**Instituto Tecnológico y de Estudios Superiores de Occidente**

**Departamento de Matemáticas y Física**

**Sustentabilidad y tecnología**

**PROYECTO DE APLICACIÓN PROFESIONAL (PAP)**

**Programa de Modelación Matemática para el Desarrollo de Planes y Proyectos de Negocio**



**PAP 4J09: Ciencia de datos aplicado a información de negocios y desarrollo social**

**Pronóstico general del sistema eléctrico (GCROC, Gerencia de Control Regional de Occidental)**

**PRESENTAN**

Programas educativos y Estudiantes

Ing. En Nanotecnología Gregorio Alberto Alvarez Alvarez   
Ing. En Nanotecnología Alan Jesús Cortes de la Torre  
Ing. Financiera Diana Laura Ramírez Hinojosa

Profesor PAP: Pablo Benavides Herrera

Tlaquepaque, Jalisco, mayo de 2020

Contenido

[REPORTE PAP 2](#_Toc39938396)

[Presentación Institucional de los Proyectos de Aplicación Profesional 2](#_Toc39938397)

[Resumen 4](#_Toc39938398)

[1. Ciclo participativo del Proyecto de Aplicación Profesional 4](#_Toc39938399)

[1.1 Entendimiento del ámbito y del contexto 5](#_Toc39938400)

[1.2 Caracterización de la organización 5](#_Toc39938401)

[1.3 Identificación de la(s) problemática(s) 6](#_Toc39938402)

[1.4. Planeación de alternativa(s) 7](#_Toc39938403)

[1.5. Desarrollo de la propuesta de mejora 8](#_Toc39938404)

[1.5.1 Variables Ficticias 8](#_Toc39938405)

[1.5.2 Serie de Fourier 9](#_Toc39938406)

[1.5.3 Regresores no lineales 9](#_Toc39938407)

[1.5.4 Separación de base de datos por día de la semana 12](#_Toc39938408)

[1.5.5 Lógica difusa 12](#_Toc39938409)

[1.5.6 Modelos de variables climáticas 14](#_Toc39938410)

[1.5.7 Efecto en el consumo por días contiguos a los festivos 15](#_Toc39938411)

[1.5.8 ARIMA 16](#_Toc39938412)

[1.5.9 Bosques aleatorios 17](#_Toc39938413)

[1.5.10 Predicción de un año 18](#_Toc39938414)

[1.5.11 Implementación de algoritmo PCA a variables climáticas 18](#_Toc39938415)

[1.6. Valoración de productos, resultados e impactos 18](#_Toc39938416)

[1.6.1 Valoración de productos 18](#_Toc39938417)

[1.6.2 Resultados 21](#_Toc39938418)

[1.7. Bibliografía y otros recursos 28](#_Toc39938419)

[2. Productos 29](#_Toc39938420)

[3. Reflexión crítica y ética de la experiencia 32](#_Toc39938421)

[3.1 Sensibilización ante las realidades 32](#_Toc39938422)

[3.2 Aprendizajes logrados 33](#_Toc39938423)

[3.2.1 Diana 33](#_Toc39938424)

[3.2.2 Alan 34](#_Toc39938425)

[3.2.3 Gregorio 34](#_Toc39938426)

# REPORTE PAP

## Presentación Institucional de los Proyectos de Aplicación Profesional

*Los Proyectos de Aplicación Profesional (PAP) son experiencias socio-profesionales de los alumnos que desde el currículo de su formación universitaria- enfrentan retos, resuelven problemas o innovan una necesidad sociotécnica del entorno, en vinculación (colaboración) (co-participación) con grupos, instituciones, organizaciones o comunidades, en escenarios reales donde comparten saberes.*

*El PAP, como espacio curricular de formación vinculada, ha logrado integrar el Servicio Social (acorde con las Orientaciones Fundamentales del ITESO), los requisitos de dar cuenta de los saberes y del saber aplicar los mismos al culminar la formación profesional (Opción Terminal), mediante la realización de proyectos profesionales de cara a las necesidades y retos del entorno (Aplicación Profesional).*

*El PAP es un proceso acotado en el tiempo en que los estudiantes, los beneficiarios externos y los profesores se asocian colaborativamente y en red, en un proyecto, e incursionan en un mundo social, como actores que enfrentan verdaderos problemas y desafíos traducibles en demandas pertinentes y socialmente relevantes. Frente a éstas transfieren experiencia de sus saberes profesionales y demuestran que saben hacer, innovar, co-crear o transformar en distintos campos sociales.*

*El PAP trata de sembrar en los estudiantes una disposición permanente de encargarse de la realidad con una actitud comprometida y ética frente a las disimetrías sociales. En otras palabras, se trata del reto de “saber y aprender a transformar”.*

*El Reporte PAP consta de tres componentes:*

*El primer componente refiere al ciclo participativo del PAP, en donde se documentan las diferentes fases del proyecto y las actividades que tuvieron lugar durante el desarrollo de este y la valoración de las incidencias en el entorno.*

*El segundo componente presenta los productos elaborados de acuerdo con su tipología.*

*El tercer componente es la reflexión crítica y ética de la experiencia, el reconocimiento de las competencias y los aprendizajes profesionales que el estudiante desarrolló en el transcurso de su labor.*

## Resumen

Se creo un algoritmo con tres fases para la predicción de datos de la gerencia general de CENACE occidente. Primeramente, se analizaron las series de tiempo para mejorar el modelo anteriormente creado. Se hizo uso de predicción de variables climáticas, variables de días especiales, variables obtenidas de regresión logística y de análisis de espectro singular.

Se compararon los resultados con los obtenidos del algoritmo anterior y un algoritmo de bosques aleatorios. Se obtuvieron mejores resultados que con el modelo anterior. Se obtuvo un MAPE menor con el modelo de bosques aleatorios que con el modelo mejorado; sin embargo, con el primero se obtuvieron errores diarios más grandes que con el segundo.

Se creo otro modelo para predecir consumo de los días de cuarentena.

## Ciclo participativo del Proyecto de Aplicación Profesional

El PAP es una experiencia de aprendizaje y de contribución social integrada por estudiantes, profesores, actores sociales y responsables de las organizaciones, que de manera colaborativa construir sus conocimientos para dar respuestas a problemáticas de un contexto específico y en un tiempo delimitado. Por tanto, la experiencia PAP supone un proceso en lógica de proyecto, así como de un estilo de trabajo participativo y recíproco entre los involucrados.

El reporte presentado expone lo realizado durante el semestre de primavera 2020. Se pretende explicar, primero que nada, el contexto en donde se desarrolla, la problemática a solucionar, los objetivos a alcanzar, el producto esperado, la teoría necesaria para entender el proyecto, para finalizar con el proceso que se desarrolló a lo largo del semestre en conjunto con profesores del ITESO, el CENACE, y los aprendizajes obtenidos.

Debido a que el semestre anterior se obtuvieron resultados satisfactorios para el CENACE, se procedió a observar las desventajas del modelo obtenido. Con la información recabada de este paso, se diseñó el proyecto que se llevó a cabo en el presente semestre. Basados en dicho plan, se modificó la metodología a lo largo del semestre a base de prueba y error, con el fin de mejorar el modelo. La idea central fue crear un algoritmo en donde todo se pudiera explicar mediante un fundamento teórico, evitando prácticas que nos llevaran a obtener una “Caja Negra”.

## Entendimiento del ámbito y del contexto

La predicción del consumo energético de la Gerencia Regional Occidental de Control (GROC) representa un reto considerable para el Centro Nacional de Control de Energía (CENACE). La magnitud de este reto radica a nivel nacional, se está hablando de 9 estados de la República, en donde un mal pronóstico crearía dos de las opciones siguientes:

* Pronosticar una demanda menor a la real, llevando a que se genere energía de improvisto, elevando el costo ya establecido por los proveedores, pues involucra calentar las maquinarias necesarias rápidamente.
* Pronosticar una demanda de energía mayor a la real, dando como resultado una mayor producción de la necesaria, es decir, pérdidas para la empresa por tener que reservar la energía para usar en otro momento.

Ambas situaciones no son favorables para el CENACE debido al costo energético, lo que se traduce en perdidas millonarios tanto para el sector mayorista, como para la población en general.  Debido a esto se precisa contar con modelos que permitan obtener predicciones más acertadas al consumo real, disminuyendo con esto las pérdidas para el CENACE. Aunque el error actual es muy bajo, siempre hay espacio para reducirlo, y mejorar las condiciones actuales del pronóstico.

En retrospectiva, todos los habitantes de los estados que controla la GROC son los beneficiarios de la resolución del problema, y los afectados en caso de que algo fallara. El trabajo del CENACE es fundamental para un bien tan necesario como la energía.

## 1.2 Caracterización de la organización

El Centro Nacional de Control de Energía es un organismo público descentralizado cuyo objeto es ejercer el Control Operativo del Sistema Eléctrico Nacional (CENACE, 2019). La misión fundamental es “Sustentar eficientemente las transacciones de los productos de la electricidad entre generadores y consumidores” (CENACE, 2019).

Debido a la aprobación de la reforma energética en 2013 por el gobierno del presidente Enrique Peña Nieto, el CENACE, pasa a ser un organismo dependiente del gobierno encargado de la regulación del mercado mayoritario energético, esto supone la mediación en la compra/venta, regulación en los precios entre otros. Con la finalidad de tener un control óptimo de sus operaciones el CENACE divide el país en 9 gerencias regionales, las cuales se encargan de la regulación en varios estados que pudiesen estar conectados de alguna manera.

FormaLa gerencia de Occidente (en la que el PAP se está llevando a cabo) se conforma por 9 estados de la República que son: Jalisco, Nayarit, Zacatecas, Aguascalientes, Colima, Guanajuato, Michoacán, Querétaro y San Luis Potosí.

Cuadro de textoEl departamento de Evaluación y Estadística se dedica a hacer pronósticos de la demanda de energía, lo cual sirve para abastecer de manera correcta a todas las localidades dentro de los estados antes mencionados. El error del pronóstico que ellos manejan a nivel gerencial es entre 1% y 3% en MAPE (más adelante se detallará acerca de esta medida), siendo el 4% el máximo que pueden alcanzar en días comunes.

Dentro del departamento de Evaluación y Estadística, se encuentra el jefe de este, Camilo Narváez, el cual es el encargado de la relación del CENACE con ITESO. Durante el desarrollo de nuestro proyecto, el supervisor operativo del departamento, Marcos Emmanuel Padilla González, nos apoyó con retroalimentación diaria, de manera que se pudiera ajustar el modelo desarrollado de una mejor manera.

## 1.3 Identificación de la(s) problemática(s)

Dado que no fue el primer semestre trabajando en este PAP, se conocía el problema más frecuente, que es la falta de actualización de los datos por parte del CENACE. No obstante, se decidió desde la primera semana del semestre, por parte del equipo al igual que por parte del CENACE, actualizar los datos de manera constante, haciendo el compromiso de hacer pronósticos diarios y mandarlos para poder tener un registro del desarrollo del proyecto.

Teniendo solucionado lo anterior, las problemáticas de este semestre se basaron en la que el semestre pasado le faltó al proyecto: optimización, las variables de días festivos y de clima, así como la búsqueda de variables no lineales.

Aunado a esto, surgieron problemáticas referentes al pronóstico de días especiales (como son conocidos en el CENACE) por fechas de las que no se tenía anterior registro: El paro del día de la mujer, y el periodo que aún se vive de distanciamiento social por el COVID-19.

## 1.4. Planeación de alternativa(s)

Debido a la necesidad de CENACE por correr los modelos en corto tiempo, se propuso el uso de otras herramientas de Python, así como alternativas de los algoritmos utilizados para crear el modelo. Esto con el fin de disminuir el tiempo de cómputo en CENACE, así como para optimizar el tiempo de los experimentos que se llevaron a cabo. Igualmente, se buscó un orden del código que fuera más sencillo de comprender para el usuario.

Los objetivos establecidos para este semestre, y que fueron seguidos a lo largo del desarrollo del proyecto son los siguientes:

* Objetivo general:
  + - Reducir el error de los pronósticos del modelo creado.
* Objetivos específicos:
  + - Agregar regresores no lineales al modelo.
    - Encontrar la manera de incluir el clima como regresor.
    - Trabajar con nuevas librerías para hacer la regresión.
    - Mejorar modelo para residuales.
    - Aplicar mismo modelo a zonas de carga diarias.
    - Optimización del modelo.

## 1.5. Desarrollo de la propuesta de mejora

### 1.5.1 Variables Ficticias

Conocidas como dummies, se utilizan para explicar variables cuantitativas, ya que las regresiones aceptan solamente números, es decir, variables cuantitativas. Sirven para que el modelo distinga una característica que puede tener dos o más opciones, por ejemplo, el sexo de una persona. ¿Cómo se utiliza? Con números binarios. Tan simple como agregar una variable que contenga un 1 si es mujer y un 0 si es hombre, o viceversa. En el caso de la demanda de energía, había factores a considerar como los días de la semana, los meses del año, y los días festivos. Esto llevaba a crear muchas variables para que el programa distinguiera estos parámetros, quedando de la siguiente manera:

* Días de la semana: Seis variables dummies
* Mes del año: Once variables dummies
* Días festivos: Los días festivos fueron divididos en dos secciones:
  + - Aquellos que solo se veían afectados el mismo día:
      * Día del Trabajo
      * Elecciones
      * Los Óscares
    - Aquellos que parecían tener un efecto en días anteriores y posteriores (realizado con lógica difusa explicada en la sección 1.5.5):
      * + Navidad
        + Nochebuena
        + Año nuevo
        + Año viejo
        + Jueves santo
        + Viernes santo
        + Sábado santo
        + Domingo santo
        + Independencia de México
        + Día de la madre
        + Día de la constitución
        + Día de la virgen de Zapopan
        + Día de la virgen de Guadalupe
        + Día de muertos
        + Toma de poder presidencial
        + SuperBowl
        + Mundial de Futbol
        + Revolución mexicana
        + Natalicio Benito Juárez

### 1.5.2 Serie de Fourier

El teorema de Fourier establece lo siguiente:

*Cualquier función periódica, con periodo T, se puede representar como suma de sinusoides de frecuencias f, 2f, 3f, .., llamadas armónicos. (La relación entre el periodo y la frecuencia es f=1/T) (Barrière, 2011).*

En pocas palabras, la serie de Fourier busca descomponer una función en sumas de senos y cosenos, con frecuencias de las señales de la función inicial. En el modelo que se buscó desarrollar, una parte fundamental fue el análisis de las frecuencias para poder determinar los armónicos a usar en l descomposición de la demanda de energía. Es importante recalcar que este teorema se puede utilizar por la forma de la serie, ya que es periódica.

### 1.5.3 Regresores no lineales

Se sabe que al implementar algoritmos de regresión lo que se obtiene es una regresión lineal de la forma en dónde cada X representa cada una de las variables o regresores que conforman el modelo, y cada ꞵ es una constante determinada por el mismo algoritmo.

Sin embargo, puede haber relaciones entre variables que no necesariamente sean lineales, por ejemplo, puede ser que el comportamiento que se intentan modelar corresponda a la forma y una regresión lineal por sí sola no puede modelar estas relaciones. Lo que se hace en estos casos es agregar una nueva variable (o nueva columna) que sea calculada a partir de las relaciones no lineales entre los regresores iniciales. De esta manera el algoritmo de regresión nos regresará una combinación lineal en dónde una de las X describirá una relación no lineal, manteniendo la simplicidad de una regresión lineal.

El problema ahora reside en poder encontrar dichas relaciones, y para ello se propuso implementar el uso de la regresión simbólica. Esta se puede definir como un algoritmo de búsqueda de funciones que describan el comportamiento de puntos dados. Esta clase de regresión tiene la ventaja de que se pueden escoger las funciones que se tomarán en cuenta para esta búsqueda, como senos y cosenos, funciones exponenciales, logarítmicas, o incluso alguna función definida por el propio usuario.

Una vez que la regresión simbólica determinó las relaciones funcionales entre las variables, pueden ser implementadas como regresores en la regresión lineal como se expuso anteriormente. Las herramientas utilizadas para esto fueron el usó la librería de Python “gplearn” y el software Eureqa Pro, siendo este último la herramienta principal utilizada.

Primero se definieron las variables que usarían en este proceso y como se añadirían al modelo para que captara la mayor información posible sin necesidad de sobrecargarlo con regresores.

Las variables que se usaron fueron:

1. Temperatura máxima (max)
2. Temperatura mínima (min)
3. Velocidad del viento (velocidad)
4. Nubosidad (nubosidad)
5. Cantidad de lluvia (lluvia)
6. Día de la semana (day)
   1. En este caso de decidió que este regresor fuera una variable categórica en la cuál iba de lunes a domingo en dónde el lunes era representado por un 0 hasta llegar al domingo representado por un 6.
   2. Esto se hizo para que le modelo tomara la información de todos los días en vez de tomar solamente los días que consideraba más relevantes.
7. Mes del año (month)
   1. Al igual que el día de la semana, se consideró este regresor como una variable categórica con valores del 1 al 12 representando los meses de enero a diciembre.
8. Semana del año (week)
   1. Al igual que el día de la semana, se consideró este regresor como una variable categórica con valores del 1 al 53 representando el número de la semana en la que se encuentra la fecha dada en el año.
9. Tiempo (num)
   1. Una variable que va del 1 hasta n, siendo n el número de entradas que tenemos registradas.
10. Festivo (festivo)
    1. Variable binaria en dónde 1 representa que el día en cuestión se considera festivo y el 0 representando que no se considera festivo.

Este proceso se repitió 2 veces, la primera vez se usaron las variables antes descritas sin mayor modificación, mientras que la segunda vez se les aplicó un proceso de normalización en el que se dividieron todos los datos (de cada variable) entre su máximo registrado. De esta manera obtenemos regresores no lineales que no necesitan de normalización de las variables que los integran y regresores que si necesitan de este paso previo.

Las regresores no lineales que no necesitan de normalización previa a su cálculo son:

Mientras los regresores que si necesitan de la normalización se muestran a continuación:

### 1.5.4 Separación de base de datos por día de la semana

Con el método de series de potencia, se busca encontrar las periodicidades ocultas y sumarlas a modo de series de Fourier, que en nuestro modelo son regresores de la regresión lineal.

Se busca una forma con la que el algoritmo de regresión lineal sea capaz de encontrar de manera más precisa, los pesos para de los regresores más importantes sin que el modelo deje de predecir adecuadamente, esto se logra al disminuyendo el número de periodicidades (regresores) así disminuyendo el ruido.

Aunque modelos que disminuyen el peso de los regresores, como Lasso o Ridge, evitando el un sobreajuste, no se logra despreciar completamente el efecto de todos los regresores, por lo que se procedió a seleccionar dichas variables separando la base de datos, en espacios de periodicidades conocidas, como son los días de la semana o horas del día.

En nuestro caso debido a la naturaleza de los datos, se procedió a separar los datos por día de la semana. Este proceso desarrollado fue una idea por la que se optó, sin embargo, no resultó favorecedora a los pronósticos y se eliminó.

### 1.5.5 Lógica difusa

Al utilizar los dummies, se obtiene una especie de regresor de tipo booleano, Sí o no, 1 o 0, Verdadero o Falso. El problema con estos tipos de datos es que, al aplicarlos en muchos casos reales, no funcionan. La razón es que el dato puede ser poco claro, puede contener un poco de verdad o de falso, y reducirlo a un término binario causaría términos inciertos.

La lógica difusa permite dar valores que sean parcialmente verdaderos y falsos y el proceso (GeeksForGeeks, 2020), ejemplificado con las variables de clima para el proyecto, es el siguiente:

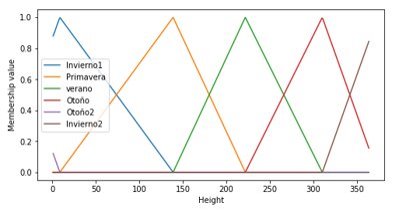
En un principio, se buscó hacer 4 columnas de regresores de primavera, verano, otoño e invierno, según los días en donde se cambia de estación. Se encontró este problema: No es posible que un día sea solo invierno o solo primavera, y es imposible que, de un día a otro, se cambie de verano a otoño. Es un cambio gradual que conlleva diferencias pausadas. La manera en que la lógica difusa funciona visualmente es el observado en la Figura 1:

Figura 1 Variables difusas. Elaboración propia

Como se observa en la Figura 1, lo que se crea son días, del 0 al 355, en donde los días pueden tener .5 de verano y .5 de otoño, o .1 invierno y .9 de primavera. Para comprobar que esto en realidad ayuda al modelo, se buscó predecir la demanda utilizando solamente estas variables difusas, obteniendo el siguiente resultado presentado en la Figura 2:

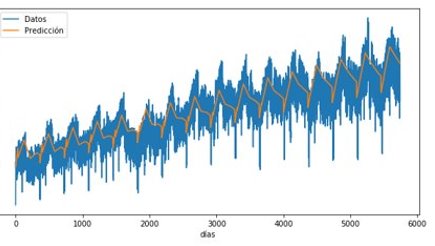


Figura 2 Predicción de la demanda con variables difusas. Elaboración propia

Aunque evidentemente el resultado no es el mejor, se puede observar que las variables difusas pueden encontrar un patrón en la demanda, que se acrecienta en los días calurosos y baja en los días fríos. Por lo tanto, esto funciona y mejora el MAPE del modelo.

Todos los elementos mencionados anteriormente se utilizaron en al producto final, que se encuentra en el apartado 2. Producto.

### 1.5.6 Modelos de variables climáticas

Anteriormente se percibió que el clima era un factor determinante a la hora de predecir la demanda energética, es por ello por lo que se empezaron a usar datos climáticos que contenían las variables:

* Temperatura máxima
* Temperatura mínima
* Temperatura promedio
* Nubosidad
* Lluvia
* Velocidad del viento

Estas variables se emplearon como regresores por si solos, y como datos de entrada para la obtención de los regresores no lineales como se explicó anteriormente. No obstante, el clima es una variable que cambia rápidamente, y las altas frecuencias en vez de ayudar al modelo de predicción, pueden empeorarlo. Es por eso por lo que para eliminar esos efectos de altas frecuencias y suavizar las variables climáticas se decidió elaborar un modelo de predicción para cada una de las variables antes descritas.

Este modelo de predicción se elaboró empleando una regresión lineal y tomando en cuenta las siguientes variables de entrada:

* Variables de lógica difusa. Esto es las variables que representan que estaciones están afectando parcialmente el clima de un día determinado.
* Senos y Cosenos de las frecuencias más importantes de la serie de tiempo de cada una de las variables climáticas.

Una vez terminado este proceso, son estos modelos los que se usan tanto de variables de entrada para la obtención de regresores no lineales, así como el empleo de algunos de ellos como regresores en sí para el modelo predictivo.

Las variables que se usan para el modelo de predicción como regresores son:

* Temperatura mínima
* Temperatura máxima
* Nubosidad

Estas 3 variables se escogieron con base en la experiencia empírica, debido que agregar las demás implicaba un aumento en el error de la predicción.

### 1.5.7 Efecto en el consumo por días contiguos a los festivos

Se sabe que los días en las que cae alguna festividad importante, el consumo es afecto. Este efecto ya se tenía contemplado en los regresores que se usan para el modelo, siendo cada uno de los días festivos una variable binaria.

No obstante, el comportamiento de los datos indica que días anteriores y posteriores a la festividad también se registra un efecto en la cantidad de energía consumida en el día. Es por eso por lo que se optó por modificar las variables binarias de las festividades por una nueva estructura que tome en cuenta el efecto de los días aledaños.

En este caso, el día en el que oficialmente ocurre la festividad será denotado con un valor de .5 en el regresor que represente a este día, mientras que el día anterior y posterior, tendrán un valor de .2, y por último dos días antes y dos días después obtendrán una valoración de .05.

La representación visual del cambio propuesto se puede observar en la Figura 3.

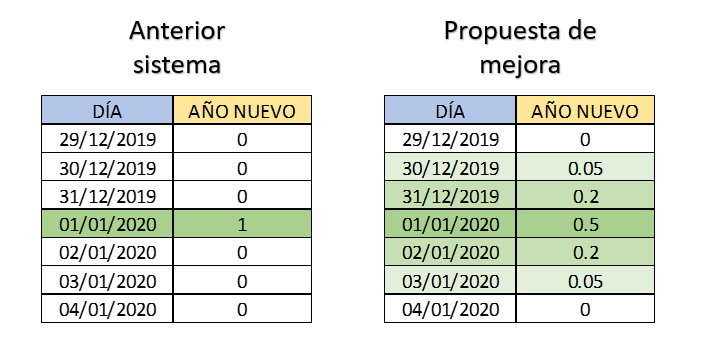


Figura 3 Propuesta de mejora para modelar la afectación por días festivos

### 1.5.8 ARIMA

ARIMA (autorregresivo, integrado, de promedio móvil por sus siglas en inglés) se basa en encontrar los patrones de autocorrelación de los datos.

Este modelo se compone de tres partes, un modelo autorregresivo, uno de diferenciación y otro de media móvil.

#### 1.5.8.1 Modelo Autorregresivo

A este modelo se le conoce como autorregresivo porque se hace una regresión a datos pasados de la misma serie. (1)

|  |  |  |
| --- | --- | --- |
|  |  | () |

Donde y corresponden a los rezagos del dato anterior y el anterior a este del dato que se tiene que predecir, al ruido blanco y los coeficientes de la regresión.

Este modelo se denota como AR(p)

#### 1.5.8.2 Diferenciación

La diferenciación se basa en encontrar el cambio al restar cada dato con el dato o los datos anteriores. Para una sola diferenciación. (2)

|  |  |  |
| --- | --- | --- |
|  |  | () |

El número de diferenciaciones es dependiente de la naturaleza de los datos, se utiliza la prueba de la raíz unitaria con el fin de saber si es necesario este proceso. Se denota como I(d). Ya que el modelo tiene media cero y parece tener forma normal, no se utilizó la diferenciación, dejando el modelo como un ARMA.

#### 1.5.8.3 Modelo de promedio móvil

Este modelo utiliza los errores de las predicciones anteriores a modo de regresión lineal. (3)

|  |  |  |
| --- | --- | --- |
|  |  | () |

Donde es ruido blanco. Se denota como MA(q). (ARIMA models, 2013)

El modelo ARIMA fue aplicado a los residuos de la serie, después de aplicar regresión lineal con Kronecker de línea en las variables dependientes. Los coeficientes de la ARIMA (p, d, q) se encontraron mediante una función iterativa. Igualmente se utilizó ARIMA junto con variables no lineales como variables exógenas, para predecir los datos de la serie reales; sin embargo, para este último es necesario un estudio más profundo.

### 1.5.9 Bosques aleatorios

Los bosques de decisión son comúnmente utilizados como algoritmos de clasificación.

Estos funcionan a base de heurística y en la prueba y error, iterando sobre cada variable.

Comúnmente este tipo de algoritmo no es bueno para predecir a modo de regresión lineal, por lo que se prefiere el uso de bosques aleatorios.

Los bosques aleatorios se componen de muchos bosques, cada uno con sus respectivas ramificaciones, en este caso, cada uno trata de sobre ajustar uno de los parámetros de los datos, con el fin de evitar el sobreajuste, se promedian los factores de cada árbol.

La variabilidad de este modelo se debe a dos partes del algoritmo, la primera es la elección de variables al crear la base del árbol y la segunda es la creación de ramificación al, donde se eligen otras variables aleatoriamente. Para este caso se utilizó XGBRegression de la librería de Python XGBoost. (Müller & Guiado, 2016)

### 1.5.10 Predicción de un año

Para comprobar la eficiencia del modelo, mediante una metodología iterativa se predijeron 354 días, 7 días a la vez.

### 1.5.11 Implementación de algoritmo PCA a variables climáticas

Anteriormente, el modelo usaba la mayoría de las variables climáticas, no obstante, esto puede resultar contraproducente si estos regresores tuvieran una alta correlación entre ellos, como es el caso de “Nubosidad” y “Lluvia”, lo que puede meter ruido al modelo de predicción. Por lo tanto, se propuso la implementación del “análisis de componentes principales”, el cual nos permitiría reducir el número de variables climáticas sin perder demasiada información, resultando en nuevos regresores sin una alta correlación entre ellos.

Para este modelo se decidió que se iban a reducir las variables a un número que explicara por lo menos el 90% de la varianza de la información brindada por el total de las 5 variables climáticas. La idea fue reducir las variables del set de clima, y una vez encontradas estas nuevas componentes, se generó un modelo predictivo para cada variable como se explicó anteriormente.

Es importante recalcar que esta metodología no quedó en el producto final, ya que no logró mejorar el error del modelo antes de PCA.

## 1.6. Valoración de productos, resultados e impactos

### 1.6.1 Valoración de productos

Diariamente, el equipo mandó el pronóstico diario y semanal al CENACE, que, con el paso del desarrollo del proyecto, quedó como se puede observar en la Figura 4.

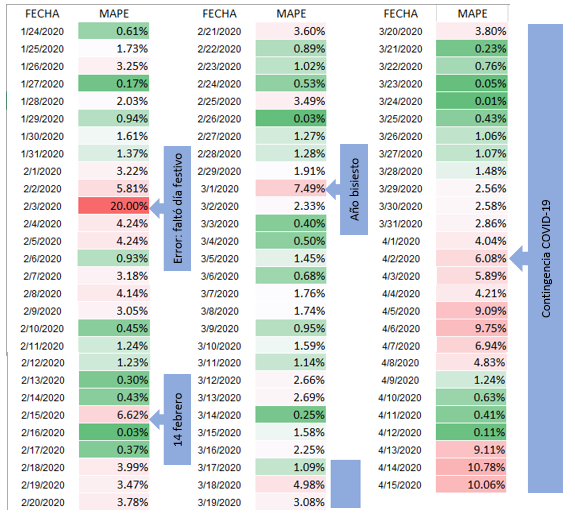


Figura 4 MAPE del pronóstico a 1 día

Como se puede observar, en los últimos días han existido errores debido a la contingencia del COVID-19. Los otros errores marcados se han debido principalmente a días festivos no contemplados, o eventos con poca información como años bisiestos.

Para analizar los objetivos establecidos al inicio del PAP, se realizó la Tabla 1, que explica brevemente cómo se logró cada objetivo:

|  |  |  |
| --- | --- | --- |
| Objetivo | Estatus | Mejoras realizadas |
| Agregar regresores no lineales al modelo. | Logrado | Se logró envolver regresores no lineales como un regresor en un modelo lineal, así se incluyeron estas combinaciones sin perder lo explicativo del modelo lineal. |
| Encontrar la manera de incluir el clima como regresor. | Logrado | Se agregó de dos maneras: Se realizó un modelo de predicción del clima, que resultó mejorar el MAPE más que los datos reales. Así mismo, se implementó lógica difusa para diferenciar las estaciones del año. |
| Trabajar con nuevas librerías para hacer la regresión. | Logrado | Se consideraron nuevas librerías para probar el modelo, sin embargo, aquella que logró el menor MAPE a 1 día para pronóstico continuó siendo Lasso. |
| Mejorar modelo para residuales. | Logrado | Se automatizó la búsqueda de un modelo ARIMA, por lo que, en cada corrida, este decide cuál es el modelo que ayudará mejor al pronóstico. |
| Aplicar mismo modelo a zonas de carga diarias. | Pendiente | - |
| Optimización del modelo. | Logrado | Se logró bajar el tiempo de espera de entre 6-7 minutos, a 2-3 minutos, siendo el proceso de selección de ARIMA el más tardado. |

Tabla 1 Objetivos propuestos para el semestre primavera 2020

Como comentario y recomendación, el siguiente PAP podría enfocarse en los días festivos en particular. A pesar de reducir el error al volverlos variables difusas, para este modelo en particular, se podría enfocar en hacer la predicción para cada día festivo, y calcular qué porcentaje se debería de dar a cada día festivo, pre y post.

### 1.6.2 Resultados

#### Modelo Lasso-ARIMA pre-pandemia

Se encontró que con este modelo prediciendo 52 semanas, se obtiene un MAPE de 2.48 % (Figura 5).

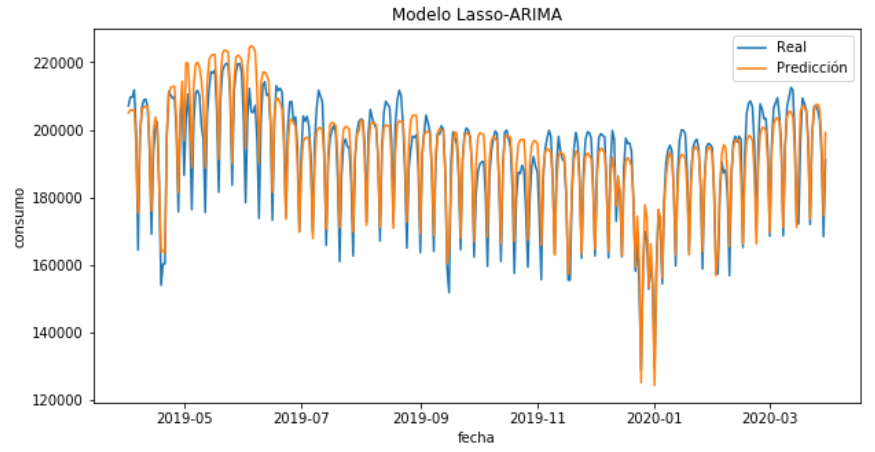


Figura 5 Consumo y predicción con modelo Lasso-ARIMA de 52 semanas.

Se obtienen 78 días con MAPE mayor a 4.0 %. entre estos, los días de la semana que se repiten con mayor frecuencia son domingo y el lunes.

Solo 11 días de este grupo pertenecen a los días especiales registrados, sin embargo, existen

6 días pos-festivos y 5 pre-festivos.

Se obtuvo mayor error el día 5 de junio del 2019 con un MAPE de 9.6 %. (Tabla 2)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **fecha** | **MAPE** | **fecha** | **MAPE** | **fecha** | **MAPE** | **fecha** | **MAPE** |
| 07/04/2019 | 6.66% | 04/06/2019 | 5.7% | 03/10/2019 | 4.39% | 23/12/2019 | 4.34% |
| 14/04/2019 | 4.06% | 05/06/2019 | 9.6% | 04/10/2019 | 4.2% | 24/12/2019 | 5.23% |
| 19/04/2019 | 6.4% | 06/06/2019 | 9.41% | 05/10/2019 | 4.59% | 27/12/2019 | 4.7% |
| 01/05/2019 | 5.56% | 07/06/2019 | 7.75% | 06/10/2019 | 5.46% | 28/12/2019 | 4.01% |
| 02/05/2019 | 7.9% | 08/06/2019 | 7.45% | 20/10/2019 | 6.01% | 01/01/2020 | 4.03% |
| 03/05/2019 | 4.37% | 09/06/2019 | 9.44% | 21/10/2019 | 5.33% | 03/02/2020 | 4.56% |
| 04/05/2019 | 5.09% | 10/06/2019 | 5.76% | 22/10/2019 | 4.24% | 06/02/2020 | 4.43% |
| 05/05/2019 | 7.06% | 16/06/2019 | 4.73% | 23/10/2019 | 5.32% | 09/02/2020 | 5.46% |
| 06/05/2019 | 5.29% | 10/07/2019 | 5.27% | 24/10/2019 | 4.02% | 15/02/2020 | 5.32% |
| 07/05/2019 | 4.02% | 11/07/2019 | 4.47% | 25/10/2019 | 4.78% | 18/02/2020 | 4.08% |
| 09/05/2019 | 4.03% | 21/07/2019 | 6.21% | 26/10/2019 | 6.87% | 19/02/2020 | 4.61% |
| 10/05/2019 | 6.68% | 28/07/2019 | 4.42% | 27/10/2019 | 5.02% | 20/02/2020 | 5.05% |
| 11/05/2019 | 5.4% | 19/08/2019 | 4.12% | 28/10/2019 | 6.33% | 21/02/2020 | 4.77% |
| 12/05/2019 | 7.54% | 21/08/2019 | 4.28% | 31/10/2019 | 4.03% | 22/02/2020 | 4.3% |
| 13/05/2019 | 4.58% | 25/08/2019 | 4.66% | 01/11/2019 | 4.5% | 25/02/2020 | 4.33% |
| 14/05/2019 | 4.05% | 27/08/2019 | 4.42% | 02/11/2019 | 7.4% | 16/03/2020 | 9.36% |
| 19/05/2019 | 5.37% | 16/09/2019 | 9.28% | 03/11/2019 | 6.51% | 17/03/2020 | 4.21% |
| 01/06/2019 | 6.56% | 30/09/2019 | 4.87% | 18/11/2019 | 6.8% | 30/03/2020 | 4.22% |
| 02/06/2019 | 8.97% | 01/10/2019 | 5.35% | 10/12/2019 | 4.02% |  |  |
| 03/06/2019 | 7.46% | 02/10/2019 | 4.83% | 11/12/2019 | 4.67% |  |  |

Tabla 2 MAPE mayor a 4.0 % de pronóstico de 52 semanas a 7 días con modelo de Lasso-ARIMA

#### 1.6.2.2 Modelo bosques aleatorios XGBoost pre-pandemia

Prediciendo 52 semanas con el algoritmo de bosques aleatorios se obtuvo un MAPE de 2.21 % (Figura 6).

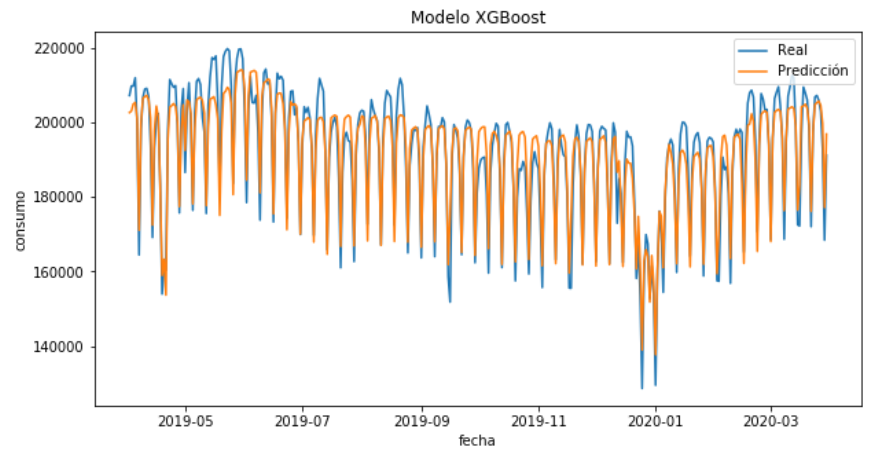


Figura 6 Consumo y predicción de 52 semanas, con modelo de bosques aleatorios

Con este modelo, se obtuvieron 57 días con MAPE mayor a 4%. De este grupo, los días que más se repitieron, fueron lunes y miércoles con 10 días cada uno.

10 días festivos registrados, 2 pre-festivos y 6 festivos entran en esta categoría.

El día con mayor error en la predicción es el 16 de septiembre del 2019, con un MAPE de 15.36 %. (Tabla 3)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **fecha** | **MAPE** | **fecha** | **MAPE** | **fecha** | **MAPE** | **fecha** | **MAPE** |
| 07/04/2019 | 4.03% | 11/07/2019 | 4.16% | 23/10/2019 | 5.43% | 16/01/2020 | 4.15% |
| 21/04/2019 | 4.19% | 20/07/2019 | 4.72% | 24/10/2019 | 4.22% | 17/01/2020 | 5.05% |
| 15/05/2019 | 4.98% | 22/07/2019 | 4.15% | 25/10/2019 | 4.23% | 18/01/2020 | 4.53% |
| 16/05/2019 | 4.61% | 27/07/2019 | 4.02% | 26/10/2019 | 5.28% | 03/02/2020 | 8.66% |
| 17/05/2019 | 5.96% | 21/08/2019 | 4.63% | 28/10/2019 | 5.71% | 04/02/2020 | 5.41% |
| 20/05/2019 | 5.68% | 16/09/2019 | 15.36% | 02/11/2019 | 4.58% | 06/02/2020 | 4.9% |
| 21/05/2019 | 4.55% | 30/09/2019 | 4.72% | 18/11/2019 | 12.05% | 09/02/2020 | 4.23% |
| 22/05/2019 | 4.91% | 01/10/2019 | 5.18% | 19/11/2019 | 4.46% | 15/02/2020 | 6.87% |
| 23/05/2019 | 4.72% | 02/10/2019 | 4.42% | 12/12/2019 | 7.91% | 19/02/2020 | 4.02% |
| 24/05/2019 | 4.75% | 03/10/2019 | 4.34% | 20/12/2019 | 4.51% | 08/03/2020 | 4.57% |
| 05/06/2019 | 4.08% | 04/10/2019 | 4.24% | 23/12/2019 | 4.53% | 16/03/2020 | 6.37% |
| 06/06/2019 | 4.24% | 05/10/2019 | 6.07% | 24/12/2019 | 5.63% | 29/03/2020 | 5.2% |
| 08/06/2019 | 4.58% | 06/10/2019 | 4.11% | 25/12/2019 | 8.01% |  |  |
| 09/06/2019 | 4.22% | 21/10/2019 | 5.06% | 01/01/2020 | 6.31% |  |  |
| 10/07/2019 | 4.97% | 22/10/2019 | 4.83% | 05/01/2020 | 4.28% |  |  |

Tabla 3 MAPE mayor a 4.0 % de pronóstico de 52 semanas a 7 días con modelo de bosques aleatorios

#### 1.6.2.3 Comparación de modelos

Con el modelo de bosques aleatorios se obtuvieron errores más grandes que con el modelo de Lasso y ARIMA; sin embargo, la cantidad de días, con error sobre 4.0 % es mayor para este último que para el modelo de bosques aleatorios.

34 valores coinciden con predicción arriba de 4.0 %, de estos datos, 11 datos pertenecen a días especiales, pos-festivos y pre-festivos. (Tabla 4)

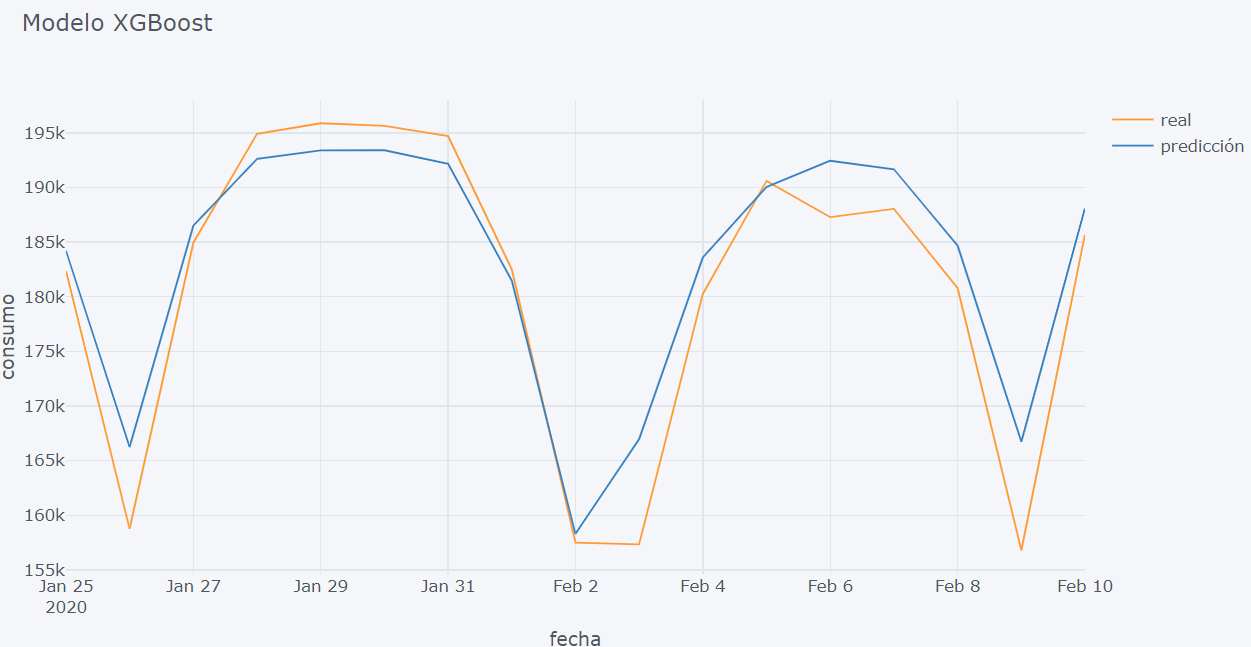
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **fecha** | **MAPE RF** | **MAPE LA** | **fecha** | **MAPE RF** | **MAPE LA** |
| 07/04/2019 | 4.03% | 6.66% | 22/10/2019 | 4.83% | 4.24% |
| 05/06/2019 | 4.08% | 9.6% | 23/10/2019 | 5.43% | 5.32% |
| 06/06/2019 | 4.24% | 9.41% | 24/10/2019 | 4.22% | 4.02% |
| 08/06/2019 | 4.58% | 7.45% | 25/10/2019 | 4.23% | 4.78% |
| 09/06/2019 | 4.22% | 9.44% | 26/10/2019 | 5.28% | 6.87% |
| 10/07/2019 | 4.97% | 5.27% | 28/10/2019 | 5.71% | 6.33% |
| 11/07/2019 | 4.16% | 4.47% | 02/11/2019 | 4.58% | 7.4% |
| 21/08/2019 | 4.63% | 4.28% | 18/11/2019 | 12.05% | 6.8% |
| 16/09/2019 | 15.36% | 9.28% | 23/12/2019 | 4.53% | 4.34% |
| 30/09/2019 | 4.72% | 4.87% | 24/12/2019 | 5.63% | 5.23% |
| 01/10/2019 | 5.18% | 5.35% | 01/01/2020 | 6.31% | 4.03% |
| 02/10/2019 | 4.42% | 4.83% | 03/02/2020 | 8.66% | 4.56% |
| 03/10/2019 | 4.34% | 4.39% | 06/02/2020 | 4.9% | 4.43% |
| 04/10/2019 | 4.24% | 4.2% | 09/02/2020 | 4.23% | 5.46% |
| 05/10/2019 | 6.07% | 4.59% | 15/02/2020 | 6.87% | 5.32% |
| 06/10/2019 | 4.11% | 5.46% | 19/02/2020 | 4.02% | 4.61% |
| 21/10/2019 | 5.06% | 5.33% | 16/03/2020 | 6.37% | 9.36% |

Tabla 4 Comparación de MAPEs mayores a 4.0 % obtenidos de predicción con modelo Lasso-ARIMA (LA)

Se muestra en la Figura 7 la predicción del día especial 3 de febrero, para el cual, en ambos modelos el error fue grande. Se encontró que los días especiales celebrados entre semana, aun siendo agregados como regresores para el modelo, no se predicen correctamente.



a)



b)

Figura 7 Datos reales y predicciones de modelo con a) Lasso y ARIMA y b) bosques aleatorios del 25 de enero al 10 de febrero del 2020

#### 1.6.2.4 Modelo en pandemia

Los efectos de la pandemia de COVID-19 son claros, como se puede ver en la Figura 8 hay una disminución en la demanda que se hace presente los primeros días del mes de abril, por lo que el error de nuestro modelo aumenta considerablemente.



Figura 8 Comparación de las predicciones hechas con nuestro modelo contra el consumo real afectado por la pandemia

Para poder sobrellevar esta afectación se propusieron 3 metodologías:

* Multiplicar nuestra predicción por un algún valor menor a 1 (por ejemplo .85) con la finalidad de ajustar nuestros valores con los datos reales.
* Observando la proporción entre lo real y lo predicho, escalar todos nuestros datos de entrenamiento de antes de la pandemia por esta proporción, así el resultado de la predicción estaría considerando este decremento.

En primer lugar, el primer punto fue descartado debido a que la razón de proporción entre lo real y la predicción no es una constante, lo que implica que al implementar esta estrategia estaríamos disminuyendo el error, pero no de una manera precisa y eficiente.

En cuanto al segundo punto, se calcularon todas las proporciones de la división real/predicción de los datos disponibles y se encontró un promedio de aproximadamente .91, el cuál fue el factor de escalamiento para los datos de entrenamiento anteriores al 30 de marzo del 2020.

Sin embargo, podemos observar en la Figura 9 y en la Tabla 5 que a pesar de que esto efectivamente ayudó a la disminución del error, no fue suficiente comparando con los resultados mostrados en la Figura 4.

Es por eso por lo que se decidió dejar el modelo como estaba, ya que empleando estas metodologías se podían disminuir en lo general el error, no obstante, sacrificando los MAPEs de hasta .11% obtenidos con el modelo sin modificación alguna.

Algunas desventajas de esta metodología fueron:

* El factor de escalamiento no era constante en todos los días, nosotros usamos el promedio, pero porque ya sabíamos de antemano todos los datos, en la predicción día con día no sería tan eficiente.
* Los errores porcentuales disminuyeron dejando MAPEs entre 2 y 6 para la mayoría de los días, siendo que con el modelo solo había días en que se obtenían MAPEs menores a 1%.



Figura 9 Resultado de la predicción empleando la metodología de escalar los datos de entrenamiento

|  |  |
| --- | --- |
| DÍA | MAPE % |
| 08/04/2020 | .5327 |
| 09/04/2020 | 4.9774 |
| 10/04/2020 | 4.13455 |
| 11/04/2020 | 3.381431 |
| 12/04/2020 | 1.5785 |
| 13/04/2020 | 4.0479 |
| 14/04/2020 | 6.1359 |

Tabla 5 MAPEs obtenidos empleando la metodología de escalamiento a los datos de entrenamiento

#### 1.6.2.5 PCA

El proceso de PCA, conservando el 90% de la varianza, dio como resultado 2 componentes principales. Mas adelante se generaron modelos predictivos de estas nuevas variables para al final incorporarse al modelo de predicción de la demanda como regresores.

Se usó un sistema iterativo para probar los beneficios de esta metodología; esto consistió en correr iterativamente semana por semana los dos modelos, con y sin PCA, por 364 días. Esto quiere decir que los modelos se corrieron 52 veces desde el 1 de abril del 2019 hasta el 30 de marzo del 2020.

La comparativa entre ambos modelos se puede observar en la Tabla 6.

Tabla 6 Comparación del modelo con y sin la implementación del PCA

|  |  |  |
| --- | --- | --- |
|  | MODELO SIN PCA | MODELO CON PCA |
| MAPE promedio de los 364 días | 2.4787 | 2.5844 |
| MAPE máximo | 9.5971 | 9.9352 |
| Días en que la predicción fue mejor | 212 | 152 |
| Promedio de la disminución del error en los días en que el modelo fue el mejor | 0.46899 | 0.401030 |

Aunque hubo 152 días en los que el modelo con PCA mejoró al que no lo implementaba, se puede notar que el error aumenta los otros 212 días es mayor, produciendo a su vez un promedio de MAPE más alto.

Por lo tanto, a la luz de este análisis, se decidió no incorporarlo a la entrega final al no brindar más beneficios que lo que perjudicaba al modelo. Sin embargo, se podrían buscar otras soluciones si se quisiera seguir por esta línea de investigación:

* Usar menos varianza, un ejemplo el 85%.
* Incorporar el PCA en otros momentos del modelo predictivo, por ejemplo, antes de la regresión.

## 1.7. Bibliografía y otros recursos

Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE, Joaquín Amat Rodrigo. Disponible en: [https://www.cienciadedatos.net/documentos/35\_principal\_component\_analysis](https://www.cienciadedatos.net/documentos/35_principal_component_analysis9) [09 de mayo del 2020]

ARIMA models. (2013). In H. R, & A.G, Forecasting: Principles and Practice (pp.67-73). Melbourne: OTexts.

Berrière, Lali (2011). Serie de Fourier en “Universitat Politècnica de Catalunya ”. Disponible en: <https://mat-web.upc.edu/people/lali.barriere/as/serie-fourier.pdf> [31 de octubre 2019]

Camacho, Máximo (s.f.) Variables Ficticias en “Universidad de Murcia”. Disponible en: <https://www.um.es/econometria/adeydcho/transpar/07-ficticias_Maximo.pdf> [28 de octubre 2019]

Fuzzy Logic | Introduction en “GeeksForGeeks” (2020). Disponible en: <https://www.geeksforgeeks.org/fuzzy-logic-introduction/> [23 de abril 2020]

Quiénes Somos en “CENACE.GOB” 2019. Disponible en: <https://www.cenace.gob.mx/CENACE.aspx> [26 de octubre 2019]

Müller, A. C., & Guido, S. (2016). Ensamble of Decision Trees. In introduction to Machine Learning with Python (A guide for data scientist) (pp. 85 - 90). California: O'Reilly.

## Productos

Figura 10 Diagrama de flujo del algoritmo utilizado

Explicando la Figura 10, el programa realiza los siguientes pasos:

1. Importar datos: Se importan 4 documentos diferentes. La demanda, Los datos de clima, los días festivos que se ven afectados en días pre y post, y los festivos que solo se afecta el día en cuestión.
2. Se crea un modelo de las variables climáticas nubosidad, temperatura mínima y máxima. Para lo anterior, se genera una regresión teniendo como parámetros de entrada las estaciones en lógica difusa, y senos-cosenos como se especifica en el teorema de Fourier.
3. Se genera la primera regresión de la demanda con los siguientes parámetros de entrada: días festivos como dummies (que incluye los afectados en días pre y posteriores y los no afectados)
4. Se obtienen los residuales de la regresión 1, y se hace un periodograma con los mismos, para encontrar los efectos que las dummies no alcanzaron a detectar.
5. Se hace la segunda regresión con los siguientes parámetros de entrada:
   1. Producto kronecker por línea de dos tablas:
      1. Tabla 1: Estaciones en lógica difusa, días festivos como dummies (que incluye los afectados en días pre y posteriores y los no afectados)
      2. Tabla 2: Senos y cosenos de los periodos encontrados en el periodograma, variables no lineales del clima
   2. El modelo de clima del paso 2.
6. Se obtienen los residuales de la segunda regresión.
7. Sobre estos residuales, se encuentra el mejor modelo ARIMA que los ajuste, siendo este un ARMA, indicando que el modelo ya está diferenciado.
8. Se hace una tercera regresión con Lasso, para reducir la cantidad de variables presentes en el modelo.
9. Con esta tercera regresión, se hace la predicción de la demanda, a la cual se le agrega la predicción del ARMA propuesto en cada corrida.

Para realizar todo el proceso, el programa contiene las siguientes funciones que son llamadas recursivamente entre ellas para dar lugar al modelo de predicción:

* Mape(y,y\_pred). Esta función toma de entrada los valores reales de demanda y los valores arrojados por el proceso de predicción, y devuelve el error porcentual absoluto medio.
* kronecker(data1, data2). Necesita 2 argumentos: data1 y data2 los cuales tiene que ser DataFrames de pandas. Esta función devuelve un DataFrame obtenido haciendo el producto kronecker por fila entre data1 y data2.
* triangular(). Esta función crea las variables climáticas difusas y no requiere de argumento alguno y regresa un DataFrame con estos valores.
* pba(j,n). Toma como argumentos de entrada el nombre de la variable climática a modelar y el número de días que se quieren predecir usando ese modelo. Regresa un DataFrame con los valores modelados de dicha variable climática.
* modelosclima(n). De argumento lleva solo el número de días a predecir con los modelos climáticos. Esta función hace el proceso iterativo de llamar a la función pba para cada variable climática.
* festivos(). No requiere de argumento y regresa un DataFrame con los regresores correspondientes a los días festivos, considerando el efecto de los días contiguos como se explicó anteriormente.
* dummies(n). Toma de argumento el número de días a predecir. Concatena la información de los variables de tiempo (día, mes) con los regresores no lineales climáticos que no requieren de normalización para posteriormente normalizar todos los datos. Una vez hecho esto se agregan los regresores no lineales que requieren de inputs normalizados y se generan dummies de las variables “mes” y “año”. Regresa un DataFrame con todos los regresores estipulados anteriormente.
* separar(n, X, y). Esta función requiere de 3 argumentos, el número de día a predecir, los regresores para la regresión (X), y los datos del consumo. La función separa los datos en sets de entrenamiento y de prueba. Regresa dos tuplas, siendo la primera la información de la variable X (X\_train, X\_test) y la segunda la información de la variable y (y\_train,y\_test).
* reg(n, pre\_forc). La función que genera la primera regresión. Toma de argumento el número de días a predecir e información acerca de si quiere aplicar el modelo entrenado a los mismos datos de entrenamiento (general) o si se desea probar el modelo entrenado en los datos de prueba (forecasting). Regresa un DataFrame con dos columnas, que representan los valores reales y los obtenidos con el modelo.
* visual1(n). Toma el número de días a predecir y muestra 3 gráficas: El comportamiento de los datos del consumo real contra los datos de la predicción sobre el set de entrenamiento, los residuales que se genera de la resta de estos dos sets, y la comparación entre el set de prueba contra los días predichos por el modelo.
* perio(n, fg). Los argumentos son: el número de días a predecir y si se quiere usar el modo “general” o “forecasting” de reg. Calcula el periodograma de los residuos obtenidos de la regresión y regresa las frecuencias encontradas.
* kron(n). Toma de argumento el número de días a predecir, y genera una matriz de senos y cosenos con las frecuencias obtenidas de perio(n,fg). Regresa el producto kronecker entre los regresores X, y la matriz de senos y cosenos.
* reg2(n, pre\_forc). Hace los mismo que reg1 solo que en este caso se manda a llamar internamente a la función kron(n), para entrenar el modelo con los datos generados por el producto kronecker.
* visual2(2). Análogo de visual1(n) pero para la visualización de las gráficas que se generan por la regresión entrada con el producto kronecker.
* aroma(n). Toma los residuos generados de la diferencia entre el consumo real y el comportamiento de reg2 sobre los datos de entrenamiento y genera un modelo ARIMA. Esto se hace con la finalidad de predecir dichas diferencias. Regresa la predicción de las diferencias para los días que se requiere predecir el consumo.
* reg3(n). Misma funcionalidad que reg2(n), solo que se agrega al resultado de la predicción los valores determinados por aroma(n), y se emplea una regresión con regularización tipo Lasso.

## 3. Reflexión crítica y ética de la experiencia

El RPAP tiene también como propósito documentar la reflexión sobre los aprendizajes en sus múltiples dimensiones, las implicaciones éticas y los aportes sociales del proyecto para compartir una comprensión crítica y amplia de las problemáticas en las que se intervino.

## 3.1 Sensibilización ante las realidades

Durante este proyecto, gracias al apoyo del CENACE, se pudo experimentar de primera mano los problemas a los que día con día se enfrentan. Primero que nada, como usuarios de energía, no se sabe todo el trabajo que hay detrás de prender la luz en la casa o en la universidad. Se da por sentado para millones de usuarios que, al pagar la luz, se tiene luz, y no se entiende que muchas personas trabajan en conjunto para hacer esto posible. Se conoce de la CFE, pero es raro saber que hay un organismo que le dicta a la CFE cuánta demanda sí y cuánta no.

Un factor muy interesante que nos abre los ojos es las pérdidas que tiene el CENACE cuando hay malos pronósticos. Esta empresa tiene contratos con otras que se encargan de proveer la energía en las cantidades que el CENACE da, por lo que ellos deben ser capaces de dar la mejor respuesta posible acerca de cuánta energía se usará en el Occidente del país. Lo más importante es que nunca se llegará a: “Se hizo un mal pronóstico, no habrá energía”. Al contrario, en caso de que se diera un pronóstico menor a la demanda, los proveedores cobran más caro, ya que es necesario prender maquinarias de manera apresurada para evitar que colonias y personas se queden sin luz. Son millones de pesos que el CENACE debe perder, pero luz sí habrá.

Aunado a lo anterior, el CENACE informó a el equipo las problemáticas que fueron surgiendo en este semestre. No fue una vez si no dos en donde México se enfrentó a situaciones que no estaban previamente contempladas: El paro del día de la mujer y la contingencia del COVID-19. Hablando del paro, los encargados del CENACE tuvieron que hacer investigaciones previas de cómo creían que iba a impactar en cada zona el paro, ya que en algunas partes del país se veía más fuerte que en otras. Aun así, zonas de México que se esperaban grandes cambios como en la zona centro del país, transcurrieron normalmente mientras en el sur, cuyo impacto no creían que fuera significante, tuvo bajas en la demanda. Ante esta situación, el equipo decidió que en el Occidente era mejor dejar el día sin tomarlo como festivo, ya que nuestro modelo detectaría una baja por la marcha del día anterior y se ajustaría a ello.

Por el otro lado, ante la contingencia, es muy poco lo que como equipo se pudo hacer. No hay información previa más que de la influenza, y al ser problemas diferentes, se esperan resultados diferentes. Sí hubo baja en la demanda, y sorprende ver la ardua labor que el CENACE tiene que llevar a cabo para pronosticar de la mejor manera posible, sin parar como muchas empresas deben hacerlo, ya que es de vital importancia la luz para llevar a cabo de la mejor manera posible las actividades de las y los mexicanos.

## 3.2 Aprendizajes logrados

### 3.2.1 Diana

Profesionalmente, el PAP me enseñó algunas herramientas con las que no estaba familiarizada como fue la idea de meter regresores no lineales en una regresión linear, aprendí cómo hacer un modelo con variables difusas y a incorporarlo al modelo que ya se tenía, aprendí un poco del concepto de otro tipo de regresiones como CUR, PCA, árboles de decisión, Regresión con Ridge, que aunque al final no se usaron para el modelo en sí, fueron pruebas que se realizaron para buscar alternativas. Este proyecto me ayudó a reafirmar qué es en lo que me gustaría especializarme profesionalmente.

Dentro del ámbito social, aprendí los diferentes retos a los que una empresa se puede encontrar al enfrentarse a situaciones que, como se conoce comúnmente en finanzas, son riesgos del mercado, es decir, afectan a todos sin importar el sector en el que se encuentren, y no hay nada que hacer para evitarlas. La manera de sobrellevarlas es intentar predecir, pero en situaciones así, al menos para nuestro equipo, es básicamente suerte si se pronosticó correctamente.

Finalmente, personalmente aprendí a trabajar de manera eficiente con un equipo, creo que al principio costó unir todas las ideas que cada uno tenía, al final todas eran buenas aportaciones y era necesario buscar una manera de unirlos. Fui capaz de organizar mis tiempos ya que, en la contingencia, a pesar de estar en casa, fue necesario establecer horarios para trabajar en el PAP.

### 3.2.2 Alan

Aunque ya había tenido experiencia en predicciones y análisis de datos por las materias complementarias que he inscrito, esta fue mi primer acercamiento profesional al campo de la ciencia de datos, lo que me proveyó de herramientas muy valiosas que pueden servirme en otros ámbitos como el análisis de datos experimentales en mi vida como investigador.

Simplemente considerando los aprendizajes nuevos en cuanto a la regresión simbólica, regresiones regularizadas como Lasso y Ridge, modelos autorregresivos como el ARIMA y el uso de árboles de decisiones (aunque este último fue descartado en nuestro modelo), quedo satisfecho con lo aprendido en el proyecto.

En cuanto a otro tipo de aprendizajes, también trabajar con mi equipo me ayudo a fortalecer mis habilidades interpersonales, así como el mejoramiento de mi organización al dividirnos el trabajo y avanzar de una manera independiente para abonar al mismo objetivo.

Además, hacer conciencia en cuanto al impacto que tiene nuestro trabajo en la vida en sociedad, y no solo hacer por hacer como comúnmente se hace en clase, por lo que consideró esencial la implementación de prácticas como lo hace el ITESO.

### 3.2.3 Gregorio

Durante este semestre aprendí la matemática de la mayoría de los algoritmos que se utilizaron, como regresión lineal Lasso, con lo que aprendí a optimizar los valores de entrada y los hiper parámetros de dichos algoritmos.

Aprendí a implementar nuevos algoritmos y nuevas metodologías, como regresión simbólica y análisis espectral de potencias, lo que, a su vez, me ayudo a conocer otros algoritmos de aprendizaje de máquinas que no se utilizaron en este proyecto como redes neuronales y PCA, también aprendí las bases de ciencia de datos.

Aprendí a montar experimentos mediante el código de programación y mejoré mis habilidades para programar en el lenguaje Python. Del mismo modo, reafirmé mis conocimientos sobre planeación de proyectos además de varios conocimientos que obtuve durante la carrea, como probabilidad y estadística y algebra lineal.

Finalmente, durante ese proyecto pude mejorar mis habilidades para trabajar en equipo, al estar consciente de las habilidades y necesidades de cada uno de los participantes, así como obtuve nociones de liderazgo, al observar la labor de nuestros tutores y sobre tener un horario de trabajo, al reunirnos en un horario especifico.