Manual de Usuario librería DANNE.EMMET

2024-09-11

Introducción

Esta librería realiza las tareas de procesamiento mensual de la Encuesta Mensual Manufacturera con Enfoque Territorial EMMET, la cual fue construida en lenguaje de programación R. desarrollando scripts que ayudena la optimización, flujo y entendimiento de los procesos.

Esta librería es diseñada a partir de la caracterización de la operación estadística donde se evidencia el uso de múltiples herramientas analíticas, como: Python, Excel, macros de Excel, visores, SAS y Word. En el siguiente gráfico se muestra el uso de herramientas para cada una de las tareas definidas en el proceso estadístico.

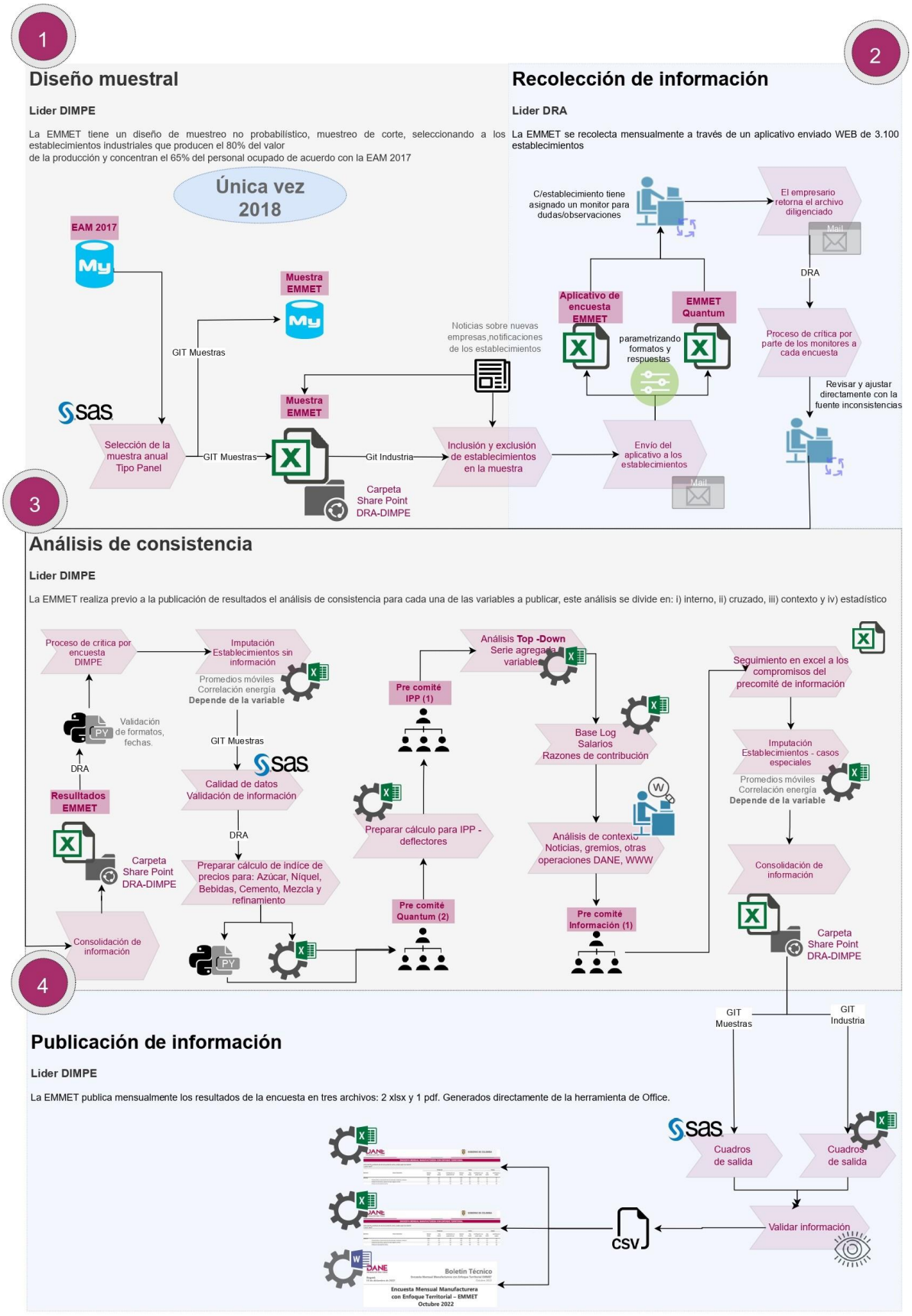


Figure 1: Flujo_libreria

Para la creación de este piloto se hizo una recopilación de bases de datos de interés, junto con un diccionario de las variables que contienen cada base. Apoyados en el diccionario, se construyó la guía para el tratamiento de estas bases, su uso e importancia. Con esto, se definió qué variables se repetían en las bases, cuáles son las llaves para enlazar las diferentes bases y cuáles de éstas permanecían por la importancia en el proceso y cuales se descartaban.

Revisando los procesos de la EMMET, se concluyó que varias tareas se podían automatizar y estandarizar en un único software libre como R o Python y de acuerdo con las capacidades del equipo técnico se priorizoel uso de R. Con base en esto y teniendo en cuenta la caracterización ya mencionada, se definió la ruta a seguir; en una primera instancia esto implicaba realizar el acople entre la base denominada "Logística"; esta base recopila el histórico de datos hasta el mes anterior a tratar, con la base original, es decir la base del mes de interés con los datos que entregaron las fuentes, sin embargo, debido al uso del aplicativo por parte de DRA, ahora se recibe una base panel ya con la integración hasta el mes de interés.

Usando todos los recursos disponibles, se transcribió un código que estandarizaba las variables numéricas y de tipo carácter. Este código estaba realizado en SAS, y se transcribió a lenguaje de R, para implementarlo dentro de la ruta que se estableció. Adicionalmente, se realiza un ajuste de fechas para las fuentes que no rindieron esta información, colocando el periodo en el que normalmente reportan.

Después de caracterizada la operación se define la construcción de las siguientes funciones:

- Integración de bases: Se toma la base panel, que contiene toda la información necesaria para ejecutar los diferentes procesos y se integra con la base paramétrica, adicionalmente se estandarizan las variables numéricas cambiando los datos faltantes o NA por 0.
- Detección de alertas: Analiza los datos entregados en el mes de interés y detecta comportamientos anómalos en comparación a la serie histórica de cada empresa.
- Imputación: De acuerdo con el comportamiento de cada empresa, identifica mejor valor por el cual se reemplaza los establecimientos identificados anteriormente como alertas.
- Creación base Temática: Realiza la creación de las nuevas variables, a partir de las encontradas en la base origina, que contiene la base temática.
- Anexo nacional: Crea los cuadros de salida del anexo nacional a partir de la base temática.
- Anexo territorial: Crea los cuadros de salida del anexo territorial a partir de la base temática.
- Cuadros de dominios: Crea cuadros de salida en donde se desagrega la información de producción, ventas y empleados por dominios.
- Cuadros de regiones: Crea cuadros de salida en donde se desagrega la información de producción, ventas y empleados por áreas metropolitanas, ciudades.
- Boletín: Crea el boletín de publicación, a partir de la base temática.

Adicionalmente se crearon dos productos, el primero es un tablero de control en Shiny que permite a los usuarios interactuar con los datos de manera fluida y dinámica, brindando una visión clara y en tiempo real de los resultados de la encuesta. El segundo producto es un tablero en power BI el cual resumen la información del boletín y la presenta de una manera más interactiva para el usuario.

Esta librería representa un avance significativo en la automatización y optimización de la Encuesta Mensual Manufacturera. Está diseñada para ahorrar tiempo, reducir errores y brindar una mayor comprensión de los procesos involucrados.

Caracterización y uso de las funciones:

Las funciones de la librería desarrollan partes claves que actúan como engranajes para el desarrollo de esta. A continuación, explicamos en detalle cada una de las funciones mencionadas anteriormente.

Las funciones y características clave de esta librería incluyen:

- **Función Inicial (f0_inicial)**

  - Descripción: Esta función instala y carga todos los paquetes necesarios para el correcto funcionamiento de la Liberia. Además, busca en el directorio suministrado si existen las carpetas en donde se almacenarán los archivos de salida de cada una de las siguientes funciones, en caso de que no existan, las crea. Finalmente, la función inicial crea un archivo llamado parámetros_boletin.xlsx en la carpeta de s6_boletin, el cual contiene dos hojas, la primera contiene los valores de los parámetros "IC_prod", "IC_ven", "IC_empl", "TNR", "TI_prod", "TI_ven", "TI_empl", "Anio_grafico"; los cuales son necesarios para la construcción del boletín. La segunda hoja contiene los nombres de las variables que se necesitan para el funcionamiento de la librería del archivo mandado por DRA.
  - Ejemplo: f0_inicial(directorio="Documents/DANE/Procesos DIMPE /PilotoEMMET", mes=9,año=2024)
  - Entrada: Nada.
  - Salida: archivo xlsx, “parámetros_boletin.xlsx”

Guía para actualizar archivo de alertas: Luego de ejecutar esta función se creara un archivo de Excel tipo csv, este contiene variables de identificación de los establecimientos, los valores en cada una de las variables de interés (capitulo 2 y capitulo 3), y si son posibles casos de imputación, las variables en las que tengan un valor diferente a continua son los que en la función de imputación pasaran por ese proceso

Para modificar el archivo debe modificar el valor en la variable que desee y busque la columna cuyo nombre es nombrevariable_caso_de_imputacion y modifique el valor por “continua_critica”, para tener registro deque valores se modificaron, si desea, en la última columna puede realizar un comentario

Ejemplo, queremos modificar en valor de la producción del establecimiento con ID_numord 23; por lo tanto primero se busca la fila cuyo valor de id_numord es 23, en la columna “AJU_III_PE_PRODUCCION” cambiaremos el valor numérico por el valor que deseamos (evite usar decimales), luego proceda a buscar la columna “AJU_III_PE_PRODUCCION_caso_de_imputacion”, ahí modifique el valor de la casilla por “continua_critica”, sin importar si el valor anterior era “continua”, “imputacion_deuda” o “imputacion_caso_especial”.

- **Función Integración (f1_integracion)**

  - Descripción: Esta función permite la consolidación y unificación de dos fuentes de datos en una sola, la base logística, que se recibe por parte de DRA y la base paramétrica histórica, para que, a partir de esta ejecuten los demás procesos, adicionalmente en los campos numéricos que posean espacios nulos, se reemplazaran por 0, estandarizando todas las variables numéricas de la base.
  - Ejemplo:f1_integracion(directorio="Documents/DANE/Procesos DIMPE/PilotoEMMET", mes=9,año=2024)
  - Entrada: EMMET_parametrica_historico.xlsx y EMMET_PANEL_imputada_mes_anio.xlsx (base logística)
  - Salida: archivo csv, “EMMET_PANEL_trabajo_original_mes_anio.csv”

- **Función Identificación de alertas (f2_identificacion_alertas)**

- Descripción: La función de alertas se encarga de analizar los comportamientos de las empresas registradas en la encuesta, con el fin de alertar cambios inusuales o bruscos en su patrón de respuesta, marcando las alertas de la siguiente manera: Si la novedad de la empresa es 5 se marcará como imputación_deuda y si el comportamiento es inusual o anómalo se marcará como imputación_caso_especial.
- Ejemplo: f2_identificacion_alertas(directorio="Documents/DANE/Procesos DIMPE/PilotoEMMET", mes=9,año=2024,avance=100)
- Entrada: "EMMET_PANEL_trabajo_original_mes_anio.csv"
- Salida: archivo csv, "EMMET_PANEL_alertas_mes_anio.csv"

• Función Imputación (f3_imputacion)

- Descripción: La función de imputación reconoce las casillas marcadas en la función de alertas y separa las variables en capítulo 2 y capitulo 3, para las variables de capítulo 2 se imputa el valor con el valor del mes anterior, mientras que para las de capítulo 3 realiza una metodología de imputación en la cual analiza el histórico del establecimiento y a partir de eso da un valor. Con esto se puede concluir que la librería brinda capacidades de imputación automáticas para completar o modificar la información de manera confiable. La función crea dos archivos uno en donde estan los valores ya imputados para el mes de interés, el otro archivo contiene las reglas de consistencias para las diferentes variables, con el fin de identificar que los valores imputados mantengan una consistencia.
- Ejemplo: f3_imputacion (directorio="Documents/DANE/Procesos DIMPE/PilotoEMMET", mes=9,año=2024,avance=100)
- Entrada: "EMMET_PANEL_trabajo_original_mes_anio.csv" y "EMMET_PANEL_alertas_mes_anio.csv"
- Salida: archivos csv, "EMMET_PANEL_imputada_mes_anio.csv" y "EMMET_reglas_consistencia_mes_anio.csv"

• Función Base temática (f4_tematica)

- Descripción: La librería facilita la generación de anexos nacionales y territoriales de manera automatizada, simplificando la generación de informes detallados. Estos anexos o cuadros de salida muestran información complementaria a la registrada en el boletín de prensa con el fin de brindar la información a un nivel más desagregado tanto total nacional como desagregado a nivel de departamentos, áreas metropolitanas y principales ciudades del país.
- Ejemplo: f5_anacional (directorio="Documents/DANE/Procesos DIMPE/PilotoEMMET", mes=9,año=2024)
- Entrada: "EMMET_PANEL_imputada_mes_anio.csv"
- Salida: archivo csv, "EMMET_PANEL_tematica_mes_anio.csv"

• Funciones anexos f5_anacional y f6_aterritorial)

- Descripción: Esta función genera la base temática. Esta contiene la información de los datos procesados, de acuerdo con la metodología de la operación, en este sentido se presentan los datos reales a partir de los nominales y a su vez la información ponderada y agrega variables de identificación de los dominios por los cuales se publica.
- Ejemplo: f4_imputacion (directorio="Documents/DANE/Procesos DIMPE/PilotoEMMET", mes=9,año=2024)
- Entrada: "anexos_nacional_emmet_formato.xlsx", "EMMET_PANEL_tematica_mes_anio.csv" y "anexos_territorial_emmet_formato.xlsx"
- Salida: archivo csv, "anexos_nacional_emmet_mes_anio.xlsx" y

"anexos_territorial_emmet_mes_anio.xlsx"

- **Función cuadro dominios (f7_cdominios)**

  - Descripción: La librería crea un archivo de excel en donde se genera una hoja por cada dominio que muestra la variación y contribución anual de las variables de producción, ventas, empleados, sueldos y horas.
  - Ejemplo: Cuadros_Dominios (directorio="Documents/DANE/Procesos DIMPE/PilotoEMMET", mes=9,año=2024)
  - Entrada: "EMMET_PANEL_tematica_mes_anio.csv" y "DEFLACTOR_mesanio.xlsx"
  - Salida: archivo csv, "cuadros_nacionales_mes_anio.xlsx"

- **Función cuadro regiones (f8_cregiones)**

  - Descripción: La librería crea un archivo de excel en donde se genera una hoja por cada desagregación de región que muestra la variación y contribución anual de las variables de producción,ventas, empleados, sueldos y horas.
  - Ejemplo: Cuadros_regiones (directorio="Documents/DANE/Procesos DIMPE/PilotoEMMET", mes=9,año=2024)
  - Entrada: "EMMET_PANEL_tematica_mes_anio.csv" y "DEFLACTOR_mesanio.xlsx"
  - Salida: archivo csv, "cuadros_territoriales_mes_anio.xlsx"

- **Función boletín (f9_boletin)**

  - Descripción: Esta es la etapa final de la Encuesta Mensual Manufacturera, la creación del boletín. Esta función genera un archivo pdf o Word, según se especifique, en donde el contenido de este es un informe que genera un resumen ejecutivo en el que se presentan los principales resultados con el uso de diferentes tipos de herramientas visuales para mostrar información histórica y lograr hacer comparaciones de variaciones entre las distintas actividades o desagregación regional.
  - Ejemplo: f9_boletin (directorio="Documents/DANE/Procesos DIMPE/PilotoEMMET", mes=9,año=2024,tipo= "word")
  - Entrada: "EMMET_PANEL_tematica_mes_anio.csv"
  - Salida: archivo en Word o pdf, "boletin_fecha.docx"

Con la idea de optimizar el proceso de computación y ejecución de las funciones se crearon dos macros, que corren por bloque, cada una ejecuta cierta cantidad de funciones. En ese sentido, se optimiza aún más el proceso de automatización de resultados de la EMMET. Cada macro se puede ejecutar con facilidad y están repartidasde la siguiente manera:

- **Función macro 1 (macro1)**

  - Descripción: Ejecuta la función inicial, de Integración e Identificacion_alertas.
  - Ejemplo: macro1 (directorio="Documents/DANE/Procesos DIMPE/PilotoEMMET", mes=9,año=2024,avance= "100")

- **Función macro 2 (macro2)**

- Descripción: Ejecuta en orden las siguientes funciones; Imputacion, Tematica, aterritorial, anacional, c_regiones, c_dominios y boletín. Los argumentos que solicita la macro 2 es la unión de los argumentos de las funciones que se necesitan individualmente.
- Ejemplo: macro2 (directorio="Documents/DANE/Procesos DIMPE/PilotoEMMET", mes=9,año=2024,avance= “100”,tipo= “word”)

Esta partición de las funciones con la macro 1 se da para facilitar la revisión de alertas que genera la automatización y con ello, realizar la verificación por parte de los analistas. Una vez hecho esto, se ejecuta en cadena las demás funciones para generar los cuadros de salida, ahorrando tiempo y minimizando los errores en escritura de los argumentos de las funciones.

Diseño de la librería

Para comprender completamente su funcionalidad, es esencial explorar sus componentes clave, su flujo de trabajo y los parámetros que permiten su ejecución precisa.

Bases Insumo:

La librería se alimenta de tres bases de datos fundamentales:

- Base Logística: Esta base proporciona información importante para la ejecución de la encuesta, pues recopila el histórico de los datos ya tratados; es decir, con los procesos de crítica e imputación previamente realizados y adicionalmente con la información del mes a tratar. Actualmente el histórico contiene información desde 2018.
- Base Paramétrica: Esta base contiene la información asignada a cada establecimiento en cada una de las variables de identificación y publicación. Lo que permite una personalización y adaptación óptima a diferentes contextos.
- Base Deflactor: Contiene los datos de IPP _ PyC, IPP _ EXP, IPC. Estos están designados según el año, mes y código CIIU4. Se asocian a cada empresa según corresponda.

Las funciones que posee la librería se deben ejecutar en orden, ya que depende de los archivos de salida generados por la función anterior. Cada función posee unos argumentos de entrada, que permiten realizar la ejecución de acuerdo con el periodo de interés a tratar.

Los parámetros que se establecieron para las funciones son:

1. Directorio: debe otorgar la dirección de la carpeta en donde se alojarán los archivos de entrada y donde se desea que se almacenen los archivos de salida.
2. Mes: Corresponde al mes que desea analizar
3. Anio: Corresponde al año que desea analizar
4. Avance: el porcentaje de avance de recolección de la información reportada por los establecimientos
5. Tipo: parámetro exclusivo para el boletín, en el cual se especifica si la salida es en Word o pdf

Estos deben ser iguales en los argumentos de todas las funciones.

El proceso general se presenta de la siguiente manera:

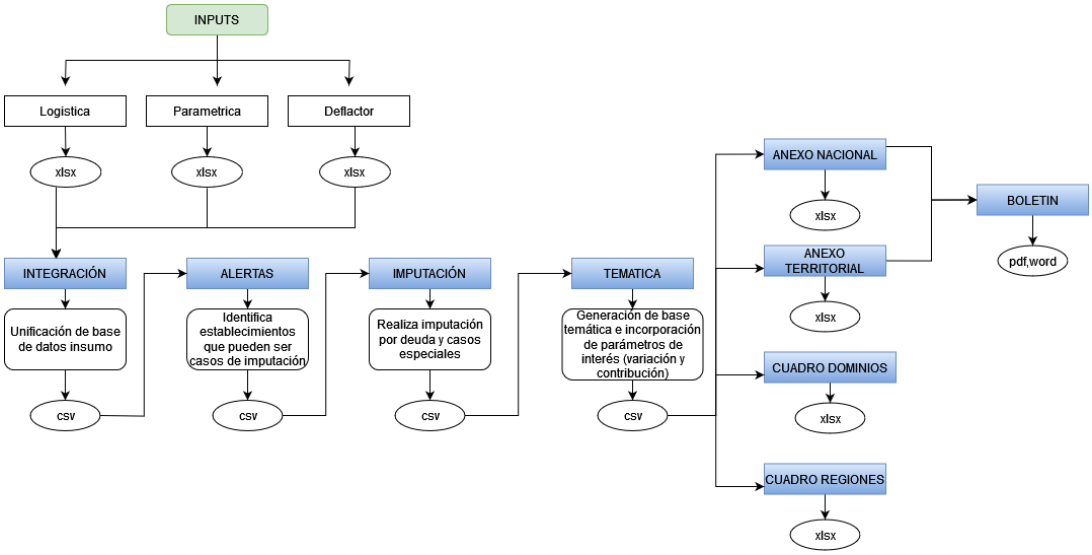


Figure 2: Flujo_libreria

Para poder hacer uso de la librería lo único que debemos hacer es ejecutar los siguientes 2 comandos en R

- Instalar librería en el R:

remotes::install_github("sub-dane/EMMET")

- cargar la libreria:
  library(DANE.EMMET)

Pruebas Metodológicas

Se realizaron varias pruebas metodológicas para determinar la función de alertas más efectiva y la función de imputación más adecuada. Estas pruebas incluyeron la evaluación de distintos algoritmos y técnicas, así como la comparación de su desempeño con métricas de calidad. Para la función de alertas, se analizaron factores como la precisión en la detección de eventos críticos y la capacidad de minimizar falsos positivos y negativos. En cuanto a la función de imputación, se exploraron diferentes métodos para rellenar datos faltantes, evaluando su impacto en la calidad y coherencia de los datos imputados. Los resultados obtenidos permitieron identificar las metodologías que ofrecían el mejor equilibrio entre precisión, eficiencia y aplicabilidad en el contexto específico del problema, asegurando así una mejora en la capacidad de respuesta y en la integridad de los datos.

- ## Metodologías identificación de alertas

Con el objetivo de definir una regla de identificación de los establecimientos y variables a imputar de la EMMET en cada uno de los periodos, se probaron las siguientes metodologías:

- **TSOUTLIERS**

Es un algoritmo usado para detectar y ajustar valores atípicos (outliers) de una serie de tiempo.

El algoritmo descompone la serie temporal, por sus componentes:

Tendencia
Estacionalidad
Residuo

En términos generales se buscar eliminar la estacionalidad y tendencia de la serie para detectar outliers del residuo. Los valores atipicos que pueden ser detectados, son:

Cambios permanentes en el nivel de la serie de tiempo (outliers tipo AO).
Cambios permanentes en la pendiente de la serie de tiempo (outliers tipo TC).
Cambios temporales en el nivel de la serie de tiempo (outliers tipo TC).
Cambios temporales en la pendiente de la serie de tiempo (outliers tipo IO).

- **CARTA DE CONTROL DE CALIDAD**

Es una herramienta utilizada en el control estadístico para monitorear un proceso a lo largo del tiempo y detectar cualquier variación significativa que pueda indicar un valor atípico; estas variaciones pueden ser atribuidas a causas comunes o especiales. Si el punto oscila dentro de los limites asumimos que la variabilidad es debido a causas comunes, si no, se asume como una causa especial, en este caso que el dato es un outlier. Los limites de la carta están construidos de la siguiente manera:

$$(\bar{x} - 1.96 * S\,, \bar{x} + 1.96 * S)$$

Donde $\bar{x}$ es el promedio de la serie histórica de la variable y establecimiento de interés y S es la desviación estándar de la serie histórica de la variable y establecimiento de interés.

- **ALGORITMO DE TUKEY**

También conocido como método de los rangos de Tukey o procedimiento de Tukey para la identificación de valores atípicos, es un método estadístico utilizado para detectar valores atípicos en un conjunto de datos univariados.
Este algoritmo se basa en la idea de que los valores atípicos son observaciones que se desvían

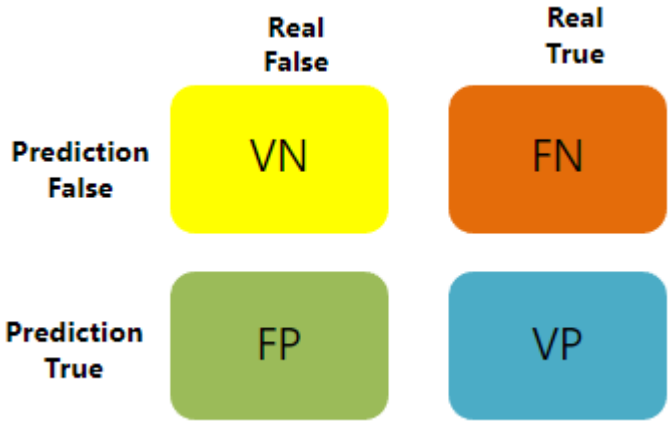
significativamente del resto de los datos. Para detectar estos valores, el algoritmo utiliza una medida de dispersión llamada rango intercuartílico (RIC), que se define como la diferencia entre el tercer y y primer cuartil de los datos. Los límites del intervalo están construidos de la siguiente manera:

$$(Q_1 - 1.5 * RIC\ , Q_3 + 1.5 * RIC)$$

Para poder realizar una comparación entre modelos se decidió utilizar las métricas de calidad de una matriz de confusión

- **MATRIZ DE CONFUSIÓN**

Es una tabla que muestra el número de resultados falsos y verdaderos que produjo el algoritmo/modelo en comparación con los resultados reales. La matriz de confusión suele estar formada por cuatro cuadrantes: verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP) y falsos negativos (FN). Cada cuadrante muestra la cantidad de veces que el modelo clasifico correcta o incorrectamente un resultado.



Las métricas que se utilizaron son las siguientes:

1. Accuracy: $\frac{VN+VP}{VN+VP+FN+FP}$
2. Recall: $\frac{VP}{VP+FN}$ o $\frac{VN}{VN+FP}$
3. Precisión: $\frac{VP}{VP+FP}$ o $\frac{VN}{VN+FN}$
4. F1-score: $\frac{2*recall*precisión}{recall+precisión}$

Finalmente se presentan las métricas por cada una de las diferentes variables a imputar en la encuesta

INFORMACIÓN PARA TODOS

Número de Aprendices y pasantes producción en etapa práctica (II_PP_AP_APEP)

TSOUTLIERS
Accuracy = 99,7%
Recall = 95,3%
Precisión = 96,2%
F1 = 95,7%

Predicción\Real	Sin cambios	Imputar
Sin cambios	3007	5
Imputar	4	101

CARTA DE CONTROL – 24 MESES
Accuracy = 98,7%
Recall = 99,1%
Precisión = 72,4%
F1 = 83,7%

Predicción\Real	Sin cambios	Imputar
Sin cambios	2971	1
Imputar	40	105

CARTA DE CONTROL - HISTÓRICO
Accuracy = 99,1%
Recall = 97,2%
Precisión = 79,8%
F1 = 87,7%

Predicción\Real	Sin cambios	Imputar
Sin cambios	2985	3
Imputar	26	103

ALGORITMO DE TUKEY
Accuracy = 98,9%
Recall = 98,1%
Precisión = 75,9%
F1 = 85,6%

Predicción\Real	Sin cambios	Imputar
Sin cambios	2978	2
Imputar	33	104

Se encontró que para las variables de capítulo 2 la mejor metodología es la carta de control teniendo en cuenta solo los últimos 2 años, mientras que para las variables de capítulo 3 la mejor metodología era la función tsoutliers, sin embargo, luego se encontró que esta metodología presentaba algunas fallas y a veces no detectaba valores atípicos que eran muy claros, por lo que se decide cambiar al comando locate.outliers.iloop. Finalmente, el procedimiento para identificar posibles establecimientos a imputar es:

1. Si la novedad es igual 5 sabemos que el individuo se va a imputar por deuda en todas sus variables o si el valor reportado en la variable de interés es cero y el mes inmediatamente anterior fue diferente de cero también se identifica como imputación deuda
2. Si la novedad es diferente a 5 se observa la variación con respecto al mes anterior; si es menor o igual a 20% no se imputa.
3. Si la variación es mayor a 20% y su valor es diferente de cero, entonces, primero se observa si el dato es igual a alguno de los datos reportados por el establecimiento en meses anteriores; en caso de que esté presente, se decide no imputar, si el dato no estuvo anteriormente, entonces se procede a identificar si la variable es referente a capítulo 2, si es así, se realiza una carta de control con los últimos 24 meses, si esta identifica el valor como atípico se imputa por caso especial; si la variable es referente a capítulo 3, entonces, se realiza una prueba mezclando el comando (locate.outliers.iloop) y la carta de control de 24 meses, el comando locate.outliers.iloop identifica los valores que pueden ser outliers en series de tiempo, si el valor fue identificado como un valor atípico se imputa por caso especial, en caso de que no fuera identificado como un valor atípico no se imputa.

Recordemos que las variables se dividen de la siguiente forma

Variables capítulo 2

```
II_PA_PP_NPERS_EP
AJU_II_PA_PP_SUELD_EP
II_PA_TD_NPERS_ET
AJU_II_PA_TD_SUELD_ET
```

II_PA_TI_NPERS_ETA
AJU_II_PA_TI_SUELD_ETA
II_PA_AP_AAEP
AJU_II_PA_AP_AAS_AP
II_PP_PP_NPERS_OP
AJU_II_PP_PP_SUELD_OP
II_PP_TD_NPERS_OT
AJU_II_PP_TD_SUELD_OT
II_PP_TI_NPERS_OTA
AJU_II_PP_TI_SUELD_OTA
II_PP_AP_APEP
AJU_II_PP_AP_AAS_PP
AJU_II_HORAS_HORDI_T
AJU_II_HORAS_HEXTR_T

Variables capítulo 3

AJU_III_PE_PRODUCCION
AJU_III_PE_VENTASIN
AJU_III_PE_VENTASEX
III_EX__VEXIS

- **Metodologías imputación de atípicos**

Con el objetivo de definir una regla de imputación de los establecimientos y variables de la EMMET en cada uno de los periodos, se probaron las siguientes metodologías iniciales:

- **HOT DECK**

Es un método de imputación que reemplaza los valores faltantes en un conjunto de datos utilizando valores conocidos de observaciones similares en la misma muestra. El funcionamiento consta de los siguientes pasos:

1. Selección de variables relevantes
2. Selección de las observaciones similares
3. Selección del valor para imputar

- **MICE**

Es una técnica de imputación múltiple que utiliza modelos estadísticos para imputar valores faltantes en los datos. El proceso implica dividir los datos en subconjuntos y luego imputar los valores faltantes en cada subconjunto por separado. Luego los datos imputados se utilizan en la siguiente iteración para imputar valores faltantes en otra variable. Este proceso se repite para cada variable con valores faltantes.

En resumen, MICE es una técnica que permite imputar valores faltantes en los datos mediante la estimación de valores basados en los patrones de los datos observados.

- **KNN**

Es una técnica de imputación de datos que utiliza los k registros más cercanos para predecir los valores faltantes. En resumen, esta metodología funciona de la siguiente manera:

1. Calcular la distancia entre los registros
2. Seleccionar los k registros más cercanos
3. Utilizar los valores de los vecinos para predecir los valores faltantes
4. Repetir para cada registro con valores faltantes

- **METODOLOGÍAS EMMET**

Existen varias metodologías dentro de la EMMET para la imputación por deuda e imputación por casos especiales, entre estas están:

1. Entre dominios: Se agrupan los establecimientos por su valor en CIIU4, se observa la variación que hay entre los establecimientos de ese dominio y se aplica esa variación al mes anterior.
2. Energía: Se observa la correlación que hay entre el consumo de energía y la variable a imputar, se observa la variación que hubo en el consumo y se aplica al valor del mes anterior.
3. Estacionalidad: Se observa la variación que hubo el año pasado para el mes actual y el mes anterior, luego se aplica esa variación al mes anterior de este año.

Para poder realizar la comparación entre modelos se tienen en cuenta dos métricas la raíz del error cuadrático medio y el MAPE

- **RAÍZ DEL ERROR CUADRÁTICO MEDIO**

Es la raíz de la diferencia cuadrática promedio entre los valores observados y los valores predichos por el modelo. Matemáticamente, el RECM se calcula cono la raíz de la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos, dividido por el número total de observaciones menos uno:

$$RECM = \sqrt{\frac{1}{n-1}} * \sum (y_i - \hat{y}_i)^2$$

Donde n es el número total de observaciones, y_i es el valor observado en la i-esima observación y $\hat{y}_i$ es el valor imputado en la i-esima observación.

- **MAPE**

El error de porcentaje medio absoluto es una medida de error relativa que utiliza valores absolutos para evitar que los errores positivos y negativos se cancelen entre si y utiliza errores relativos para permitir la comparar la precisión de previsión entre métodos de serie de tiempo

$$MAPE = \frac{1}{n} * \sum_{i=1}^{n} \left| \frac{Real_i - Predicción_i}{Real_i} \right|$$

Se realizan pruebas para las observaciones a imputar de los meses de prueba y adicionalmente se toma

una pequeña muestra de establecimientos para probar las diferentes metodologías.

INFORMACIÓN PARA TODOS

Número de empleados permanentes (II_PA_PP_NPERS_EP)

Metodologías de imputación	
Método	**ECM**
Hot-deck ID_NUMORD	15,335
Hot-deck CIIU4	96,712
MICE	72,495
KNN ID_NUMORD	10,939
KNN CIIU4	102,376
Método 1 (por CIIU4)	3,679
Método 2 (energía)	8,406
Método 3 (variación año anterior)	3,944
Imputación Mes anterior	3,606
TSOUTLIERS	34,020

Dado que para las variables de capítulo 2 siempre se ha imputado por el valor del mes anterior y a solicitud de los temáticos de la EMMET se dejo está metodología; para las variables en capítulo 3 se encontró que la mejor metodología de esas pruebas iniciales era un KNN combinado.

El KNN combinado consta de los siguientes pasos:

1. Realizar un primer KNN imputando los individuos atípicos en las variables de interés (KNN1).
2. Calcular la variación con respecto al mes anterior.
3. Realizar un segundo KNN para la variación con respecto al mes anterior de los establecimientos del mismo dominio (KNN2).
4. Calcular el valor final de la siguiente manera:

$$\hat{y} = KNN_1 * (1 + KNN_2) \quad (1)$$

Donde $\hat{y}$ es el valor resultante con el que se imputara la variable de interés.

Para el caso de imputación deuda en capítulo 3 se sigue la siguiente formula:

$$\hat{y} = MA * (1 + KNN_2) \quad (2)$$

Esto es que el valor resultante con el que se imputara la variable de interés es igual al valor del mes anterior, multiplicado por 1 más la variación con respecto al mes anterior de los establecimientos del mismo dominio.

Sin embargo, se encontró que para las imputaciones por deuda la metodología de vecinos cercanos presentaba fallas, ya que al no tener ninguna información del establecimiento no hay como encontrar los vecinos más cercanos. Por esto se decide implementar unas nuevas metodologías: ARIMA, ARIMAX y filtro de Kalman.

Se decide implementar estas metodologías por las siguientes razones:

1. Estructura temporal
2. Estacionalidad
3. Robustez y flexibilidad

AJU_III_PE_PRODUCCION - CASOS ESPECIALES

INFORMACIÓN PARA TODOS

Metodologías de imputación								
Método	**JUL-EMC**	**JUL-MAPE**	**AGO-EMC**	**AGO-MAPE**	**SEP-EMC**	**SEP-MAPE**	**MUESTRA-ECM**	**MUESTRA-MAPE**
KNN	4.382.608,65	35,712773	601.301,91	36,8293632	953.783,73	27,7400764	4.047.309,86	30,4400581
KNN-VAR-MES	4.118.637,03	36,6793801	3.034.124,62	61,3282655	2.721.493,57	53,4259293	4.013.556,43	31,6613003
ARIMA-HISTORICO	1.221.482,14	9,14697021	1.142.709,64	14,1585593	1.409.904,78	49,966997	4.055.801,00	31,3141588
ARIMA-2AÑOS	1.235.122,93	9,47173423	1.199.143,30	13,628633	1.906.259,51	7,90306212	4.356.422,62	34,0801382
ARIMAX	1.408.971,43	11,135646	809.932,35	13,3539314	1.594.020,75	36,4633311	4.470.957,32	31,6483452
FILTRO-KALMAN	1.014.023,90	9,33496035	1.056.516,97	18,2915074	1.397.380,85	16,4426078	4.433.638,31	32,6438678

Se encontró que tanto las metodologías de la familia ARIMA como el filtro de Kalman eran buenas opciones, dado que en los años del COVID-19 esta información genera ruido, se decidió implementar el ARIMA teniendo en cuenta solo los dos últimos años, para esto se usa el comando autoarima, el cual tiene un defecto y es que cuando el modelo propuesto por autoarima es de la forma (0,1,0) la imputación es el valor del mes anterior, lo cual para las variables del capítulo 3 se requiere que sea diferente; para solucionar esto se hace un hibrido entre ARIMA-2AÑOS y FILTRO-KALMAN

Finalmente, la metodología de imputación es la siguiente:

1. Para las variables de capítulo 2 la metodología a usar es la imputación por el mes anterior.
2. Para las variables de capítulo 3 la metodología a usar para imputación es un ARIMA teniendo en cuenta los últimos 2 años y si el modelo propuesto es un modelo de la forma (0,1,0) se usa la metodología de filtro de Kalman.