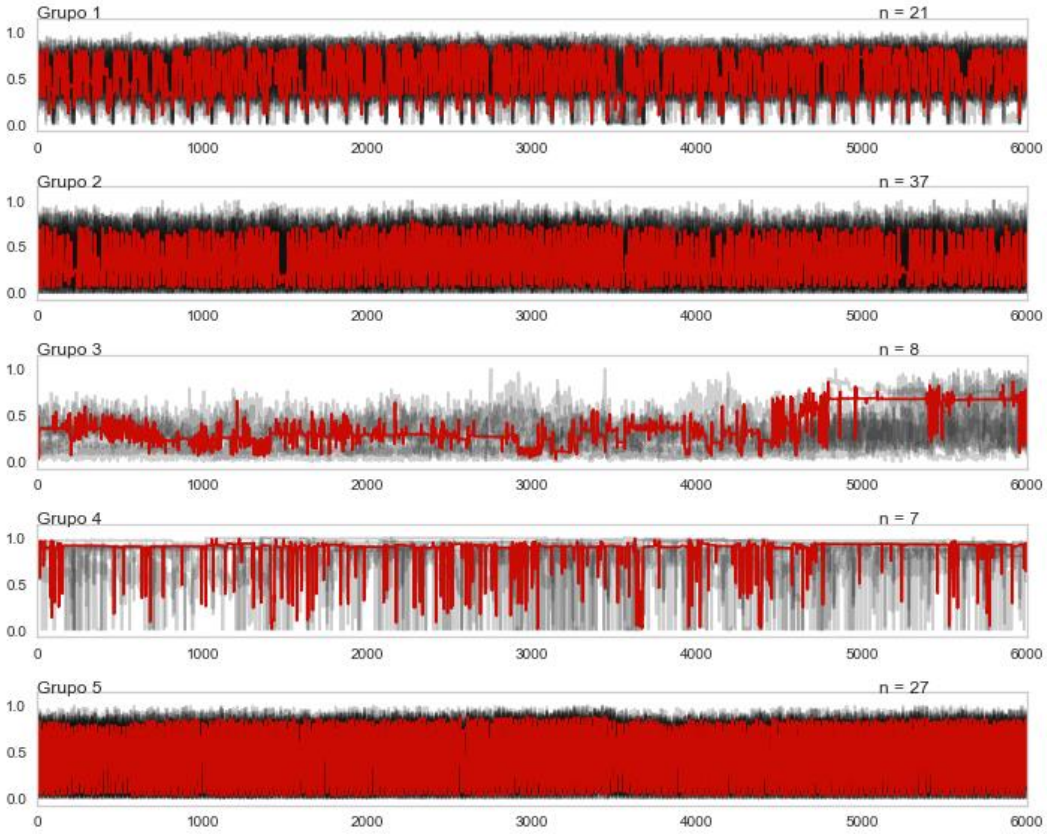


Figura B.15: Seleccionado el índice de validación: DBA k-Means, siendo los valores de los grupos para cada cliente y los valores de los centroides los mostrados a continuación:

```
meth2_LaV_k5=np.reshape(dba_km.cluster_centers_,(clusters,int(df.size/100)))
Cent_LaV_DBA_k5=pd.DataFrame(np.array(meth2_LaV_k5))
Cent_LaV_DBA_k5
```

	0	1	2	3	4	5	6	7	8	9
0	0.542542	0.408781	0.708679	0.805420	0.805298	0.812012	0.532764	0.788696	0.842896	0.864551
1	0.302903	0.171267	0.143829	0.092651	0.180143	0.222306	0.419450	0.555044	0.633439	0.661075
2	0.605402	0.645410	0.601453	0.607248	0.561185	0.452193	0.648372	0.761230	0.798648	0.798623
3	0.255346	0.126033	0.072491	0.154320	0.229526	0.364531	0.512521	0.409298	0.394209	0.394209
4	0.121356	0.053795	0.034120	0.042198	0.040261	0.040730	0.051999	0.119702	0.304631	0.492702

5 rows × 6000 columns

```
y_LaV_sdtw_5=y_pred_sdtw_km
y_LaV_sdtw_5
```

```
array([1, 1, 3, 4, 0, 1, 3, 4, 4, 1, 4, 4, 2, 0, 1, 0, 4, 0, 0, 0, 4, 1,
       4, 0, 0, 4, 1, 1, 2, 4, 0, 0, 2, 1, 2, 1, 3, 4, 1, 2, 0, 0, 4, 4,
       0, 3, 2, 4, 0, 0, 1, 3, 3, 1, 4, 4, 1, 4, 0, 1, 4, 4, 4, 1, 0, 1,
       1, 1, 4, 1, 3, 1, 1, 1, 0, 0, 4, 2, 4, 1, 1, 4, 0, 4, 0, 1, 1, 4,
       1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 4, 1], dtype=int64)
```

ANEXO C

GRUPOS CONFORMADOS DE LA CLASIFICACIÓN DE LOS DÍAS.

Tabla C.1: Detalles de los Clientes Seleccionados de la EEASA

Agencia	N°	Cuenta Del Cliente	Factor De Potencia	Consumo kWh	Nombre Del Cliente	Dirección	Demanda Máxima kW
1	1	189789	0.944791	1245600	PLASTICAUCHO INDUSTRIAL S.A.	PIAIV ETAPA	3024
1	2	143453	0.925238	793423	FAIRIS CA	PANAMERICANA NORTECUNCHIBAMBA	2523.02
1	3	149813	0.982653	469200	EPEMAPAA	QUILLAN ALEMANIA	782
1	4	264869	0.97968	355600	INMOBILIARIA LAVIE S. A	PIO BAROJA Y LOPE DE VEGA	840
1	5	194420	0.931082	261163	TEIMSA SA	SAN JOSESANTA ROSA	885.82
1	6	11196	0.928477	220500	HOSPITAL PROVINCIAL DOCENTE AMBATO	CESAR VITERI Y PASTEUR	427
1	7	227361	0.962739	124600	CIUDAD DEL AUTO CIAUTO CIA LTDA	CAMINO REAL PUCARUMI A.N. MARTINEZ	784
1	8	220156	0.977521	142800	CURTIDURIA TUNGURAHUA	PIA III8	560
1	9	213779	0.954857	113400	ARCOS MIRANDA LIDIA MARLENE	PIAIV Y F	294
1	10	217614	0.975916	119000	MILL POLIMEROS	PUERTO ARTURO	602
1	11	236720	0.992278	106400	EP PETROECUADOR	JULIO CESAR CAÑAR Y RIO MACHANGARA	462
1	12	252472	0.958798	56700	AVIPAZ CIA LTDA AVICOLA	CAMINO REAL SAMANGA	826
1	13	206193	0.99116	106680	AVIPAZ CIALTDA	4 ESQUINASSANTA FESAMANGA	567
1	14	70907	0.993151	99960	INDUSTRIAS CATEDRAL	SAN VICENTEIZAMBA	277.2
1	15	100750	0.970843	97200	HOSPITAL GENERAL AMBATO	RODRIGO PACHANO Y LOS GUAYTAMBOS	252
1	16	191280	0.923814	93240	EPEMAPAA POZO	HUACHI SAN FRANCISCO	147
1	17	121806	0.939793	80960	EL PERAL COMPAÑIA LIMITADA	SAN JOSEPISQUE	193.2

1	18	81656	0.999888	84000	UNIVERSIDAD TECNICA AMBATO	RIO PAYAMINO Y RIO TALATAG	226.8
1	19	8846	0.990992	79920	EPEMAPAA CURIQUINGUE	CARRETERA A BAÑOS	129.6
1	20	102086	0.93068	71930	EPEMAPAA TAMBILLO	UNIDAD NACIONAL TAMBILLO	119.14
1	21	201699		77760	MEDIDOR TOTALIZADOR CT13	AV. CEVALLOS Y JUAN L. MERA	220.8
1	22	95018	0.999955	75600	CNT EP ...	LOS SHYRIS Y CHIAQUITINTA	136.8
1	23	201700		76800	MEDIDOR TOTALIZADOR CT21	M. EGUEZ Y JUAN B. VELA	208.8
1	24	198142	0.958747	81203	MILBOOTS CIALTDA	VCALLE F Y AVENIDA IV	200.88
1	25	267337		61740	GAD MUNICIPALIDAD DE AMBATO	PASO LATERAL Y AGUSTIN GUERRERO	163.8
1	26	136434	0.963159	64680	MILPLAST CIA LTDA	PIACALLE 5 Y AVENIDAD F	277.2
1	27	237699	0.955286	35280	MONTALVO AQUINO DESIDERIO	TIUGUAPICAIHUA	239.4
1	28	252004	0.955779	54600	ALUVIDGLASS S. A	SAN PEDRO SANTA ROSA	644
1	29	201717		56000	MEDIDOR TOTALIZADOR CT03	AV. CEVALLOS Y GUAYAQUIL COL. LUIS A. MARTINEZ	163.2
1	30	201705		48800	MEDIDOR TOTALIZADOR CT05	BOLIVAR Y QUITO	139.2
1	31	146653	0.903601	48354	EPEMAPAA ESTBOMBEO ATAHUALPA	DESTACAMENTO MILITAR NUMBATKAIME	105.64
1	32	226596		36360	EEASA CAMARA 12 DE NOVIEMBRE Y ABDON CALDERON	AV. CEVALLOS Y ABDON CALDERON	109.2
1	33	201710		47840	MEDIDOR TOTALIZADOR CT18	ROCAFUERTE Y LALAMA MEDALLA MILAGROSA	136
1	34	105537	0.997624	44308	GAD MUNICIPALIDAD DE AMBATO MERCADO MODELOIMA	AV. CEVALLOS Y TOMAS SEVILLA	106.49
1	35	115563	0.990889	43260	EMPRE MUNIC MERCADO MAYORISTA AMBATO EMA	EL CONDOR Y LINEA FERREA SUR	121.8
1	36	201696		43320	MEDIDOR TOTALIZADOR CT141	VARGAS TORRES Y AV. 12 DE NOVIEMBRE SS HH	116.4
1	37	32887	0.984587	42560	ESFORSFT	EL PISQUE	103.6

1	38	101842	0.96448	43960	ENI ECUADOR SA	SAN JACINTOMONTALVO	184.8
1	39	201694		41600	MEDIDOR TOTALIZADOR CT22	AV. 12 DE NOVIEMBRE Y ESPEJO EEASA	100
1	40	56868	0.999056	41400	PRODUTEXTI CIA LTDA	PIA IV ETAPA 1	180
1	41	201702		43040	MEDIDOR TOTALIZADOR CT125	MALDONADO Y PRIMERA IMPRENTA	129.6
1	42	201708		40800	MEDIDOR TOTALIZADOR CT118	AYLLON Y PRIMERA IMPRENTA	120
1	43	201715		42720	MEDIDOR TOTALIZADOR CT19	ESPEJO Y BOLIVAR D. EDUCACION	132.8
1	44	201695		41680	MEDIDOR TOTALIZADOR 8219PAD	TOMAS SEVILLA Y CEVALLOS	135.2
1	45	235354	0.96408	41160	GAD MUNICIPAL DE AMBATO	ATAHUALPA Y RIO CUTUCHI	113.4
1	46	134695	0.968277	39060	RUBBERSHOES INDUSTRIAL CIA. LTDA.	PIA 2 Y F	189
1	47	203725		40160	MEDIDOR TOTALIZADOR CT067 ANDINATEL	CASTILLO Y BOLIVAR	99.2
1	48	203718		40440	MEDIDOR TOTALIZADOR CT99	DARQUEA Y 5 DE JUNIO	108
1	49	203719		36000	MEDIDOR TOTALIZADOR CT263	M. EGUEZ Y JUAN B. VELA	120
1	50	139815		38400	PLASTICAUCHO INDUSTRIAL S.A.	R PACHANO Y CARTAGO LA VICTORIA	192
1	51	203717		36840	MEDIDOR TOTALIZADOR CT93 CT93	TOMAS SEVILLA Y BOLIVAR	109.2
1	52	252613	0.999308	45150	ANDESFOOD CIA LTDA	PANAMERICANA NORTE KM 61/2 PISQUE	126
1	53	222593	0.973547	41160	ECUAMATRIZ CIA LTDA	SAN PEDRO SANTA ROSA	214.2
1	54	201713		38400	MEDIDOR TOTALIZADOR CT09	OLMEDO Y CASTILLO GRADAS	103.2
1	55	201720		37440	MEDIDOR TOTALIZADOR CT20	CEVALLOS Y M. EGUEZ TEOFILO LOPEZ	117.6
1	56	203720		37080	MEDIDOR TOTALIZADOR CT077	MALDONADO Y BOLIVAR	104.4

1	57	200666	0.972806	35280	SOLCA NUCLEO DE TUNGURAHUA	PEDRO VASCONEZIZA MBA	112
1	58	245979		34920	CAMARA TOTALIZADORA EEASA	MALDONADO Y BENJAMIN ARAUJO	82.8
1	59	262988		33782	CAMARA TOTALIZADORA CT 14 EEASA	ELOY ALFARO Y FRANCISCO DE ARAUJO	79.97
1	60	193695	0.905846	35924	EPEMAPAA	PASO LATERALTERR EMOTO	61.81
2	1	28795	0.924387	553000	PRODEGEL SA	PACHANLICA BENITEZ PELILEO	1106
2	2	259376	0.969122	473200	BIOALIMENTAR COMPANIA LTA	PACHANLICA - VIA A BENITEZ-PELILEO	1876
2	3	28796	0.95261	149290	HOLVIPLAS SA	PACHANLICA BENITEZ PELILEO	501.82
2	4	41117	0.986975	95340	INDUSTRIAS LACTEAS CHIMBORAZO CIA LTDA	PELILEO GRANDE	256.2
2	5	146966	0.97057	43411	INDUSTRIA INLECHE AMPLIACION	PELILEO GRANDEPELILEO	213.79
2	6	120531	0.912818	31920	GUADALUPE SA	GUADALUPE PELILEO	142.8
2	7	220690	0.988936	22400	BIOALIMENTAR PLANTA	PACHANLICABENITEZPELILEO	145.6
2	8	134035	0.945902	12332	SAILEMA GOMEZ OLIVIA JEANET	EL TAMBO PELILEO	92.72
2	9	177354	0.76906	14572	GAD MUNICIPAL DEL CANTON SAN PEDRO DE PELILEO MERCADO REPUBLICA DE ARGENTINA	JORGE CHACON Y QUIZ QUIZ PELILEO	53.28
2	10	130726	0.992799	15225	INDUSTRIAS LACTEAS CHIMBORAZO CTA LTDA	PELILEO GRANDE	82.48
3	1	163260	0.978623	28927	MOYA SEGUNDO	ROCAFUERTE ATIPILLAHUAZOPILLARO	69.77
3	2	161495	0.915712	23337	SANCHEZ JACOME MARTHA SUSANA	SECTOR ROCAFUERTE PILLARO	87.31
3	3	160597	0.97266	19737	AVALOS JAVIER	SECTOR ROCAFUERTE PILLARO	70.38
4	1	101599	0.883292	54600	CHAVEZ IVAN	AGOYANBAÑOS	285.6

4	2	252661	0.979535	30660	COMPAÑIA BIOPREMIX CIA LTDA	JUIVE LA PAMPA	260.4
4	3	177173	0.948683	21600	CORPELECTRICAC DEL ECUADOR CELECEP	AGOYANBAÑO S	216
4	4	77676	0.999199	17400	CHAVEZ SALOMON IVAN	TOMAS HALFLANTSAV PANAMERICAN A	55.39
4	5	259225	0.909243	10845	INT FOOD SERVICES CORP	16 DE DICIEMBRE Y AMBATO BAÑOS KFC	29
5	1	142975	0.993711	13391	GUEVARA CISNEROS JUAN ABSALÓN	PITULAPATATE	74.63
5	2	179035	0.997188	7058	DIRECCION DISTRITAL 18D04 PATATE SAN PEDRO DE PELILEO SALUD	BELLAVISTA - PATATE	24.07
7	1	53256	0.943257	70560	BRIGADA DE SELVA 17 PASTAZA 4	SHELL	208.8
7	2	50831	0.968236	74256	ARBORIENTE SA	AV CESLAO MARIN CHONTOA	246.84
7	3	329557	0.978234	37950	HOSPITAL BASICO EL PUYO	ANTONIO ACUÑA GONZALO PIZARRO	92
7	4	316993	0.907754	31783	INCUBADURA PASTAZA CHAVEZ SALOMON	KM 25 VIA MADRE TIERRA	80.78
7	5	324974	0.980581	21000	DIRECCION PROV DEL CONSEJO DE LA JUDICATURA	AV ALBERTO ZAMBRANO URBTRUJILLO VEINTIMILLA	75.6
7	6	313579	0.958218	24602	TIENDAS IND ASOCIADAS TIA	MARIN	59.36
7	7	327808	0.989293	19231	CORPORACION FAVORITA C.A. - GRAN AKI PASTAZA	CESLAO MARIN ALVARO VALLADARES	53.8
7	8	50076		18941	CORPORACION NACIONAL DE TELECOMUNICACIO NESCNT EP	VILLAMIL Y FCO DE ORELLANA	39.17
8	1	402709		4936	CORPORACION NACIONAL DE TELECOMUNICACIO NES CNT EP	JUAN LEON MERA Y FLORIDA	11.42
8	2	402272	0.961524	2940	DIRECCION DISTRITAL 14D01-MORONA-SALUD	AV CUMANDA Y ORELLANA	12.6
10	1	252476	0.995219	8731	DIR.DISTRITAL 18D04 PATATE SAN	CEVALLOS CENTRAL	25.7

					PEDRO DE PELILEO SALUD	CENTRO DE SALUD	
10	2	89375	0	7560	UNIVERSIDAD TECNICA AMBATO	QUEROCHACA-CEVALLOS	25.2
10	3	266531	0.79515	4609	GUERRERO BARONA MONICA SHISEL	S N Y JUAN GUEVARA AIRE LIBRE CEVALLOS	53.78
11	1	466741	0.955857	31456	TIENDAS INDUSTRIALES ASOCIADAS TIA SA	AV 15 DE NOVIEMBRE Y AV DEL CHOFER ESQUINA	89.35
11	2	467364	0.999935	24640	CENTRO CLINICO QUIRURGICO AMBULATORIO HOSPITAL DEL DIA	LOT HUERTOS FAMILIARES VIA PTO NAPO	117.6
11	3	481049	0.894427	10920	GAD MUNICIPAL DE TENA PLANTA DE TRATAMIENDO DE AGUAS	PALANDACOCHA	67.2
11	4	450001	0.998949	18589	GOBIERNO AUTONOMO DESCENTRAILZADO PROVINCIAL DE NAPO	JUAN MONTALVO Y OLMEDO	107.46
11	5	456413	0.979939	20794	COS 2	AV DOS RIOS	81.12
12	1	482211	0.919124	18048	TIA ARCHIDONA	C NAPO Y C JONDACHI	68.53
12	2	454152	0.866033	659	MC GHEE MARVIN GLENN	SHICAMA ALTO Y BAJO	6.55
12	3	454603	0	8037	GOBIERNO MUNICIPAL ARCHIDONA	AV. NAPO ENTRE TRASV. 15 y 16	35.09
12	4	472782	1	2152	AMI RUNA ECUADOR LLC	VIA RUKULLACTA ESCUELA	16.34

Tabla C.2: Grupos Conformados de los Fines de Semana

CONFORMACIÓN DE GRUPOS DE LOS FINES DE SEMANA				
Grupo 1 34 clientes	**Grupo 2 8 clientes**	**Grupo 3 22 clientes**	**Grupo 4 10 clientes**	**Grupo 5 26 clientes**
1_1	1_3	1_5	1_7	1_4
1_2	1_16	1_13	1_12	1_9
1_6	1_37	1_18	1_23	1_11
1_8	1_46	1_20	1_24	1_17
1_10	1_49	1_22	1_30	1_19
1_14	1_50	1_28	1_53	1_21
1_15	1_52	1_29	2_5	1_26
1_25	3_2	1_33	2_9	1_34
1_27		1_35	2_10	1_43
1_31		1_38	10_1	1_44
1_32		1_54		1_48
1_36		3_1		1_55
1_39		4_4		1_56
1_40		5_1		1_58
1_41		7_4		2_2
1_42		7_5		2_3
1_45		7_8		2_6
1_47		8_2		2_7
1_51		10_2		2_8
1_57		10_3		3_3
1_59		11_2		4_2
1_60		12_3		4_3
2_4				4_5
4_1				11_4
5_2				11_5
7_1				12_2
7_2				
7_3				
7_6				
7_7				
8_1				
11_1				
11_3				
12_1				

Tabla C.3: Grupos Conformados de los días de lunes a viernes

Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
4 clientes	**30 clientes**	**19 clientes**	**25 clientes**	**22 clientes**
1_32	1_1	1_3	1_2	1_4
1_45	1_6	1_5	1_10	1_8
1_52	1_12	1_7	1_13	1_9
7_6	1_15	1_14	1_22	1_11
	1_18	1_16	1_28	1_17
	1_25	1_19	1_29	1_21
	1_27	1_20	1_31	1_23
	1_36	1_24	1_33	1_26
	1_38	1_37	1_34	1_30
	1_39	1_41	1_35	1_43
	1_54	1_42	1_40	1_44
	1_60	1_46	1_47	1_48
	2_5	1_49	1_51	1_55
	2_6	1_50	1_57	1_56
	2_8	1_53	2_7	1_58
	2_9	1_59	3_1	2_2
	2_10	3_2	4_2	2_3
	3_3	4_4	5_1	2_4
	4_1	7_4	7_7	4_5
	4_3		8_2	5_2
	7_1		10_2	8_1
	7_2		10_3	12_2
	7_3		11_1	
	7_5		11_2	
	7_8		11_4	
	10_1			
	11_3			
	11_5			
	12_1			
	12_3			

ANEXO D

PROGRAMACIÓN DESARROLLADA EN JUPYTERLAB DE LOS CLIENTES FRAUDULENTOS.

- En el formato pdf con nombre Anexo D.

ANEXO E

MODELOS DE RED NEURONAL REALIZADA PARA LOS DÍAS DE LUNES A VIERNES Y LOS FINES DE SEMANA.

Figura E.1: Modelo de Red Neuronal de los Fines de Semana:

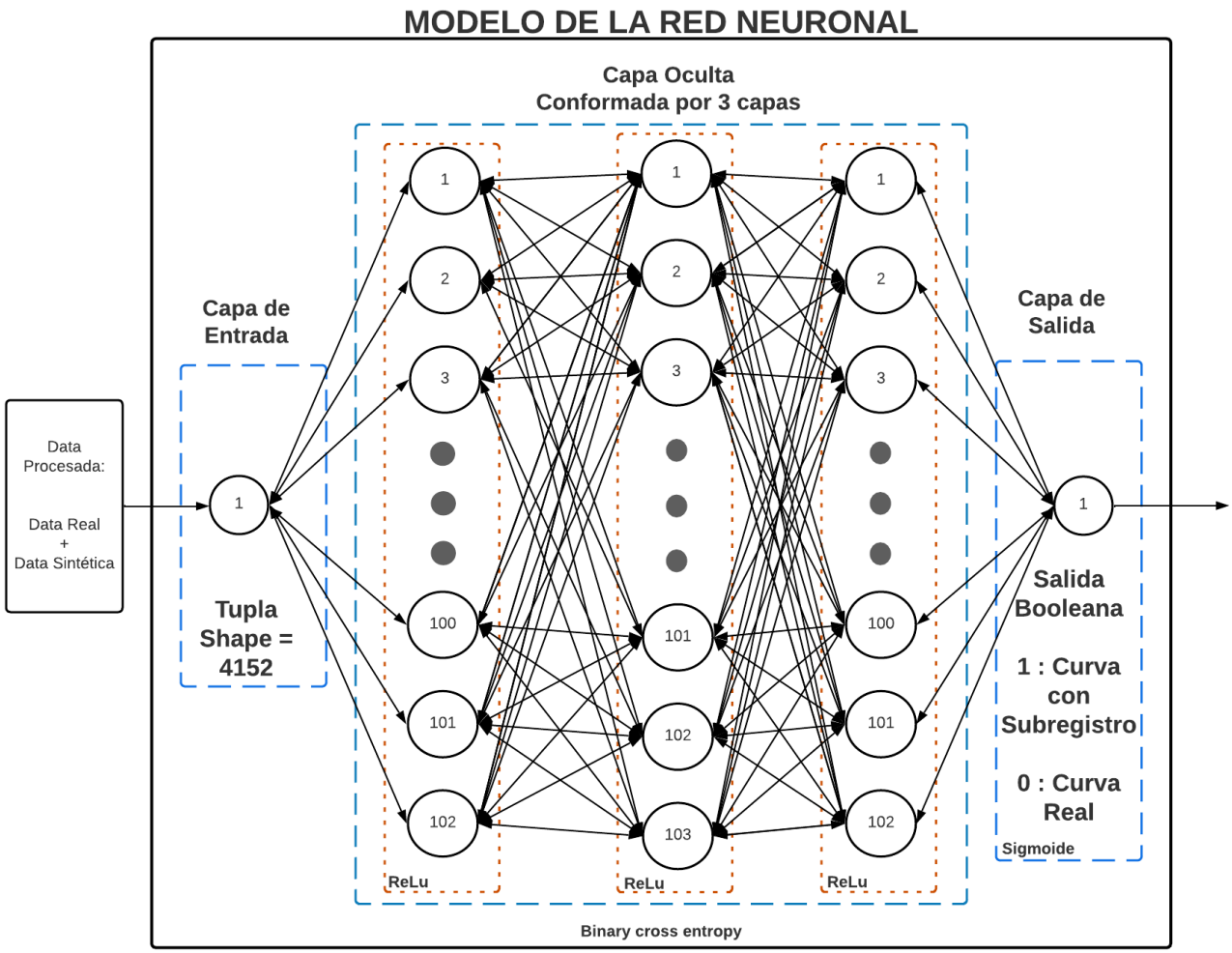


Figura E.2: Diagrama en KNIME de los días de Fin de Semana.

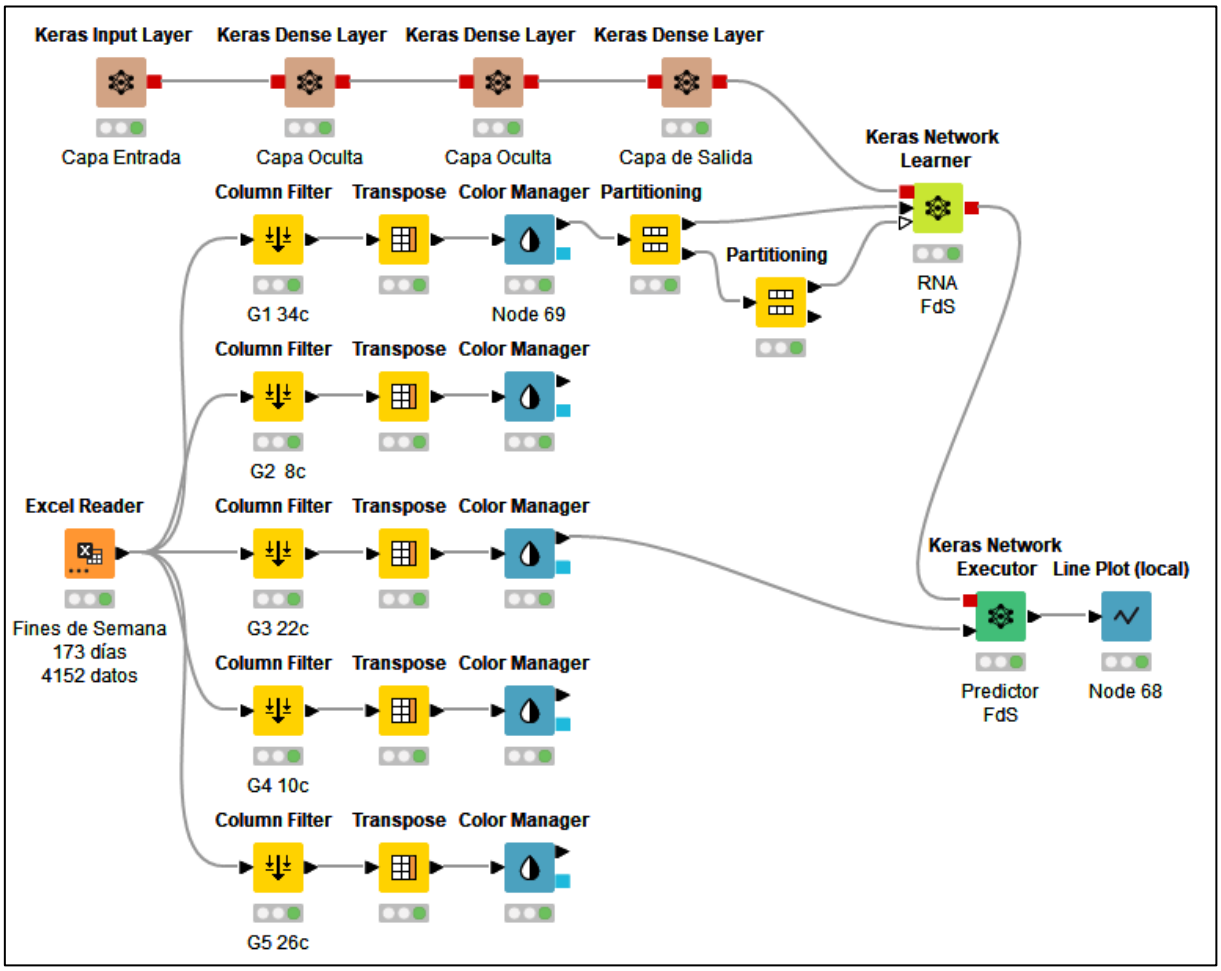


Figura E.3: Modelo de Red Neuronal de días de lunes a viernes.

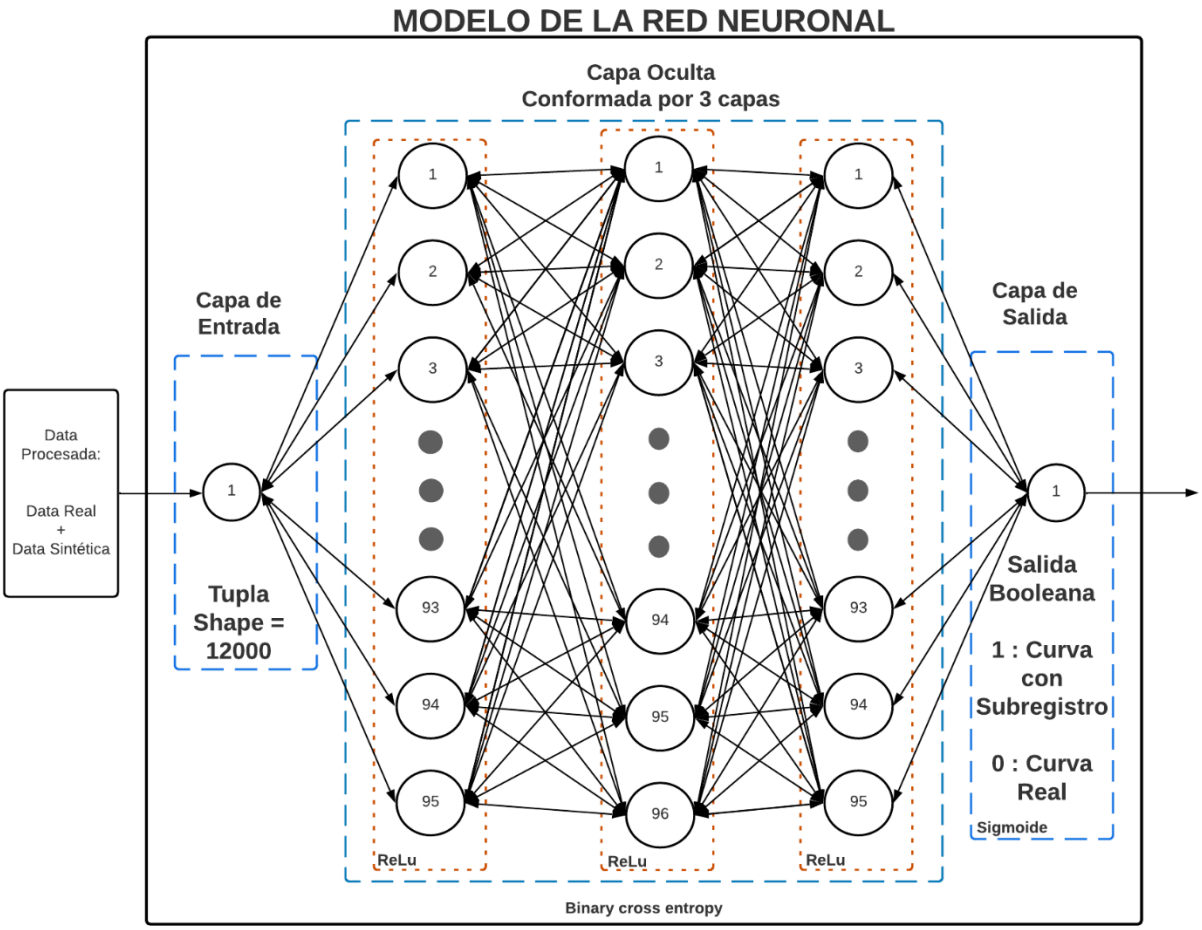
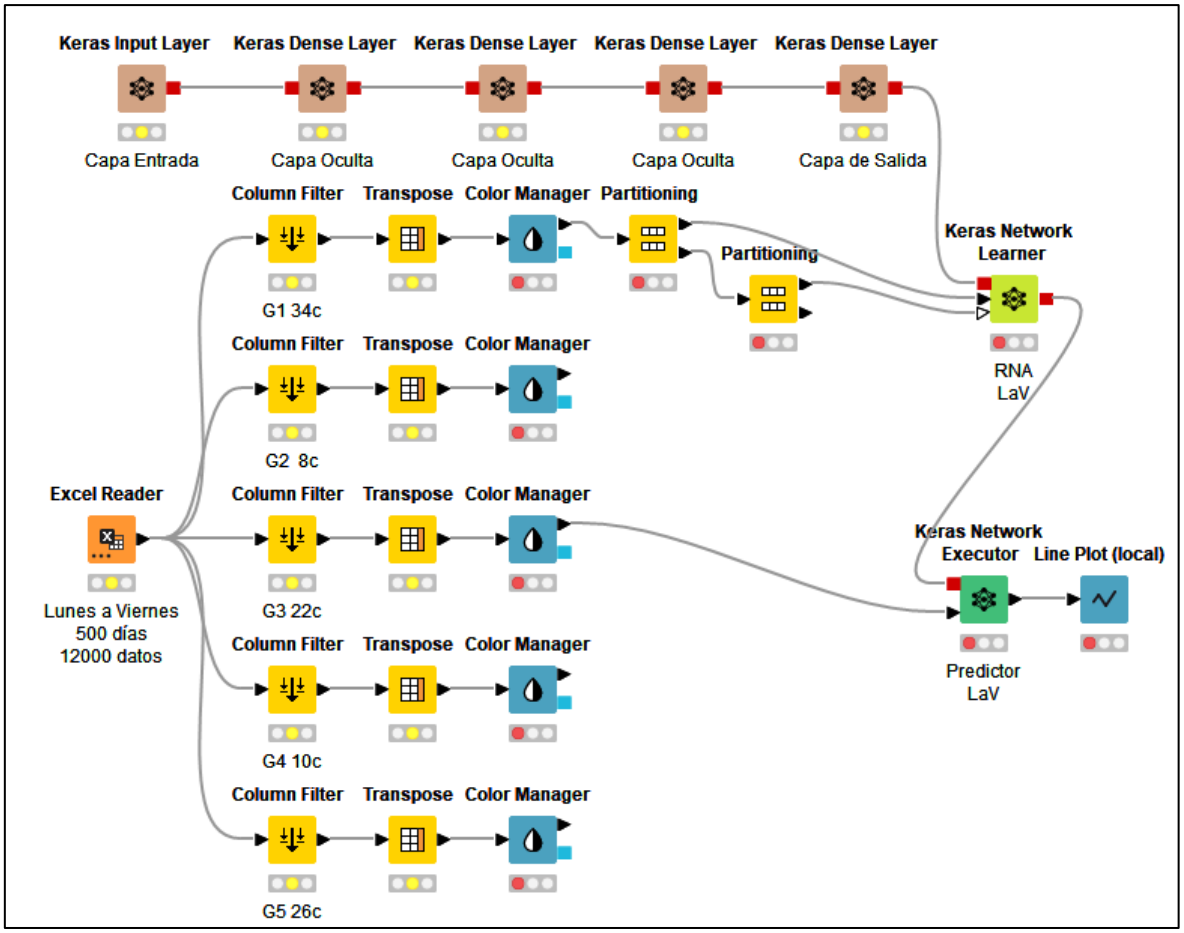


Figura E.4: Diagrama en KNIME de los días de lunes a viernes.



ORDEN DE EMPASTADO