ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

NUEVA METODOLOGÍA PARA LA DETECCIÓN DE ANOMALÍAS UTILIZANDO TÉCNICAS DE MINERÍA DE DATOS Y BOOTSTRAP: CASO DE APLICACIÓN EN EFICIENCIA ENERGÉTICA

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO MATEMÁTICO

PROYECTO DE INVESTIGACIÓN

BRYAN ANDRÉS TOBAR TORRES
bryan_tobar@hotmail.es

Director: MIGUEL ALFONSO FLORES SÁNCHEZ, PHD
miguel.flores@epn.edu.ec

QUITO, ABRIL 2021

DECLARACIÓN

Yo BRYAN ANDRÉS TOBAR TORRES, declaro bajo juramento que el trabajo aquí escrito 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 cedo mis derechos de propiedad intelectual, correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su reglamento y por la normatividad institucional vigente.

Bryan Andrés Tobar Torres

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por BRYAN ANDRÉS TOBAR TORRES, bajo mi supervisión.

Miguel Alfonso Flores Sánchez, PhD
Director del Proyecto

AGRADECIMIENTOS

Quiero agradecer a Dios por haberme puesto en el camino que estoy recorriendo, por mi familia que siempre ha estado pendiente brindando su preocupación y apoyo, por haberme permitido conocer a tantas personas tan geniales que he llegado a considerar parte de mi familia y aquellas que solo fueron pasajeras. Sobre todo quiero agradecer por la vida de mi Madre y por las oportunidades que ha brindado en mi vida y quiero dar agradecimiento especial a Miguel Flores quien me brindó su apoyo y a pesar de desacuerdos y dolores de cabeza que pude haberle a ocasionado sigue presto para el desarrollo de proyectos y del material presentado.

DEDICATORIA

A mi familia.

Índice general

Índice de figuras

Índice de cuadros

Terminología

- $\mathscr{X}$: conjunto de entrada.
- x: dato de entrada.
- $\mathscr{Y}$: conjunto de salidas.
- $f : \mathscr{X} \to \mathscr{Y}$: función objetivo.
- $\mathscr{D} \subset \mathscr{X}$: conjunto de datos, de aquí extraigo las entradas y salidas (si son conocidas) $(x_1, y_1), ..., (x_N, y_N)$ donde $Y_i = f(x_i)$ con $x_i \in \mathscr{D}$
- $\mathscr{H} = \{h : \mathscr{X} \to \mathscr{Y}\}$: conjunto de hipótesis, es el conjunto de todas las posibles funciones objetivo.
- $\mathscr{A}$: algoritmo.
- n: tamaño de la muestra.
- $\mathbb{P} = (P_1, P_2, \ldots, P_n)$: muestra.
- P_i: observación i-ésima de la muestra $\mathbb{P}$.
- $d(p_i, p_j)$: distancia entre p_i y p_j con $p_i, p_j \in \mathbb{P}$.
- $R_{\mathbb{P}}$: EL radio más grande que contiene a cada observación de la muestra de $\mathbb{P}$.

$$R_{\mathbb{P}} = \max_{p_i, p_j \in \mathbb{P}} d(p_i, p_j)$$

- $\mathcal{N}(p_i, r) = \{p \in \mathbb{P} : d(p, p_i) \leq r\}$
  Las observaciones que se encuentran dentro de la vecindad de p_i
- $n(p_i, r)$: la cardinalidad de $\mathcal{N}(p_i, r)$ i.e

$$n(p_i, r) = |\mathcal{N}(p_i, r)|$$

- $mean(n(p_i, r, \alpha))$: La suma del número de puntos que conforman las sub-vecindades $\mathcal{N}(p, \alpha r)$ con $p \in N(p_i, r)$ y $p_i \in \mathbb{P}$, dividida entre $n(p_i, r)$.

$$mean(n(p_i, r, \alpha)) = \frac{\sum_{p \in N(p_i, r)} n(p, \alpha r)}{n(p_i, r)}$$

  Donde $\alpha \in (0, 1)$ y se la conoce como relajación.

- $\sigma(p_i, r, \alpha)$: La raíz cuadrada de la suma del número de puntos que conforman las sub-vecindades $\mathcal{N}(p, \alpha r)$ menos $mean(n(p_i, r, \alpha))$ al cuadrado, dividido para $n(p_i, r)$, con $p \in N(p_i, r)$ y $p_i \in \mathbb{P}$.

$$\sigma(p_i, r, \alpha) = \sqrt{\frac{\sum_{p \in N(p_i, r)} (n(p, \alpha r) - mean(n(p_i, r, \alpha)))^2}{n(p_i, r)}}$$

- P_i^*: observación Bootstrap, que puede tomar cualquier valor de muestra.
- $\mathbb{P}^* = (P_1^*, P_2^*, \ldots, P_n^*)$: remuestra Bootstrap, esta es obtenida a partir de la muestra.
- F: distribución
- $\hat{F}$: aproximación de la distribución poblacional F,
- $R = R(\mathbb{P}, F)$: estadístico de interés.
- $R^* = R\left(\mathbb{P}^*, \hat{F}\right)$: estimador Bootstrap del estadístico.

Resumen

Se propone una nueva metodología (algoritmo LOCI-BOOTSTRAP) para la detección de valores atípicos en sistemas HVAC (calefacción, ventilación y aire acondicionado), a través del algoritmo de Correlación Local Integral (Local Correlación Integral, LOCI) y empleando técnicas Bootstrap con el objetivo de obtener una regla considerando la distribucion del score que emplea el método LOCI y poder mejorar la clasificación de las observaciones.
Esta metodología fue aplicada para el caso de las instalaciones de una tienda de ropa ubicada en Panamá, donde se registraron 24 lecturas diarias durante 434 días. En cada lectura, se monitorearon 15 variables que miden el confort térmico, calidad de aire y eficiencia energética.
Para el entrenamiento del algoritmo y evaluación, se consideran acontecimientos anómalos registrados por los operarios del sistema HVAC. En la etapa de entrenamiento, se estiman los parámetros que mejor se acoplen a los datos, creando un índice empleando el método LOCI a fin de obtener un score para cada una de las observaciones y estudiar su distribucion mediante la aplicación de técnicas Bootstrap para realizar las clasificaciones. Para la evaluación del desempeño del algoritmo se utiliza validación cruzada y a partir de estos resultados se compara con lo obtenido en estudios anteriores

Abstract

A new methodology (LOCI-BOOTSTRAP algorithm) is proposed for the detection of outliers in HVAC (heating, ventilation and air conditioning) systems, through Local Correlation Integral (LOCI) algorithm and using Bootstrap techniques in order to obtain a rule considering the score distribution used by the LOCI method and to improve the classification of the observations.
This methodology was applied to the case of the facilities of a clothing store located in Panama, where 24 daily readings were recorded during 434 days. In each reading, 15 variables measuring thermal comfort, air quality and energy efficiency were monitored.
For algorithm training and evaluation, anomalous events recorded by HVAC system operators are considered. In the training stage, the parameters that best fit the data are estimated, creating an index using the LOCI method in order to obtain a score for each of the observations and study their distribution by applying Bootstrap techniques to perform the classifications. For the evaluation of the performance of the algorithm, cross validation is used and from these results it is compared with those obtained in previous studies.

Capítulo 1

Introducción

1.1. Antecedentes

Actualmente, se ha vuelto común que dentro de las industrias (textiles, comercio, etc), se cuente con sistemas de calefacción, ventilación y aire acondicionado (heating, ventilation and air conditioning, HVAC), convirtiéndose en necesarios para la labor diaria (centros comerciales, supermercados, etc). Debido a esto es importante considerar los costes que impone su uso, por lo que es necesario el poder realizar los respectivos controles y monitorización a fin de evitar fallas. No obstante, esto no representa un problema, debido a que la tecnología actual permite almacenar grandes cantidades datos, respecto a los trabajos realizados, de una forma instantánea. Esto conduce a generar demandas con respecto al monitoreo de procesos, análisis de datos y detección de fallas. Sin embargo, la cantidad de datos que se maneja puede provocar que la elección de intervalos de mantenimiento sea corta, el potencial de las máquinas para un mayor rendimiento y eficiencia no sea utilizado lo suficiente y que en algunas ocasiones los problemas y fallas sean detectados demasiado tarde [22].

A fin de abordar esta problemática se han diseñado diferentes modelos con el objetivo de hallar los patrones, características y comportamientos que poseen estas fallas que debido a su escasa ocurrencia se conocen como anomalías (outliers) para tomar las debidas precauciones [20]. Su estudio posee una gran relevancia y es realizado por una de las ramas de la Minería de datos (Data mining), llamada conocimiento descubierto en bases de datos (Knowledge Discovery in Databases, KDD), el cual "combina técnicas del aprendizaje de máquina, reconocimiento de patrones, estadística, bases de datos, y visualización para automáticamente extraer conoci-

miento (o información), de un nivel bajo de datos (bases de datos" [12], a fin de obtener resultados sustanciales y consistentes que ayuden a la toma de decisiones.

Debido a su aplicación en diferentes contextos, el análisis de comportamientos atípicos ha provocado la creación de distintos métodos que buscan resolver esta problemática, abordados desde el punto de vista del KDD en conjunto con otras técnicas pertenecientes a la Minería de Datos (DM). Siendo clasificados en dos ramas conocidas como: Métodos de aprendizaje supervisado y no supervisado [1]. Con el fin de poder estudiar el comportamiento de las anomalías y generar normas que permitan identificarlas de forma óptima. Adicionalmente a esto se ha vuelto necesario buscar una manera intuitiva de mostrar los resultados para facilitar la toma de decisiones, relacionadas con: establecer políticas de mantenimiento que sean necesarias y oportunas, brindar eficiencia y óptimos resultados en los sistemas HVAC y facilitar la interpretación de los resultados a los involucrados en el proceso.

1.2. Justificación

El método a ser empleado es conocido como Local Correlation Integral (LOCI) [17], el cual tiene como base criterios de agrupación (clusters), distancias y de la densidad de los datos [10] al momento de colocar sus respectivas etiquetas (outlier, inlier), estudiando el caso de consumo energético en una tienda de ropa ubicada en Panamá [8]. En la cual se aplicó este algoritmo para mostrarlo como una alternativa válida para identificar las anomalías en la eficiencia energética de sistemas HVAC, siendo verificadas por gráficos de control, empleando el punto de vista del aprendizaje supervisado. Con la metodología explicada anteriormente y considerando la perspectiva usada en " Outlier detection with one-classclassifiers from ml and kdd." [12], donde se crea un índice de disimilitud empleando el método LOCI a fin de obtener un puntaje (score) para cada una de las observaciones, se pretende aplicar técnicas Bootstrap [23] sobre este puntaje y estudiar su distribución para realizar las clasificaciones. Mostrando como una de sus principales ventajas que al pertenecer a los métodos del KDD [4][13] posee la facilidad de adaptarse a nuevos comportamientos, generando las etiquetas de las anomalías sin necesidad de realizar nuevamente un estudio para volver a realizar un modelo.

1.3. Objetivos

1.3.1. General

Proponer una metodología de Minería de Datos para la detección de anomalías en el caso de eficiencia energética de sistema HVAC, que permita mejorar los planes de mantenimiento.

1.3.2. Específicos

- Elaborar una regla para la identificación de patrones que siguen aquellos días sobre los cuales se han detectado posibles anomalías, para generar sus respectivas etiquetas.
- Comparar los resultados obtenidos con los que se hallaron en el estudio realizado en Case Study of Anomaly Detection and Quality Control of Energy Efficiency and Hygrothermal Comfort in Buildings, Carlos Eiras-Franco et al, 2019, mostrando la validez del método.
- Elaborar una aplicación, que brinde de una manera ilustrativa los resultados para la toma de decisiones.

Capítulo 2

Marco teórico.

2.1. El problema de aprendizaje supervisado

La finalidad de cualquier tipo de aprendizaje es poder clasificar el conjunto de datos o encontrar una explicación para su comportamiento. Sin embargo, desde el punto de vista supervisado, toda la información relacionada con la clasificación de las observaciones se encuentra disponible, por ende la problemática se centra en hallar una función que se acople a las etiquetas de los datos y que se adapte a cualquier tipo de procedencia de las observaciones con el fin de obtener la clase a la que pertenecen nuevos conjuntos de datos. [1]

El aprendizaje comienza por la partición del conjunto de datos en una base de entrenamiento y otra de testeo, las cuales son construidas tomando observaciones aleatorias dentro del total de datos. El primer conjunto es usado para obtener una función que aproxime de manera correcta a las etiquetas de los datos, y el segundo para validar los resultados obtenidos. No obstante, está función no es única por lo que es necesario establecer medidas que permita diferenciar aquellas cuya clasificación se acople mejor a los datos que las de otras conocidas como **medidas de error**.

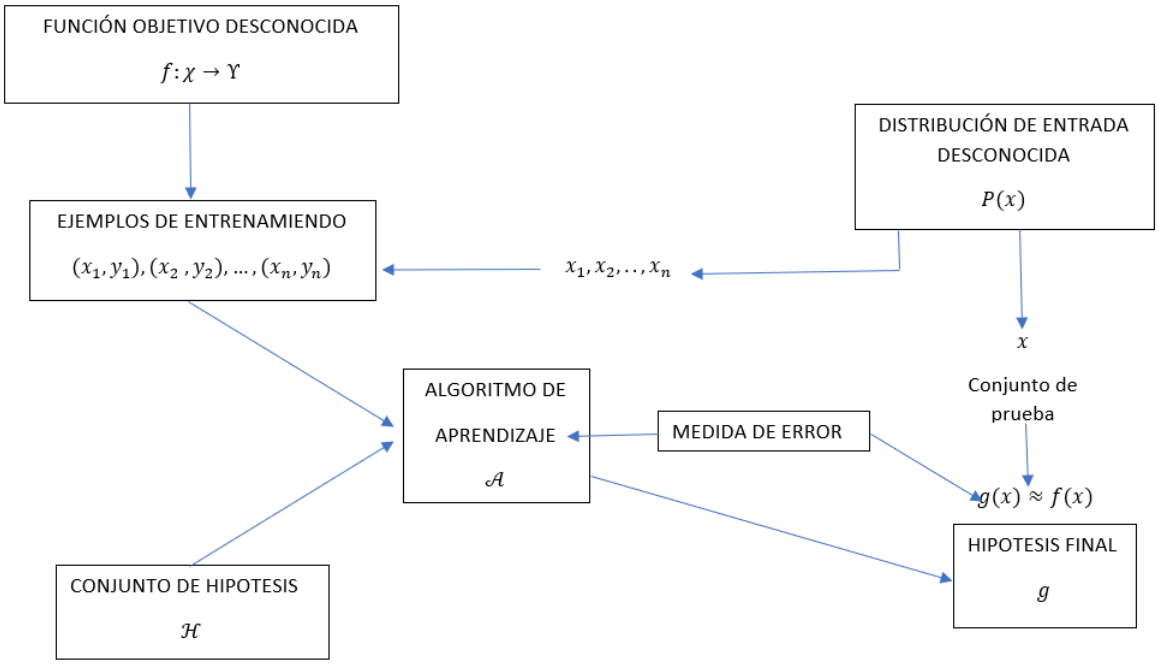

Figura 2.1: Esquema del problema de aprendizaje supervisado.
Recuperado de "Learning from data" [1].

OBSERVACIÓN 2.1 *El algoritmo $\mathscr{A}$ elige $g \in \mathscr{H}$ que mejor se ajuste a los datos.*

2.1.1. Error de entrada y salida

Si bien es posible que dentro del conjunto de Hipótesis exista una función que se acople de manera impecable a las etiquetas de los datos de entrenamiento, no quiere decir que esta sea la mejor, debido a que esto puede provocar sobre ajuste (overfitting), que consiste en que la hipótesis elegida solo se desenvolverá de manera correcta con conjuntos de datos que tengan un comportamiento equivalente a los cuales fueron considerados para su construcción. Es por ello que el objetivo es hallar una función que etiquete a los datos de manera óptima, es decir que dentro de está constarán errores de clasificación pero su cantidad será mínima. La forma de identificar los errores es a partir de una **Matriz de Confusión**, que permite comparar los valores predichos frente a los valores reales [4].
Sea f la función objetivo y $g \in \mathscr{H}$ la función elegida en el proceso de aprendizaje:

Cuadro 2.1: Matriz de confusión.

		f 0	1
g	0	Verdaderos Positivos (TP)	Falsos Negativos (FN)
	1	Falsos Positivos (FP)	Verdaderos Negativos (TN)

Métricas de rendimiento para las clasificaciones

Las métricas de rendimiento son medidas calculadas a partir de los valores obtenidos en la matriz de confusión, empleadas para evaluar la precisión de las predicciones. [4]

DEFINICIÓN 2.2 (Tasa de verdaderos positivos (sensibilidad, TPR)) *El porcentaje de datos clasificados correctamente cuando pertenecen al grupo 0.*

$$TPR = \frac{TP}{TP + FN}$$

DEFINICIÓN 2.3 (Tasa de verdaderos negativos (especificidad, TNR)) *El porcentaje de datos clasificados correctamente cuando pertenecen al grupo 1.*

$$TNR = \frac{TN}{TN + FP}$$

DEFINICIÓN 2.4 (Precisión global (ACC)) *El porcentaje de datos clasificados correctamente.*

$$ACC = \frac{TP + TN}{TP + FN + TN + FP}$$

DEFINICIÓN 2.5 (Precisión balanceada (BA)) *El promedio entre la TPR y TNR.*

$$BA = \frac{TPR + TNR}{2}$$

DEFINICIÓN 2.6 (Area Under The Curve Receiver Operating Characteristics (AUC–ROC)) *El área bajo la curva ROC es una medida de rendimiento para los problemas de clasificación binaria al variar el umbral. [5] La ROC es una curva de probabilidad que permite representar gráficamente a la sensibilidad frente a la razón de falsos positivos (1-especificidad), mientras que el valor de AUC representa el grado o la medida de separabilidad, indicando en qué medida el modelo es capaz de diferenciar entre grupos. Cuanto más alto sea el AUC, mejor será el modelo para distinguir entre clases (predecir 0s como 0s y 1s como 1s).*

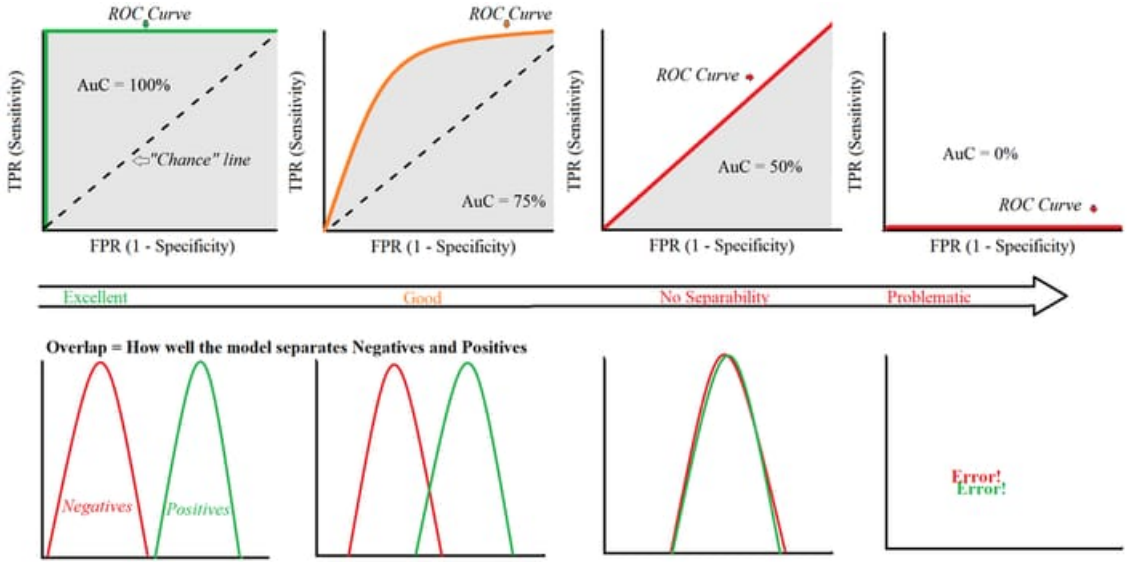

Figura 2.2: Comportamiento de la curva AUC-ROC.
Recuperado de:
`Stephanie Glen on March 9, 2019. Data Science Central, https://www.datasciencecentral.com/profiles/blogs/roc-curve-explained-in-one-picture.`

La curva ideal se encuentra en la parte izquierda de la imagen y posee un AUC del 100 %, lo que significa que va a ser capaz de distinguir entre grupos el 100 % de las veces. Cuanto más a la derecha de la imagen se encuentre la curva, peor es la detección. La curva del extremo derecho tiene el peor desempeño realizando clasificaciones al azar, mezclando los negativos y los positivos, lo que implica que probablemente se tiene un error en el modelo.

La hipótesis idónea es aquella que en la mayoría de etiquetas consten verdaderos positivos y negativos puesto a que esto implica altos valores en las tasas de verdaderos positivos y negativos, precisión global y balanceada. Por lo tanto el objetivo es que a partir del conjunto de entrenamiento hallar una función que asegure que la cantidad de falsos positivos y negativos sea mínima y que al utilizarla para clasificar otro conjunto posea un desempeño superior o similar. Este resultado es posible gracias a la desigualdad de Hoeffding.

TEOREMA 2.7 (Hoeffding) *Sean:*

- $g \in \mathscr{H}$.
- μ*: la probabilidad de que g etiquete de manera incorrecta a las observaciones en* $\mathscr{D}$.

- *ν: la proporción de observaciones etiquetas de manera incorrecta por g en la muestra $\mathscr{D}$.*
- *N el tamaño de la muestra $\mathscr{D}$.*
- *$\epsilon > 0$: tolerancia.*

A medida que N crece, se vuelve exponencialmente improbable que ν se desvíe de μ en más de ϵ.

$$Pr(|\nu - \mu| > \epsilon) \leq 2e^{-2\epsilon^2 N} \tag{2.1}$$

Para cualquier $\epsilon > 0$.

A ν y μ se los asocia con los errores de entrada y salida respectivamente y se definen de la siguiente manera:
Sean

- f la función objetivo.
- $g \in \mathscr{H}$
- $D \in \mathscr{D}$

DEFINICIÓN 2.8 (Error de entrada) *Es la fracción de observaciones en D, donde g y f difieren. Corresponde a ν.*

$$E_{in}(g) = \frac{1}{N}\sum_{n=1}^{N}[|g(x_n) \neq f(x_n)|]$$

Donde

$$[|Condicion|] = \begin{cases} 1 & \textit{si la condición es verdadera} \\ 0 & \textit{si la condición es falsa} \end{cases}$$

DEFINICIÓN 2.9 (Error de salida) *Es la probabilidad de que g y f difieran al evaluarlas en la población $\mathscr{X}$. Corresponde a μ.*

$$E_{out}(g) = P(g(x) \neq f(x))$$

Donde $x \in \mathscr{X}$

Substituyendo estos valores en la ecuación 3.1,se tiene que:

$$Pr(|E_{in}(g) - E_{out}(g)| > \epsilon) \leq 2e^{-2\epsilon^2 N} \tag{2.2}$$

En general para M hipótesis:

Sean $h_1, h_2, ..., h_M \in \mathscr{H}$ y $g \in \{h_1, h_2, ..., h_M\}$ la hipótesis para la cual se tiene la diferencia mínima de :

$$|E_{in}(g) - E_{out}(g)|(*)$$

y que cumple:

$$|E_{in}(g) - E_{out}(g)| > \epsilon(**)$$

De $(*)$ y $(**)$ se tiene que:

$$\begin{aligned} |E_{in}(g) - E_{out}(g)| > \epsilon \Longrightarrow |E_{in}(h_1) - E_{out}(h_1)| > \epsilon \qquad o \\ |E_{in}(h_2) - E_{out}(h_2)| > \epsilon \qquad o \\ .... \\ |E_{in}(h_M) - E_{out}(h_M)| > \epsilon \end{aligned}$$

De donde

$$\begin{aligned} P(|E_{in}(g) - E_{out}(g)| > \epsilon) \leq P(|E_{in}(h_1) - E_{out}(h_1)| > \epsilon \qquad o \\ |E_{in}(h_2) - E_{out}(h_2)| > \epsilon \qquad o \\ .... \\ |E_{in}(h_M) - E_{out}(h_M)| > \epsilon) \end{aligned}$$

Dado que las hipótesis son independientes de la proposición anterior se tiene:

$$P(|E_{in}(g) - E_{out}(g)| > \epsilon) \leq \sum_{j=1}^{M} P(|E_{in}(h_j) - E_{out}(h_j)| > \epsilon)$$

Finalmente de la ecuación de Hoeffding se obtiene:

$$Pr(|E_{in}(g) - E_{out}(g)| > \epsilon) \leq 2Me^{-2\epsilon^2 N} \tag{2.3}$$

para cualquier $\epsilon > 0\square$
La ecuación (2.3) puede ser reformulada de la siguiente manera:

Tomando una tolerancia δ y afirmando con una probabilidad de $1-\delta$ que:

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2N} ln(\frac{2M}{\delta})} \tag{2.4}$$

Este resultado se debe a que de la desigualdad de Hoeffding (3.3) se tiene que con una probabilidad de al menos $1-2Me^{-2\epsilon^2 N}$ i.e $\delta = 2Me^{-2\epsilon^2 N}$,de donde se tiene que:

$$|E_{out}(g) - E_{in}(g)| \leq \epsilon$$

Lo que implica que:

$$E_{out}(g) \leq \epsilon + E_{in}(g) \tag{2.5}$$

Ahora bien, dado que $\delta = 2Me^{-2\epsilon^2 N}$ se tiene que:

$$\frac{\delta}{2M} = e^{-2\epsilon^2 N}$$

$$ln(\frac{\delta}{2M}) = -2\epsilon^2 N$$

$$ln(\frac{2M}{\delta}) = 2\epsilon^2 N$$

$$\epsilon = \sqrt{\frac{2}{2N} ln(\frac{2M}{\delta})}$$

Remplazando en (2.5) se obtiene el resultado deseado □

Sin embargo, la ecuación (2.4) depende tanto del número de hipótesis como del tamaño de la muestra, para para obtener un resultado dependa únicamente del tamaño de la muestra utiliza la dimensión de Vapnik-Chervonenkis (VC) que establece: "el mayor número de puntos que un algoritmo puede separar" [Moore Andrew, 2009]. Para lo cual se requiere aclarar tanto los conceptos de dicotomía como el número máximo de dicotomías.

DEFINICIÓN 2.10 (Dicotomía) *Centrando el análisis en las funciones objetivo binarias i.e $f : \mathscr{X} \to \{0,1\}$ y sean:*

- $x_1, x_2, .., x_n \in \mathscr{X}$
- $g \in \mathscr{H}$

Una dicotomía es el resultado de aplicar g a la cantidad finita de observaciones

$x_1, x_2, .., x_n$, obteniendo $g(x_1), g(x_2), .., g(x_n)$ la cual divide al conjunto de datos en dos grupos.

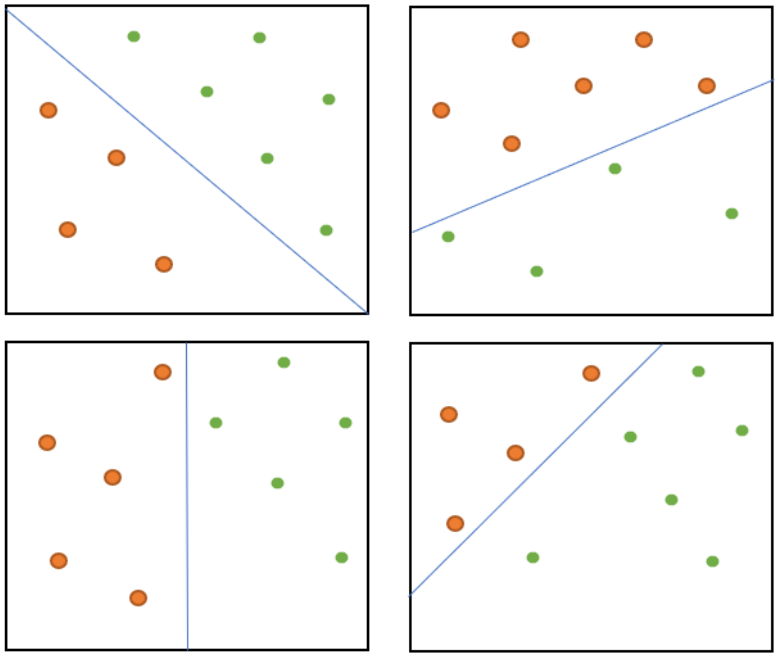

Figura 2.3: Ejemplo Dicotomías.
Recuperado de "Learning from data" [1].

DEFINICIÓN 2.11 ($m_{\mathscr{H}}(N)$) *Es el máximo número de dicotomías en N puntos.*

Por ejemplo, tomemos a $\mathscr{H}$ como el conjunto de todas as hipótesis h de la forma:

$$h(x) = \begin{cases} 1 & \text{if } x \geq a \\ 0 & \text{if } x < a \end{cases}$$

donde $a \in \mathbb{R}$.

El comportamiento de las funciones depende de que los valores de la muestra se encuentre a la izquierda o derecha de **a** de la siguiente manera:

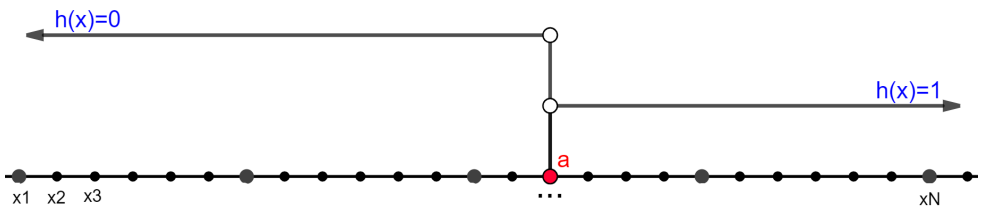

Figura 2.4: Comportamiento de la hipótesis perteneciente al conjunto $\mathscr{H}$ del ejemplo correpondiente a las definición 3.11

Recuperado de "Learning from data" [1].

Se tiene un $m_{\mathscr{H}}(N) = N + 1$, relacionándolo con la cantidad de formas diferentes en que se puede fijar al valor de **a** dividiendo a la recta en dos.

TEOREMA 2.12 (Límite de generalizacion de VC) *Para una tolerancia* $\delta > 0$,

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} ln(\frac{4m_{\mathscr{H}}(2N)}{\delta})} \tag{2.6}$$

Con una probabilidad mayor o igual $1 - \delta$.

2.1.2. Cantidad de datos de entrenamiento y testeo

Uno de los objetivos específicos del aprendizaje es lograr determinar el tamaño de la muestra N, con el fin de obtener un buen rendimiento dentro de otro conjunto de observaciones, para ello se debe establecer tanto un nivel de error ϵ y un nivel de confianza δ, el primero nos indica que tan semejantes serán las clasificaciones entre la hipótesis seleccionada y la función objetivo, mientras que la segunda la cantidad de veces en la que esto no sucederá. "La rapidez con la que N crece a medida que ϵ y δ se vuelven pequeños indica cuántos datos se deben tener para obtener una buena generalización."[1].
Aclarando que dentro del ámbito del aprendizaje en general, cuando únicamente se cuenta con una hipótesis se trata de un problema de verificación, es decir queremos saber si la hipótesis empleada se aproxima de forma correcta a la función objetivo, y por ende se contaría solamente con una dicotomía.
A partir de la ecuación (2.6) se obtiene para un δ fijo que:

$$E_{out}(g) - E_{in}(g) \leq \sqrt{\frac{8}{N} ln(\frac{4m_{\mathscr{H}}(2N)}{\delta})}$$

Por lo tanto con el objetivo de que el lado izquierdo sea lo más pequeño posible,

se necesita que:

$$\sqrt{\frac{8}{N} ln(\frac{4m_{\mathscr{H}}(2N)}{\delta})} \leq \epsilon$$

De donde:

$$N \geq \frac{8}{\epsilon^2} ln(\frac{4m_{\mathscr{H}}(2N)}{\delta}) \tag{2.7}$$

De la ecuación (2.7) colegimos que podemos conocer el tamaño de la muestra estableciendo un error, un nivel de confianza y conociendo la cantidad máxima de dicotomías a emplearse para el estudio.

2.2. Anomalía

En la minería de datos, una anomalía se considera una observación o un grupo de observaciones de escasa ocurrencia, cuyo comportamiento difiere del resto de datos.

DEFINICIÓN 2.13 *"Una anomalía es una observación que se desvía tanto de otras observaciones que despierta la sospecha de haber sido generado por un mecanismo diferente" [Hawkings, 1980].*

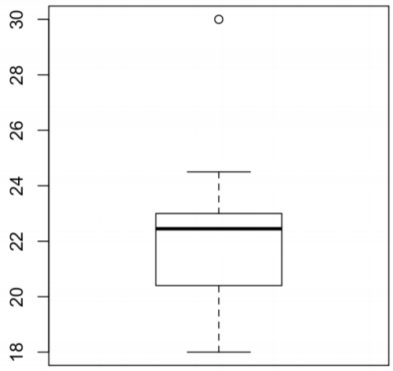

Figura 2.5: Ejemplo de anomalías en un conjunto de datos.
Elaboración: Autor

La anomalías pueden ser halladas en diversas ramas de aplicación, por ejemplo:

- Medicina: donde se detectan enfermedades a partir de síntomas los cuales se hallan fuera de los estándares normales y a partir de estos se dan los diagnósticos. [11]

- Mercados financieros: el caso de los fraudes que consisten en perjudicar a las diferentes entidades a fin de obtener un beneficio económico. [21]
- Mercado eléctrico: tomando por ejemplo el caso del consumo, y diferentes tipos de eventos que pueden causar que este decaiga o aumente dependiendo de la circunstancia, por ejemplo en el caso de una avería en la maquinaria de un proceso provocando que el consumo de energía de una máquina sea excesivo en comparación al resto. [8]
- Cyberseguridad: el caso de las perpetraciones web o hacking. [3]

2.2.1. Métodos para la detección de anomalías.

Existen diversas técnicas pertenecientes al KDD, que permiten la extracción de información, cada una emplea diferentes consideraciones requeridas para obtener clasificaciones y extraer el conocimiento del conjunto de datos, todas son abarcadas en las siguientes ramas del aprendizaje:

- **Métodos de aprendizaje supervisado:** se refiere a aquellos métodos en los que se conoce la etiqueta de cada una de las observaciones, obtenidos a partir de registros históricos dentro de los cuales deben existir casos de anomalías para que a partir de estos poder estudiar su comportamiento y generar etiquetas futuras [19].
- **Métodos de aprendizaje no supervisado:** se refiere a aquellos métodos que ayudan a describir la estructura, comportamiento y demás características de los datos, debido a que no se tiene ningún tipo de conocimiento acerca de las etiquetas de los datos. Por lo general estos métodos trabajan con ciertos supuestos que cumplen las anomalías para clasificarlas [1].

Dentro de ambos puntos de vista se pueden constatar los siguientes enfoques, cada uno de los cuales posee sus propias consideraciones para la detección de anomalías:

- **Basados en la distribución:** Se desarrollan en base a modelos estadísticos a partir de los datos. Consisten en la aplicación de una prueba estadística para determinar si el comportamiento de un registro corresponde al modelo al que pertenece el resto de observaciones [9].

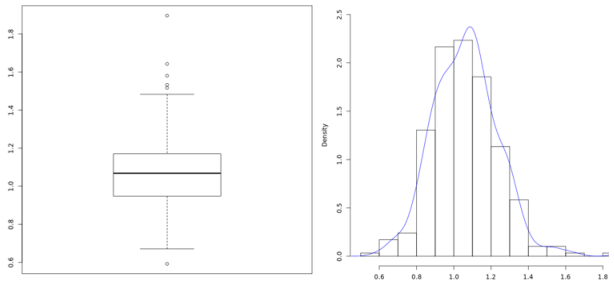

Figura 2.6: Ejemplo de clasificación basada en la distribución empleando el Test de Grubbs.

Elaboración: Autor

En la gráfica anterior se muestra el diagrama de caja y el histograma de un conjunto de datos, como se puede observar el histograma puede asociarse a una distribución normal y los valores que se encuentran fuera de los límites del diagrama son aquellos que corresponden a los valores extremos del histograma, es decir aquellos valores que se encuentran a una distancia de 3 desviaciones estándar de la media.

- **Basados en la distancia:** Se establece la distancia con la cual se realizará el análisis (Euclidea,Mahalanobis) y la cantidad de agrupaciones que se desean, finalmente se asocia a cada una de las observaciones de acuerdo al grupo donde se encuentren la mayoría de observaciones más cercanas a esta. Una anomalía es considerada como aquella que se encuentra más lejana al resto de su grupo [6].

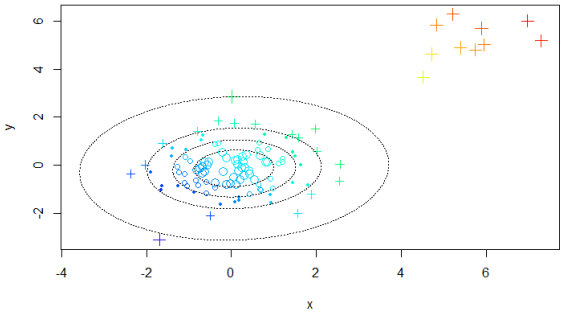

Figura 2.7: Ejemplo clasificación basada en la distancia empleando el algoritmo de vecinos más cercanos.

Elaboración: Autor

- **Basados en la densidad:** Como su nombre indica estos métodos se basan en las zonas con altas concentraciones de observaciones o regiones densas y aquellas con pocas de observaciones o regiones de baja densidad [20], para su análisis se tiene en cuenta las siguientes consideraciones:

  1. **radio** (ϵ): el radio de la vecindad
  2. **minpts** (umbral): la cantidad mínima de observaciones que debe tener una vecindad.

  Clasificando a las observaciones de la siguiente forma:

  - **Puntos Centrales:** aquellas observaciones que al ser consideradas como centro de la vecindad de radio ϵ, poseen una cantidad de vecinos mayor o igual a $minpts$. Suelen encontrarse regiones de alta densidad [10].
  - **Puntos Borde o frontera:** aquellas observaciones que al ser consideradas como centro de la vecindad de radio ϵ, poseen una cantidad de vecinos menor $minpts$, sin embargo, pertenecen a la vecindad de un punto central. Suelen encontrarse en regiones densas [10].
  - **Puntos Ruido (Anomalías) :** aquellas observaciones que no pueden ser consideras centros o fronteras. Suelen encontrarse en regiones de baja densidad [10].

  Ejemplo: Métodos LOCI,LOF.

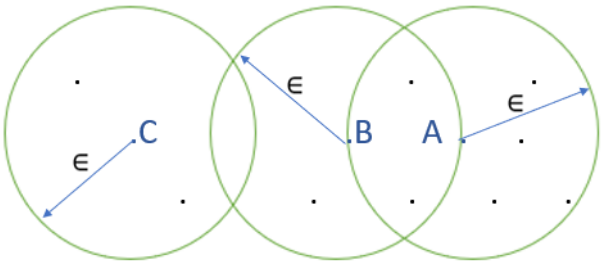

Figura 2.8: Puntos centrales, fronteras y ruidos.
Elaboración: Autor

- **Basados en agrupaciones:** Se analiza el comportamiento de las observaciones, una vez establecida la cantidad de agrupaciones se asigna a las variables de tal forma que la varianza entre grupos de observaciones sea máxima y varianza de las observaciones dentro de las agrupaciones sea mínima. [14]

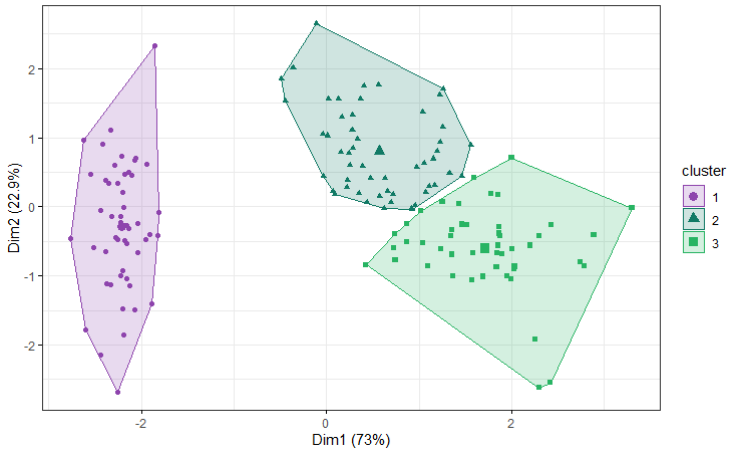

Figura 2.9: Ejemplo agrupaciones empleando el algoritmo de k-medias
Elaboración: Autor

2.3. El método LOCI.

El método LOCI pertenece a los algoritmos basados en la densidad, puesto que define a la densidad local de un punto en base al número de vecinos, se usa para la detección de datos atípicos tanto de forma individual como agrupaciones, mediante el **factor de desviación multigranular (MDEF)** que permite el análisis de las variaciones de la densidad local [2] .

En el siguiente gráfico se puede evidenciar la composición de vecindades para el punto p_0, obteniendo la cantidad de puntos que pertenecen a $\mathcal{N}(p_0, r)$, como también la cantidad de puntos que conforman las sub-vecindades $\mathcal{N}(p, \alpha r)$ con $p \in \mathcal{N}(p_0, r)$.

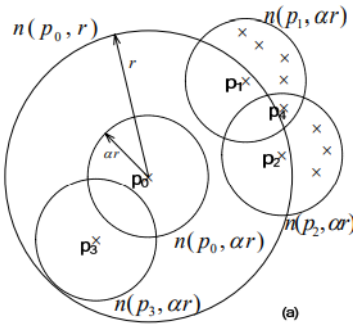

Figura 2.10: Gráfico de vecindades y sub-vecindades de p_0.
Recuperado de "Incremental outlier detection in data streams using local correlation integral." [15].

DEFINICIÓN 2.14 (MDEF) *El factor de desviación multigranular para una observación $p_i \in \mathbb{P}$, un radio r y una relajación α se define como:*

$$MDEF(p_i, r, \alpha) = \frac{mean(n(p_i, r, \alpha)) - n(p_i, \alpha r)}{mean(n(p_i, r, \alpha))}$$

$$MDEF(p_i, r, \alpha) = 1 - \frac{n(p_i, \alpha r)}{mean(n(p_i, r, \alpha))}$$

Interpretación:

- **MDEF>0**: Indica el número de observaciones dentro de la vecindad $N(p_i, \alpha r)$, es menor que el promedio de vecinos dentro de cada una de las sub vecindades ($mean(n(p_i, r, \alpha))$), cuan más cercano esté el valor del MDEF a 1, indica un mayor comportamiento anómalo.
- **MDEF$\leq$ 0**: Indica el número de observaciones dentro de la vecindad $N(p_i, \alpha r)$, es mayor o igual que el promedio de vecinos dentro de cada una de las sub vecindades ($mean(n(p_i, r, \alpha))$),este comportamiento se asocia a datos no anómalos.

DEFINICIÓN 2.15 (Desviación estándar normalizada)

$$\sigma_{mdef}(p_i, r, \alpha) = \frac{\sigma(p_i, r, \alpha)}{mean(n(p_i, r, \alpha))}$$

Consiste en llevar el valor de $\sigma(p_i, r, \alpha)$ a la misma escala del $MDEF(p_i, r, \alpha)$.

OBSERVACIÓN 2.16 *Por lo general a α se lo toma de tal manera que reduzca a la mitad el valor de r, siendo así que se considera un r_{min} que contenga al menos 20 observaciones, y un r_{max} igual a $R_{\mathbb{P}}$ que contiene a toda la muestra.*

DEFINICIÓN 2.17 *Una observación $p_i \in \mathbb{P}$ será marcada como una anomalía si: para cualquier radio $r \in [r_{min}, r_{max}]$, el valor del $MDEF$ es suficientemente grande, es decir:*

$$MDEF(p_i, r, \alpha) > k(\sigma_{mdef}(p_i, r, \alpha))$$

con $k > 0$.

En la practica se suele tomar k=3, esto debido a la aplicación de la desigualdad

de Chevyshev:

$$\begin{aligned} Pr(MDEF(p_i,r,\alpha) > k\sigma_{mdef}(p_i,r,\alpha)) &\leq Pr(|MDEF(p_i,r,\alpha)| > k\sigma_{mdef}(p_i,r,\alpha)) \\ &\leq \frac{(\sigma_{mdef}(p_i,r,\alpha))^2}{(k\sigma_{mdef}(p_i,r,\alpha))^2} \qquad Des.Chebyshev \\ &= \frac{1}{k^2} \end{aligned}$$

Cabe recalcar que este resultado se mantiene independientemente de la distribución.

ALGORITMO 2.18 *(Método LOCI)*

Pre-procesamiento

- *Para cada* $p_i \in \mathbb{P}$*:*

  Realizar una búsqueda de rango para

  $$N_i = \{p \in \mathbb{P} : d(p, p_i) \leq r_{max}\}$$

  A partir de los elementos pertenecientes a N_i*, construir una lista ordenada* D_i *de las distancias críticas y* $\alpha-$ *críticas de* p_i*.*

Post-procesamiento

- *Para cada* $p_i \in \mathbb{P}$*:*

  Para cada radio $r \in D_i$ *(ascendiente):*

  Actualizar $n(p_i, \alpha r)$ *y mean* $(n(p_i, r, \alpha))$

  Calcular

  $MDEF(p_i, r, \alpha)$ *y* $\sigma_{mdef}(p_i, r, \alpha)$

  Si $MDEF(p_i, r, \alpha) > 3\sigma_{mdef}(p_i, r, \alpha)$

  Etiquetar p_i *como un atípico.*

  Caso contrario:

  Etiquetar p_i *como una observación normal.*

2.3.1. Obtención del puntaje (score):

Debido a la forma en que el método se halla definido, se necesita realizar varias mediciones dentro de los datos para poder asegurar que un punto se trata de un dato atípico, es por ello que en [12] se propone la obtención de score llamado indice de disimilitud calculado a patir del MDEF y σ_{mdef}, como su nombre indica mide que tan diferente es una observación de las demás.

DEFINICIÓN 2.19 (Índice de disimilitud) *Sean:*

- $\mathscr{R}$*: conjunto de radios relevantes*

El índice de disimilitud se define como:

$$\delta_{LOCI} = max_{r \in \mathcal{R}} \left\{ \frac{MDEF\left(p_i, r, \alpha\right)}{\sigma_{mdef}\left(p_i, r, \alpha\right)} \right\}$$

DEFINICIÓN 2.20 (Radio relevante) *Sean* $r_1, r_2 \in [r_{min}, r_{max}]$*, si* δ_{LOCI}, r_1 *difiere significativamente de* δ_{LOCI}, r_2 *entonces* r_1, r_2 *son radios relevantes.*

OBSERVACIÓN 2.21 *Finalmente para etiquetar los datos de sigue el mismo procedimiento indicado en la definición 2.17.*

2.4. Remuestreo

El remuestreo es un método de inferencia estadística que emplea una variedad de técnicas que permiten valorar cuan preciso es un estimador de un parámetro de interés (media, varianza, mediana,etc), mediante la evaluación de sesgo y la variabilidad del estimador a partir de submuestras generadas del conjunto de muestra inicial.

Dos de los métodos más conocidos del remuestreo son los método Jackkanife [18] y Bootstrap [7].

2.4.1. Método Bootstrap

Bootstrap es un método de remuestreo propuesto por Bradley Efron en 1979 como un procedimiento informático. Este proceso se utiliza para estimar la distribución en el muestreo de un estadístico, la construcción de intervalos de confianza y realizar pruebas de hipótesis a partir de muestras aleatorias obtenidas del conjunto original. [7] Una de sus principales ventajas es la relajación sobre las hipótesis en el mecanismo por el cual fueron generados los datos facilitando la obtención de las propiedades asintóticas, adicionalmente en comparación a otros métodos posee una implementación sencilla. No obstante, este método posee potente necesidad computacional debido a la robustez de sus cálculos [23]. El método Bootstrap extrae de forma aleatoria las observaciones trabajando con B-conjuntos construidos a partir del

conjunto de datos considerando reemplazo, lo que significa que la misma observación se puede seleccionar varias veces en el mismo conjunto de datos Bootstrap, y el tamaño de cada pseudomuestra es del mismo que el tamaño de la muestra original.

Sea $\mathbb{P}$ una muestra aleatoria simple (m.a.s) con distribución F, estamos interesados en hacer inferencia sobre $\theta = \theta(F)$,Para ello necetamos conocer la distribución en el muestreo de $R(\mathbb{P}, F)$, cierto estadístico función de la muestra y de la distribución poblacional. Por ejemplo:

$$R = R(\mathbb{P}, F) = \theta(F_n) - \theta(F)$$

El método bootstrap consiste en el empleo de una estimación $\hat{F}$ de la distribución poblacional desconocida F, en este caso tomando a $\hat{F} = F_n$, donde F_n es la distribución empírica de la muestra y es utilizada para generar condicionalmente las remuestras a partir de $\mathbb{P}$ notadas por $\mathbb{P}^*$ y denominadas muestras bootstrap. Dichas muestras poseen una distribución $\hat{F}$ y sirven para obtener la distribución bootstrap de $R^* = R\left(\mathbb{P}^*, \hat{F}\right)$ que consiste en la aproximación muestral a la distribución de $R = R(\mathbb{P}, F)$.

Por ejemplo, considerando a $\mathbb{P} \sim F$ con media μ y varianza σ^2 conocida.

Desde el punto de vista de la inferencia estadística clásica:

$$\theta(F) = \mu = \int p dF(\mathbb{P}) = \int p f(p) dp$$

$$\theta(F_n) = \frac{1}{n} \sum P_i = \overline{p}$$

$$R = R(\mathbb{P}, F) = \sqrt{n} \frac{\overline{p} - \mu}{\sigma}$$

Bajo normalidad se tiene que $R \sim \mathcal{N}(0, 1)$, Si F no es normal bajo ciertas condiciones se sabe que $R \to^d \mathcal{N}(0, 1)$

Desde el punto de vista del método Bootstrap:

Considerando a $\hat{F} = F_n$, y el empleo de método Montecarlo [1] se generan B muestras bootstrap con el fin de analizar las propiedades del estadístico de la siguiente forma:

Para cada $i = 1, ..., n$ y $b = 1, .., B$, $P^{*(b)}(P_i^* = P_j) = \frac{1}{n}, j = 1, .., n$

Obtener $\mathbb{P}^{*(b)} = \{P_1^{*(b)}, P_2^{*(b)}, \ldots, P_n^{*(b)}\}$

[1] Algoritmos computacionales no deterministas que brindan una amplia variedad de soluciones a problemas matemáticos a partir del muestreo de números pseudoaleatorios

Calcular $R^{*(b)} = R(\mathbb{P}^{*(b)}, \hat{F}) = \sqrt{n}\frac{\overline{p}^{*(b)} - \overline{p}}{\sigma}$

Donde $\overline{p}^{*(b)} = \frac{1}{n}\sum P_i^{*(b)}$

Finalmente el funcionamiento del algoritmo Bootstrap mediante el empleo de técnicas Montecarlo puede ser bosquejado bajo el siguiente algoritmo:

ALGORITMO 2.22 *(Bootstrap empleando métodos Montecarlo)*

1. *Para cada $i = 1, ..., n$ arrojar P_i^* a partir de $\hat{F}$*
2. *Obtener $\mathbb{P}^* = \{P_1^*, P_2^*, \ldots, P_n^*\}$*
3. *Calcular $R^* = R\left(\mathbb{P}^*, \hat{F}\right)$*
4. *Reperir B-veces los pasos 1-3 para obtener réplicas Bootstrap $R^{*(1)}, ..., R^{*(B)}$*
5. *Utilizar estas réplicas para aproximar la distribución en el muestreo de R.*

2.4.2. Bootstrap Uniforme

El Bootstrap uniforme utiliza la distribución de empírica F_n como un reemplazo de la distribución poblacional desconocida.

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{P_i \leq x\}$$

Por lo tanto se tiene que $\hat{F} = F_n$ y $R^* = R(\mathbb{P}^*, F_n)$.

El empleo de los métodos Montecarlo se enfoca en hallar una aproximación de la distribución de R mediante R^*, empleando una gran cantidad (B) de replicas de Bootstrap de R^*.Debido a la función de distribución empírica cada una de las observaciones dentro de la muestra tendrán una probabilidad $\frac{1}{n}$ de ser elegida para formar parte de las muestras Bootstrap, de la siguiente manera:

Para cada $i = 1, ..., n$

$$P^*(P_i^* = P_j) = \frac{1}{n}, j = 1, .., n \tag{2.8}$$

donde $P_j \in \mathbb{P}$. Siendo bosquejado por el siguiente algoritmo:

ALGORITMO 2.23 *(Bootstrap Uniforme)*

1. *Para cada* $i = 1, ..., n$ *arrojar* P_i^* *a partir de* F_n*, es decir empleado la ecuación 2.8.*

2. *Obtener* $\mathbb{P}^* = \{P_1^*, P_2^*, \ldots, P_n^*\}$

3. *Calcular* $R^* = R\left(\mathbb{P}^*, F_n\right)$

4. *Reperir B-veces los pasos 1-3 para obtener réplicas Bootstrap* $R^{*(1)}, ..., R^{*(B)}$

5. *Utilizar estas réplicas para aproximar la distribución en el muestreo de R.*

Capítulo 3

Nueva metodología para la detección de anomalías

La metodología Bootstrap-LOCI consiste en el empleo de técnicas de validación cruzada para fijar un límite óptimo al momento de aplicar el método LOCI para la detección de anomalías. Este límite es determinado a partir de la creación de muestras Bootstrap sobre las cuales se ejecuta el algoritmo LOCI, donde se estudia la distribución Bootstrap del puntaje obtenido, permitiendo establecer un límite a partir de las propiedades asintóticas del puntaje en lugar del empleo de la desigualdad de Chevyshev como se mostró en la **definición 2.17**.

3.1. Propuesta del algoritmo Bootstrap-LOCI

En los estudios realizados en "Case study ofanomaly detection and quality control of energy efficiency and hygrothermalcomfort in buildings." [8], se muestra la efectividad del método LOCI al momento de etiquetar las fallas que se suscitaron en el empleo de los sistemas HVAC durante los días observados, siendo verificados mediante gráficos funcionales de control, mostrando que el método puede emplearse para el presente caso de estudio. No obstante, dentro del artículo se emplean dos umbrales diferentes al momento de aplicar el algoritmo, el primero como una alerta correspondiente al valor de 1.5 y el segundo valor como límite para las anomalías con un valor de 2.5. Sin embargo, el argumento para el empleo de estos valores se basa en la cantidad de días que eran detectados como anómalos al usarlos, siendo detectados alrededor de 1 día anómalo por cada 4 días de observación en el caso del la alerta y alrededor de 1 día anómalo por cada 10 días de observación para el límite

de las anomalías.

El método Bootstrap-LOCI busca una mejora en el umbral empleado para las clasificaciones del método LOCI mediante el estudio de la distribución de este, por tanto su enfoque no se centra en modificar el algoritmo de su método predecesor, si no en la búsqueda de un valor que permita detectar la mayoría de anomalías de forma correcta. Adicionalmente, dado que el método LOCI no posee supuestos sobre el origen de los datos el empleo de técnicas no paramétricas permite una buena aproximación de las propiedades asintóticas de los parámetros. Para hallar este límite se parte del empleo de técnicas de validación cruzada dividiendo a los datos en un conjunto de entrenamiento y otro de testeo. Sobre el primer conjunto se toma de forma aleatoria las observaciones que no fueron consideradas anómalas a fin de construir B-conjuntos de igual tamaño al conjunto de entrenamiento (muestras Bootstrap), con el objetivo de estudiar en el comportamiento de las observaciones no anómalas.

Sobre cada una de las muestras Bootstrap se ejecutará el método LOCI con intención de obtener el puntaje para cada una de las observaciones dentro de los conjuntos, con el cual se estudiará sus propiedades a fin de obtener el nuevo valor que reemplazará el límite usado en el método LOCI. Cabe aclarar que se deben establecer previamente los parámetros con los cuales se aplicará el método (radio y α). Cuando se establece el nuevo límite se procede a la aplicación del método LOCI sobre los datos de testeo obteniendo el puntaje y realizando las comparaciones necesarias para las clasificaciones. Análogamente al estudio realizado en [8] el método LOCI tendrá la misma efectividad en la detección de anomalías debido a que únicamente se modificará el límite para realizar las clasificaciones.

ALGORITMO 3.1 *(Método Bootstrap-LOCI)*

1. *Seleccionar* $\alpha \in (0, 1]$
2. *Seleccionar* $r \in [r_{min}, r_{max}]$
3. *Para cada* $i = 1, ..., n$ *arrojar* P_i^* *a partir de* F_n
4. *Obtener* $\mathbb{P}^* = \{P_1^*, P_2^*, \ldots, P_n^*\}$
5. *Calcular* $\delta^*_{LOCI,\alpha,r} = \delta_{LOCI,\alpha,r}\left(\mathbb{P}^*, F_n\right)$ *a partir del algoritmo LOCI*
6. *Reperir B-veces los pasos 3-5 para obtener réplicas Bootstrap* $\delta^{*(1)}_{LOCI,\alpha,r}, ..., \delta^{*(B)}_{LOCI,\alpha,r}$
7. *Utilizar estas réplicas para aproximar la distribución de* $\delta_{LOCI,\alpha,r}$.

8. *Obterner límite L a partir de $\delta^{*(1)}_{LOCI,\alpha,r}, ..., \delta^{*(B)}_{LOCI,\alpha,r}$*

9. *Aplicar algoritmo LOCI a $\mathbb{P}$ y obtener $\delta_{LOCI,\alpha,r}$*

10. *Decisión si $\delta_{LOCI,\alpha,r} > L$ marcar como anómalo.*

3.2. Caso de aplicación eficiencia energética

Para el caso de aplicación se trabaja con los datos tomados a partir del monitoreo realizado por Σqus a una tienda de ropa ubicada en Panamá, se cuenta con un total de 125047 observaciones tomadas diariamente cada 5 minutos durante las fechas del primero de agosto del 2017 hasta el 9 del octubre del 2018, los datos están distribuidos en 18 variables que corresponden a medidas de temperaturas interiores, impulsión y retorno, el consumo energético y la cantidad de CO2 de diferentes enfriadoras de maquinarias pertenecientes a los sistemas HVAC de la tienda.
Se procedió a agrupar los datos en horas tomando su media es decir se trabajará 24 observaciones por día, obteniendo un total 10416 registros completos equivalentes a 434 días diferentes y un con un total de 15 variables, con el fin de replicar los datos obtenidos en [8].

Las anomalías fueron etiquetadas a partir notificaciones realizadas por los encargados del mantenimiento siendo un total de 24 (11-sept, 21-22 sept, 29-sept, 16-31 oct, 17-20 nov del 2017) y la observación de la evolución del comportamiento en el transcurso del día de la potencia de enfriamiento dentro del conjunto de datos obteniendo 33 etiquetas adicionales (1-nov, 25-dic del 2017, 7-13-mar del 2018, etc.), dando un total de 57 días anómalos, que fueron marcados con el valor de 1 y a los días que se consideró que no tenían un comportamiento anómalo se los marcó con el valor de 0.

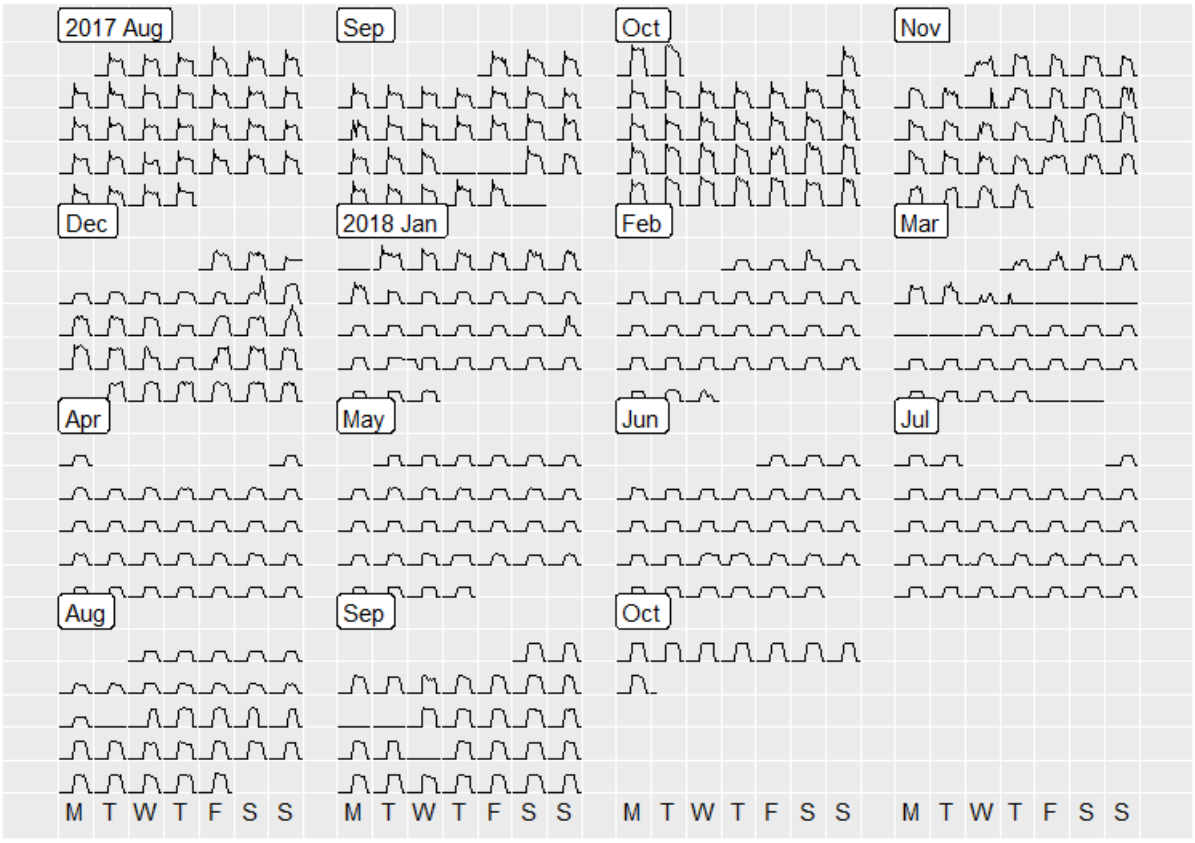

Figura 3.1: Comportamiento en el tiempo de la potencia de enfriamiento de los sistemas HVAC en la tienda.

Correspondiente a los meses de septiembre del 2017 a octubre del 2018.

Elaboración: Autor

Al aplicar la ecuación (2.7) con $m_{\mathscr{H}}(2N) = 1$ y dado que solamente contamos con una hipótesis, se obtuvo que el número de registros al variar el nivel de confianza y el error aceptado para los conjuntos de entrenamiento y testeo es el siguiente:

Cuadro 3.1: Cantidad de días y observaciones para los conjuntos de entrenamiento y testeo requeridos para tener un error ϵ y un nivel de confianza δ.

ϵ	δ	DÍAS ENTRENAMIENTO	DÍAS TESTEO	REGISTROS ENTRENAMIENTO	REGISTROS TESTEO
0.15	0.8	20	414	480	9936
0.1	0.8	44	390	1056	9360
0.05	0.8	174	260	4176	6240
0.15	0.85	22	412	528	9888
0.1	0.85	48	386	1152	9264
0.05	0.85	191	243	4584	5832
0.15	0.9	24	410	576	9840
0.1	0.9	54	380	1296	9120
0.05	0.9	214	220	5136	5280
0.15	0.95	29	405	696	9720
0.1	0.95	64	370	1536	8880
0.05	0.95	254	180	6096	4320
0.15	0.99	39	395	936	9480
0.1	0.99	87	347	2088	8328
0.05	0.99	347	87	8328	2088

Con un nivel de confianza del 90 % y un error aproximado del 5 %, se colige que lo ideal es trabajar con una cantidad mayor o igual a 214 días (5136 registros) en el conjunto de entrenamiento y 220 días (5280 registros) en el testeo, por lo tanto se procedió a trabajar el 50 % de datos para el conjunto de entrenamiento dejando el otro 50 % restante para el testeo, fijando una proporción de anomalías equivalente dentro de ambos conjuntos.

Cuadro 3.2: Distribución de los datos en los conjuntos de entrenamiento y testeo.

Conjunto	Cantidad de Días	Cantidad de Anomalías	Total Registros
Entrenamiento	217	28	5208
Testeo	217	29	5208

3.2.1. Aplicación del algoritmo LOCI

En primera instancia se procede a establecer los parámetros necesarios para obtener los resultados que se acoplen mejor a las etiquetas en los datos de entrenamiento.

Mediante el uso del lenguaje R y la librería que replica el algoritmo descrito en la sección 2.3.2 "DDoutlier" [16], con el fin de seleccionar el valor de α se escogió a aquel que al trabajar con el radio mínimo (contenga 20 observaciones dentro de su vecindad) logre capturar la mayoría de anomalías.

Cabe recalcar que para cada uno de los días se tendrá 24 puntajes asociados a cada una de las horas. Para establecer un día como anómalo basta con que el puntaje en una de estas horas supere el límite del método LOCI.

α	Cantidad Anomalias
0.1	0
0.15	0
0.2	0
0.25	0
0.3	0
0.35	0
0.4	0
0.45	0
0.5	0
0.55	2
0.6	3
0.65	3
0.7	4
0.75	5
0.8	5
0.85	4
0.9	5
0.95	5
1	4

Cuadro 3.3: Cantidad de anomalías detectadas al variar el valor de α con la cantidad mínima de vecindades.

Se estableció un $\alpha = 0{,}75$. Se procede a establecer el índice de disimilitud de la definición 2.6 calculado a partir de 100 radios diferentes que que se hallan entre r_{min} y r_{max}. Obteniendo el siguiente desempeño:

Cuadro 3.4: Matriz de confusión empleando el indice de disimilitud.

		Real	
		0	1
Predicho	0	89	100
	1	6	22

Al emplear el índice disimilitud es facil notar que se obtiene un rendimiento deficiente para la clasificación. Sin embargo, esto se puede subsanar debido a que se cuentan con 100 radios diferentes, sobre los cuales se ejecutó el algoritmo LOCI y se pueden generar las respectivas etiquetas que ya tienen asociado un puntaje. Se evaluó cada uno de estos estableciendo el que mejor se desenvuelve en el entrenamiento y será empleado en el modelo.

Cuadro 3.5: Matriz de confusión empleando los radios que contienen 3217 y 3269 observaciones con $alpha = 0{,}75$.

		Real	
		0	1
Predicho	0	178	11
	1	8	20

Para los radios mostrados en la tabla anterior se obtuvo el mejor desempeño de todo el conjunto al aplicar el método LOCI, con una tasa de aciertos del 91.24 % y una tasa de balanceada de aciertos del 82.80 %, de entre los dos se escogió el radio que contiene a 3217 observaciones puesto que es el más pequeño.

3.2.2. Obtención del límite Bootstrap-LOCI

A partir de los datos de entrenamiento se trabajó con B=500, es decir 500 muestras Bootstrap, datos construidos a partir de los 189 días considerados como no atípicos tomados al azar con reemplazo siguiendo la metodología del Bootstrap uniforme, equivalente a 4536 observaciones cada uno.

Se evaluó a cada uno de los conjuntos con el método LOCI, a fin de obtener el puntaje y estudiar la distribución Bootstrap, estableciendo el nuevo límite para las clasificaciones.

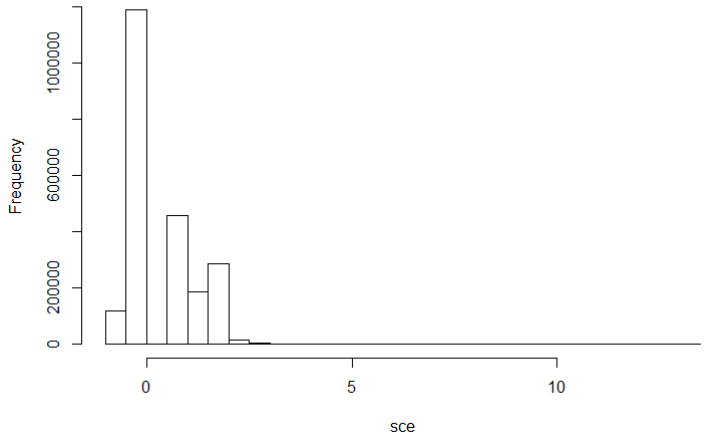

Figura 3.2: Histograma del score Bootstrap-LOCI .
Elaboración: Autor

Tomando como límite al valor que se encuentra en el percentil 99.

Cuadro 3.6: Percentiles del score Bootstrap-LOCI

50 %	75 %	90 %	99 %
-0.1402	0.7813	1.6034	2.0972

Capítulo 4

Resultados

A continuación se justificará el cumplimiento de los objetivos planteados en el presente estudio, mediante la evaluación del desempeño del método Bootstrap-LOCI para las clasificaciones y sus respectivas comparaciones.

4.1. Conjunto de entrenamiento

Se comparó los resultados obtenidos al aplicar los límites empleados en [8] (1.5 y 2.5), con el nuevo límite obtenido de la aplicación del método propuesto cuyo valor es igual a 2.0972.

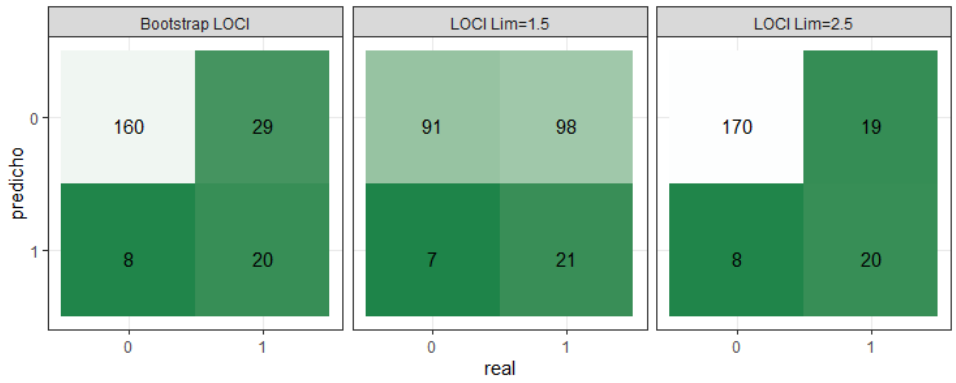

Figura 4.1: Matrices de confusión elaboradas con las predicciones del conjunto de entrenamiento con diferentes límites para las clasificaciones.
Izquierda: Límite igual a 2.0972. Derecha: Límite igual a 2.5. Centro: Límite igual a 1.5. **Elaboración:** Autor

En una breve inspección a los resultados de las matrices de confusión,parece que

el método LOCI empleando el límite de 2.5 es aquel que realiza mejor las clasificaciones, para garantizar estas aseveraciones se analizan los indicadores de desempeño.

Cuadro 4.1: Indicadores de los modelos empleando diferentes límites

Indicador	Bootstrap-LOCI Límite= 2,0972	LOCI Límite= 1,5	LOCI Límite= 2,5
sensibilidad (Tasa de aciertos positivos)	0.8466	0.4815	0.8995
especificidad (tasa de aciertos negativos)	0.7143	0.75	0.7143
precisión (tasa de aciertos)	0.8295	0.5161	0.8756
precisión balanceada	0.7804	0.6157	0.8069

Enfocando el análisis en el método Bootstrap-LOCI con respecto a los límites empleados en los estudios anteriores este presenta un desempeño superior que al usar el valor de 1.5 para clasificaciones. No obstante, al cambiar al valor de 2.5 el método Bootstrap-LOCI posee un desempeño similar en cuanto a aciertos en clasificación de anómalos, pero existe una diferencia considerable al momento de catalogar a las observaciones no anómalas, lo que conlleva a que al emplear el valor de 2.5 realice una mejor diferenciación entre los grupo i.e realice un mejor etiquetado.Lo expuesto anteriormente se puede evidenciar en las curvas AUC-ROC.

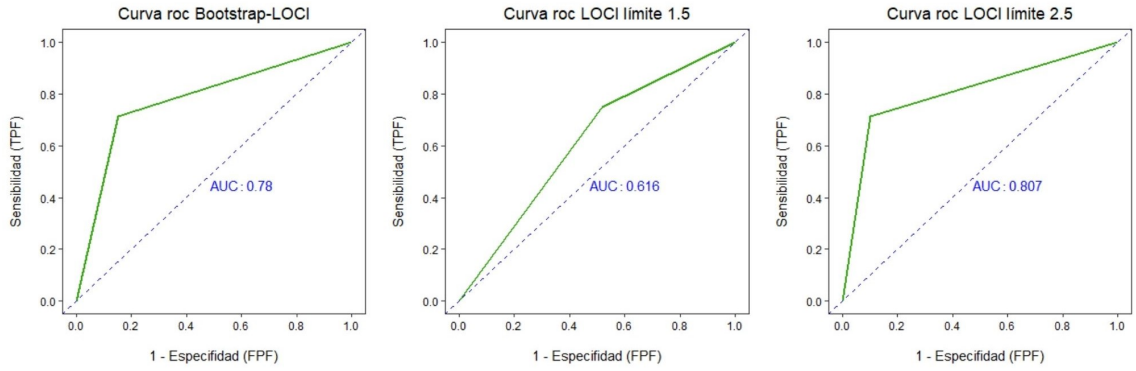

Figura 4.2: Curvas AUC-ROC evaluación del modelo con diferentes límites en el conjunto de entrenamiento.
Izquierda: Límite igual a 2.0972. Derecha: Límite igual a 2.5. Centro: Límite igual a 1.5. **Elaboración:** Autor

4.2. Conjunto de testeo

Aplicando la misma metodología para juzgar a las observaciones que en el conjunto de entrenamiento,en el testeo se obtienen los siguientes resultados:

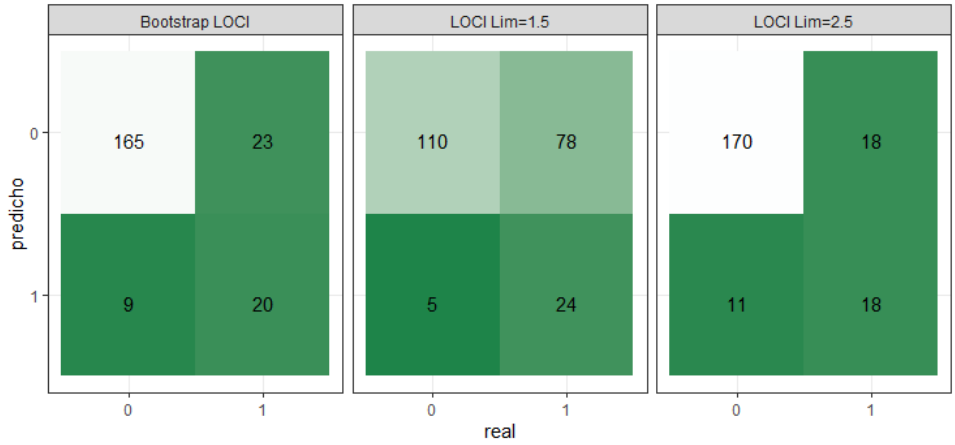

Figura 4.3: Matrices de confusión elaboradas con las predicciones del conjunto de testeo con diferentes límites para las clasificaciones.
Izquierda: Límite igual a 2.0972. Derecha: Límite igual a 2.5. Centro: Límite igual a 1.5. **Elaboración:** Autor

A diferencia del entrenamiento al observar las matrices de confusión, el método Bootstrap-LOCI presenta pequeñas diferencias en las cantidades de observaciones etiquetadas correctamente respecto del método LOCI al emplear el valor del límite igual a 2.5. Para poder asegurar que método posee el mejor desempeño se deben analizar los indicadores.

Cuadro 4.2: Indicadores de los modelos empleando diferentes límites

Indicador	Bootstrap-LOCI Límite= 2,0972	LOCI Límite= 1,5	LOCI Límite= 2,5
sensibilidad (tasa de aciertos positivos)	0.8777	0.5851	0.9043
especificidad (tasa de aciertos negativos)	0.6897	0.8276	0.6207
precisión (tasa de aciertos)	0.8295	0.6175	0.8664
precisión balanceada	0.7837	0.7063	0.7624

El método Bootstrap-LOCI presenta un mejoría en cuanto a la detección de días anómalos, a pesar de que se ve superado por el que método LOCI que posee un límite igual a 2.5 en cuanto a la detección de días no anómalos, al igual que en el entrenamiento se muestra que el peor desempeño del método resulta de tomar como límite el valor de 1.5, finalmente se tiene que en el testeo la mejor discriminación de datos se obtiene al aplicar el algoritmo Bootstrap-LOCI. Se puede evidenciar este resultado al emplear las curvas roc como se muestra a continuación.

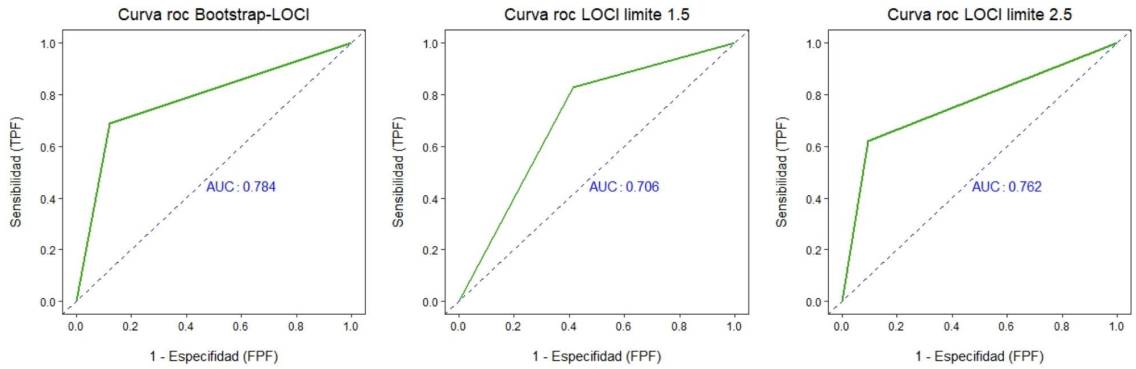

Figura 4.4: Curvas ROC evaluación del modelo con diferentes límites en el conjunto de testeo.

Elaboración: Autor

En la siguiente gráfica se muestra el comportamiento del puntaje obtenido para cada uno de los días considerados en el conjunto testeo, tomando en cuenta como anómalos aquellos que superan el límite de 2.0972, marcadas con rojo aquellas en las que el método Bootstrap-LOCI etiqueto como anómalas y de verde caso contrario, adicionalmente si la línea se presenta entrecortada indica que la etiqueta es incorrecta, los resultados mostrados son validados con la matriz de confusión.

Figura 4.5: Resultados obtenido método LOCI-BOOTSTRAP

Elaboración: Autor

Capítulo 5

Conclusiones y recomendaciones

El método Bootstrap-LOCI como una modificación de su algoritmo predecesor, hereda las ventajas que presentaba el empleo del método LOCI. Una de ellas es la posibilidad de presentar la evolución en el tiempo del comportamiento del puntaje obtenido, como se mostró en la figura 4.7. En la práctica poder observar los resultados de forma ilustrativa es conveniente dado que aparte de identificar los días en los que hubo una anomalía, permite el análisis de patrones y a su vez facilita la toma de decisiones.

El poder establecer un límite basado en la distribución brinda de manera óptima un valor que permite etiquetar a las anomalías, lo que es evidenciado al momento de discriminar entre grupos, donde el algoritmo Bootstrap-LOCI presentó una mejoría al momento de detectar los días anómalos frente al algoritmo LOCI con los límites empleados en "Case study ofanomaly detection and quality control of energy efficiency and hygrothermalcomfort in buildings. "[8] cuyo uso fue justificado en la cantidad de anomalías que eran detectadas al usar cada uno de estos. Esta mejoría puede interpretarse como la detección de un día anómalo cada 10 días, es decir que pese a tener un aumento en la discriminación de los datos no posee evidencia suficiente para colegir que el límite hallado en el método Bootstrap-LOCI supera de manera sustancial a los resultados obtenidos al usar el límite de 2.5. Pese a esto debido a los recursos computacionales que requiere el empleo del algoritmo Bootstrap y el tiempo que puede tomar obtener los resultados, puede considerarse pertinente el uso del método LOCI tradicional en lugar del método Bootstrap-LOCI.

Si se desea emplear el método Bootstrap-LOCI se recomienda analizar la viabilidad de su aplicación considerando la cantidad de datos, el tiempo que se tiene para obtener los resultados y los medios que se poseen. Adicionalmente, para automati-

zar el uso del algoritmo mediante una aplicación web del tipo shiny u otra, se debe tener en cuenta las condiciones sobre las cuales cuales se ejecutará, por ejemplo considerando al rededor de 1500 observaciones y 12 gb de memoria ram en el servidor, un α de 0.75 y el radio variando entre el r_{min} y r_{max}, el método LOCI puede demorar un entre 4 segundos y 11 minutos, estos valores son calculados desde el momento en que se inicializa el algoritmo hasta que termina de ejecutarse, por lo tanto antes de realizar la aplicación se debe solicitar al proveedor información sobre el tiempo de ejecución permitido o la posibilidad de extenderlo, con la finalidad de que el programa no se suspenda al momento de ser utilizado mediante una pagina web.

Bibliografía

[1] Y. Abu-Mostafa, M. Magdon-Ismail, y Hsuan-Tien Lin. Learning from data. 2012.

[2] Charu C. Aggarwal. *Outlier Analysis*. Springer Publishing Company, Incorporated, 2017.

[3] Fatima Benjelloun, Ayoub Ait Lahcen, y Samir Belfkih. Outlier detection techniques for big data streams: focus on cyber security. *International Journal of Internet Technology and Secured Transactions*, 9:446, 01 2019.

[4] Rubén Fernández Casal y Julián Costa. *Aprendizaje estadístico*. GitHub, 2020.

[5] Jaime Cerda y Lorena Cifuentes. Uso de curvas ROC en investigación clínica: Aspectos teórico-prácticos. *Revista chilena de infectología*, 29:138 – 141, 04 2012.

[6] Bac Cong, Jorge Rivero Pérez, y Carlos Morell. Aprendizaje supervisado de funciones de distancia: estado del arte. *Revista Cubana de Ciencias Informáticas*, 9, 04 2015.

[7] Bradley Efron y Robert J. Tibshirani. *An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, Florida, USA, 1993.

[8] Carlos Eiras-Franco, Miguel Flores, Verónica Bolón-Canedo, Sonia Zaragoza, Rubén Fernández-Casal, Salvador Naya, y J. Tarrío-Saavedra. Case study of anomaly detection and quality control of energy efficiency and hygrothermal comfort in buildings. pgs. 145–151, 01 2019.

[9] Jim Freeman. Outliers in statistical data (3rd edition). *Journal of the Operational Research Society*, 46:28–52, 08 1995.

[10] Damaris Pascual González. Algoritmos de agrupamiento basados en densidad y validación de clusters. 2010.

[11] Valko M. Kveton B. Visweswaran S. Cooper G. F. Hauskrecht, M. Evidence-based anomaly detection in clinical domains. *Annual Symposium proceedings. AMIA Symposium*, pg. 319–323, 2007.

[12] Jeroen Janssens, Ildikó Flesch, y Eric Postma. Outlier detection with one-class classifiers from ml and kdd. pgs. 147–153, 12 2009.

[13] Edwin Knorr y Raymond Ng. Algorithms for mining distance-based outliers in large datasets. *VLDB*, 06 1998.

[14] Rokach L. y Maimon O. Clustering methods. in: Maimon o., rokach l. (eds) data mining and knowledge discovery handbook. *pringer, Boston, MA.*, pg. 446, 01 2005.

[15] Xinkie Lu, Tian Yang, Zaifei Liao, Manzoor Elahi, Wei Liu, y Hongan Wang. Incremental outlier detection in data streams using local correlation integral. *SAC '09: Proceedings of the 2009 ACM symposium on Applied Computing*, pg. 1520–1521, 03 2009.

[16] Jacob H. Madsen. Distance density-based outlier detection, 05 2018.

[17] S. Papadimitriou, H. Kitagawa, P. B. Gibbons, y C. Faloutsos. Loci: fast outlier detection using the local correlation integral. In *Proceedings 19th International Conference on Data Engineering (Cat. No.03CH37405)*, pgs. 315–326, 2003.

[18] S. Sawyer. Resampling data: Using a statistical jackknife. *Washington University*, 03 2005.

[19] Mark Ryan M. Talabis, Robert McPherson, I. Miyamoto, Jason L. Martin, y D. Kaye. Chapter 1 - analytics defined. In Mark Ryan M. Talabis, Robert McPherson, I. Miyamoto, Jason L. Martin, y D. Kaye, editors, *Information Security Analytics*, pgs. 1–12. Syngress, Boston, 2015.

[20] Cristina Urgiles y Martin Amoroso. Revisión de algoritmos para la detección de valores atípicos. *Killkana Técnica*, 2:19, 06 2018.

[21] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, y A. Anderla. Credit card fraud detection - machine learning methods. In *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*, pgs. 1–5, 2019.

[22] Stefan Windmann, Alexander Maier, Oliver Niggemann, Christian Frey, Ansgar Bernardi, Conny Gu, Holger Pfrommer, Thilo Steckel, Michael Krüger, y

Robert Kraus. Big data analysis of manufacturing processes. volume 659, pg. 012055, 11 2015.

[23] Ricardo Cao Abad y y Rubén Fernández Casal. Técnicas de remuestreo. pgs. 9–36, 11 2020.

Apéndice A

Descripción del etiquetado

FECHA	MARCA	OBSERVACION
11/9/2017	1	identificadas por mantenimiento
21/9/2017	1	identificadas por mantenimiento
22/9/2017	1	identificadas por mantenimiento
29/9/2017	1	identificadas por mantenimiento
16/10/2017	1	identificadas por mantenimiento
17/10/2017	1	identificadas por mantenimiento
18/10/2017	1	identificadas por mantenimiento
19/10/2017	1	identificadas por mantenimiento
20/10/2017	1	identificadas por mantenimiento
21/10/2017	1	identificadas por mantenimiento
22/10/2017	1	identificadas por mantenimiento
23/10/2017	1	identificadas por mantenimiento
24/10/2017	1	identificadas por mantenimiento
25/10/2017	1	identificadas por mantenimiento
26/10/2017	1	identificadas por mantenimiento
27/10/2017	1	identificadas por mantenimiento
28/10/2017	1	identificadas por mantenimiento
29/10/2017	1	identificadas por mantenimiento
30/10/2017	1	identificadas por mantenimiento
31/10/2017	1	identificadas por mantenimiento
17/11/2017	1	identificadas por mantenimiento
18/11/2017	1	identificadas por mantenimiento
19/11/2017	1	identificadas por mantenimiento
20/11/2017	1	identificadas por mantenimiento

Cuadro A.1: Etiquetas realizadas a los días a partir de las observaciones realizadas por mantenimiento.

FECHA	MARCA	OBSERVACION
30/9/2017	1	identificadas por comportamiento de la potencia de enfriamiento
1/10/2017	1	identificadas por comportamiento de la potencia de enfriamiento
2/10/2017	1	identificadas por comportamiento de la potencia de enfriamiento
8/11/2017	1	identificadas por comportamiento de la potencia de enfriamiento
12/11/2017	1	identificadas por comportamiento de la potencia de enfriamiento
24/11/2017	1	identificadas por comportamiento de la potencia de enfriamiento
3/12/2017	1	identificadas por comportamiento de la potencia de enfriamiento
9/12/2017	1	identificadas por comportamiento de la potencia de enfriamiento
17/12/2017	1	identificadas por comportamiento de la potencia de enfriamiento
18/12/2017	1	identificadas por comportamiento de la potencia de enfriamiento
25/12/2017	1	identificadas por comportamiento de la potencia de enfriamiento
1/1/2018	1	identificadas por comportamiento de la potencia de enfriamiento
2/1/2018	1	identificadas por comportamiento de la potencia de enfriamiento
21/1/2018	1	identificadas por comportamiento de la potencia de enfriamiento
23/1/2018	1	identificadas por comportamiento de la potencia de enfriamiento
24/1/2018	1	identificadas por comportamiento de la potencia de enfriamiento
28/2/2018	1	identificadas por comportamiento de la potencia de enfriamiento
7/3/2018	1	identificadas por comportamiento de la potencia de enfriamiento
8/3/2018	1	identificadas por comportamiento de la potencia de enfriamiento
9/3/2018	1	identificadas por comportamiento de la potencia de enfriamiento
10/3/2018	1	identificadas por comportamiento de la potencia de enfriamiento
11/3/2018	1	identificadas por comportamiento de la potencia de enfriamiento
12/3/2018	1	identificadas por comportamiento de la potencia de enfriamiento
13/3/2018	1	identificadas por comportamiento de la potencia de enfriamiento
30/3/2018	1	identificadas por comportamiento de la potencia de enfriamiento
31/3/2018	1	identificadas por comportamiento de la potencia de enfriamiento
14/8/2018	1	identificadas por comportamiento de la potencia de enfriamiento
15/8/2018	1	identificadas por comportamiento de la potencia de enfriamiento
18/8/2018	1	identificadas por comportamiento de la potencia de enfriamiento
19/8/2018	1	identificadas por comportamiento de la potencia de enfriamiento
10/9/2018	1	identificadas por comportamiento de la potencia de enfriamiento
11/9/2018	1	identificadas por comportamiento de la potencia de enfriamiento
19/9/2018	1	identificadas por comportamiento de la potencia de enfriamiento

Cuadro A.2: Etiquetas realizadas a los días a partir de las observaciones realizadas a la potencia de enfriamiento.

Apéndice B

Aplicativo Shiny

B.1. Imágenes

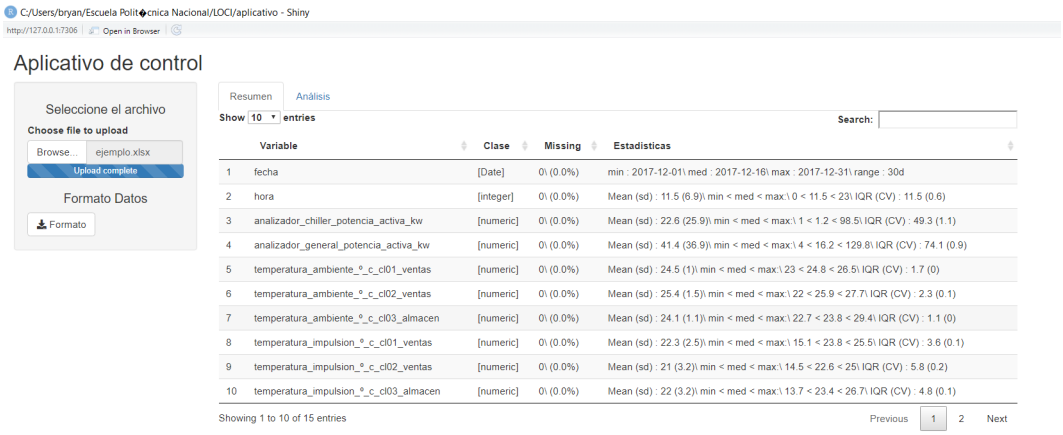

Figura B.1: Carga y resumen de datos

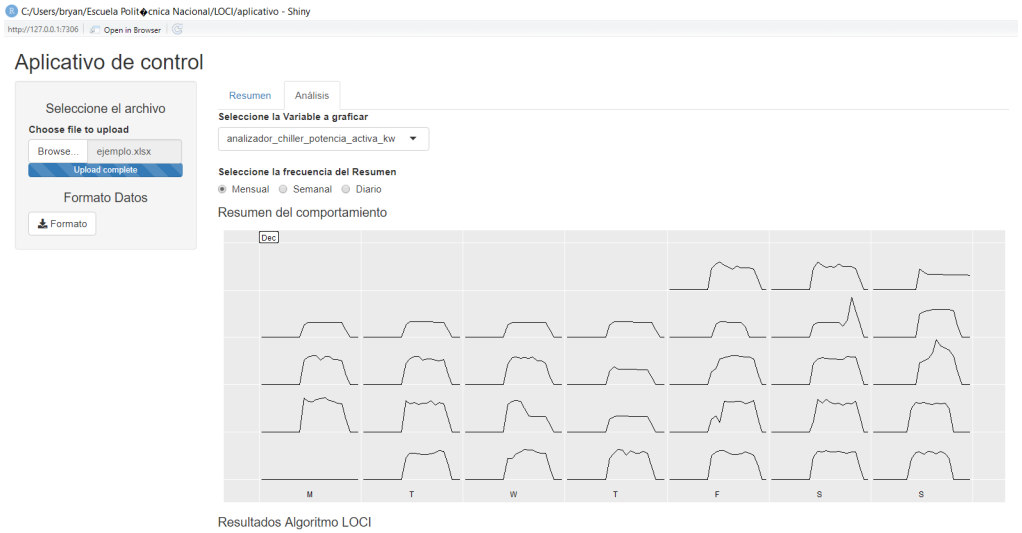

Figura B.2: Gráfica estilo calendario usando librerías ggplot2 y sugrrants

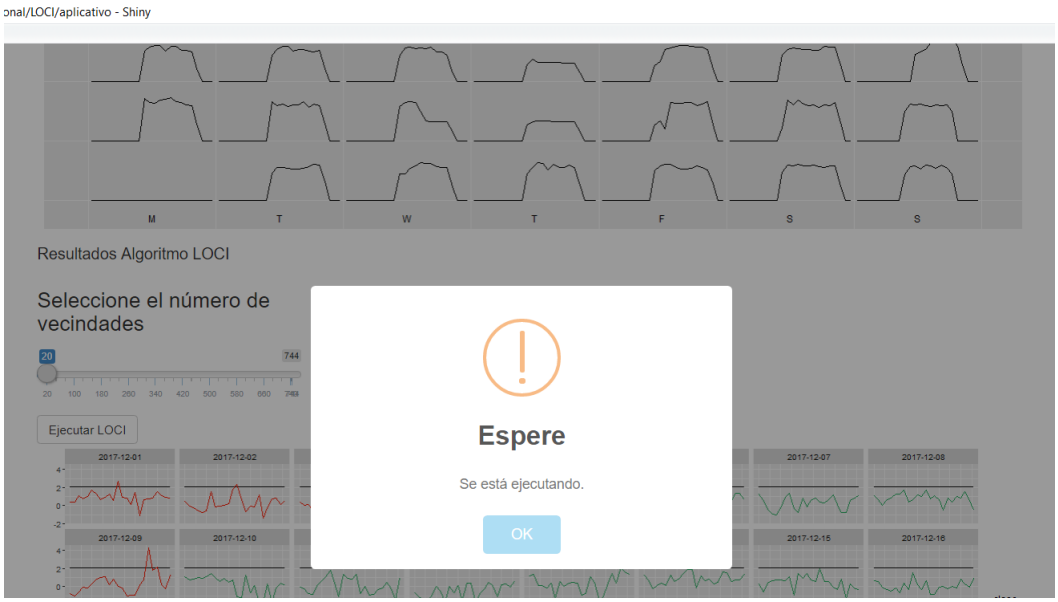

Figura B.3: Ejecución del algoritmo LOCI

Figura B.4: Gráfica de resultados del algoritmo LOCI empleando el límite Bootstrap-LOCI

B.2. Código

B.2.1. Global

```
library(readxl)
library(janitor)
library(openxlsx)
library(tidyverse)
library(DT)
library(summarytools)
library(lubridate)
library(sugrrants)
library(DDoutlier)
library(shinyalert)
```

B.2.2. UI

```
library(shiny)

# Define UI for application that draws a histogram
shinyUI(fluidPage(

    # Application title
    titlePanel("Aplicativo de control"),
```

```
    useShinyalert(),

    # Sidebar with a slider input for number of bins
    sidebarLayout(
        sidebarPanel(
            h4(align='center','Seleccione el archivo'),

            fileInput('file1', 'Choose file to upload',
                      accept = c(
                          'text/xlsx',
                          'text/Excel Microsoft Office Open XML Format Spreadsheet file',
                          '.xlsx'
                      )),
            h4(align='center','Formato Datos'),
            downloadButton("df", "Formato")
            ,width=2
        ),

        # Show a plot of the generated distribution
        mainPanel(
            tabsetPanel(
                tabPanel("Resumen", DTOutput('summa') ),
                tabPanel("An lisis", uiOutput('vari'),
                         radioButtons('frec','Seleccione la frecuencia del Resumen',
                                      c("Mensual"="monthly","Semanal"="weekly","Diario"="daily"),inline = T),
                         h4('Resumen del comportamiento'),
                         plotOutput('gcal'),
                         h4('Resultados Algoritmo LOCI'),
                         uiOutput('vecindad'),
                         actionButton('ej','Ejecutar LOCI'),
                         plotOutput('gloci'),
                         textOutput('tar'),
                         downloadButton('dr','Descargar Resultados'))
            )

        )
    )
))
```

B.2.3. Server

```
library(shiny)
shinyServer(function(input, output) {
    output$df <- downloadHandler(
        filename = function() {
            paste0("formato carga datos.xlsx")
        },
        content = function(file) {
            wb <- loadWorkbook('datos/formato carga datos.xlsx')
            owd <- setwd(tempdir())
            on.exit(setwd(owd))
            saveWorkbook(wb,file=file, overwrite = T)
        }
    )
    data <- reactive({

        inFile <- input$file1

        if (is.null(inFile))
            return(NULL)

        datos<-read_excel(inFile$datapath)
        datos<-datos %>% clean_names()
        datos[is.na(datos)]<-0
        datos %>%mutate(fecha=as.Date(fecha),hora=as.integer(hora))
    })
    resumen <-reactive({
        res<-dfSummary(data())
        res2<-res$Variable
        summa<-data.frame(matrix(unlist((strsplit(res2,split = '\\n'))
    ),ncol=2,byrow = T))
        names(summa)<-c("Variable","Clase")
        summa<-summa %>%mutate_if(is.factor,as.character)
        summa<-data.frame(summa %>% mutate(Variable=str_remove_all(
    Variable,"\\\\")),
                          Missing=res$Missing,Estadisticas=res$`Stats
    / Values`)
        summa

    })
    output$summa<-renderDT({
        req(data())
        resumen()
```

```
40      })
41      output$vari<-renderUI({
42          selectInput('var','Seleccione la Variable a graficar',names(
        data())[-c(1:2)])
43      })
44      dg<-reactive({
45          data() %>% select(fecha,hora,input$var)
46
47      })
48      g<-reactive({
49          req(data(),input$frec)
50          switch (which(names(data())[-c(1,2)]%in%input$var), dg() %>%
        frame_calendar(x = hora, y = vars(analizador_chiller_potencia_
        activa_kw), date =fecha, calendar = input$frec),
51                  dg() %>% frame_calendar(x = hora, y = vars(analizador_
        general_potencia_activa_kw), date =fecha, calendar = input$frec),
52                  dg() %>% frame_calendar(x = hora, y = vars(temperatura
        _ambiente_  _c_cl01_ventas), date =fecha, calendar = input$frec),
53                  dg() %>% frame_calendar(x = hora, y = vars(temperatura
        _ambiente_  _c_cl02_ventas), date =fecha, calendar = input$frec),
54                  dg() %>% frame_calendar(x = hora, y = vars(temperatura
        _ambiente_  _c_cl03_almacen), date =fecha, calendar = input$frec),
55                  dg() %>% frame_calendar(x = hora, y = vars(temperatura
        _impulsion_  _c_cl01_ventas), date =fecha, calendar = input$frec),
56                  dg() %>% frame_calendar(x = hora, y = vars(temperatura
        _impulsion_  _c_cl02_ventas), date =fecha, calendar = input$frec),
57                  dg() %>% frame_calendar(x = hora, y = vars(temperatura
        _impulsion_  _c_cl03_almacen), date =fecha, calendar = input$frec)
        ,
58                  dg() %>% frame_calendar(x = hora, y = vars(temperatura
        _retorno_  _c_agua_cl01_ventas), date =fecha, calendar = input$
        frec),
59                  dg() %>% frame_calendar(x = hora, y = vars(temperatura
        _retorno_  _c_agua_cl02_ventas), date =fecha, calendar = input$
        frec),
60                  dg() %>% frame_calendar(x = hora, y = vars(temperatura
        _retorno_  _c_agua_cl03_almacen), date =fecha, calendar = input$
        frec),
61                  dg() %>% frame_calendar(x = hora, y = vars(temperatura
        _  _c_entrada_agua_general), date =fecha, calendar = input$frec),
62                  dg() %>% frame_calendar(x = hora, y = vars(humedad_
        relativa_percent_ventas), date =fecha, calendar = input$frec))
63      })
64      output$gcal<-renderPlot({
```

```
65        req(data())
66        switch (which(names(data())[-c(1,2)]%in%input$var), p1<-g()
    %>% ggplot(aes(x = .hora, y = .analizador_chiller_potencia_activa_
    kw, group=fecha))+geom_line(),
67                  p1<-g() %>% ggplot(aes(x = .hora, y = .analizador_
    general_potencia_activa_kw, group=fecha))+geom_line(),
68                  p1<-g() %>% ggplot(aes(x = .hora, y = .temperatura_
    ambiente_  _c_cl01_ventas, group=fecha))+geom_line(),
69                  p1<-g() %>% ggplot(aes(x = .hora, y = .temperatura_
    ambiente_  _c_cl02_ventas, group=fecha))+geom_line(),
70                  p1<-g() %>% ggplot(aes(x = .hora, y = .temperatura_
    ambiente_  _c_cl03_almacen, group=fecha))+geom_line(),
71                  p1<-g() %>% ggplot(aes(x = .hora, y = .temperatura_
    impulsion_  _c_cl01_ventas, group=fecha))+geom_line(),
72                  p1<-g() %>% ggplot(aes(x = .hora, y = .temperatura_
    impulsion_  _c_cl02_ventas, group=fecha))+geom_line(),
73                  p1<-g() %>% ggplot(aes(x = .hora, y = .temperatura_
    impulsion_  _c_cl03_almacen, group=fecha))+geom_line(),
74                  p1<-g() %>% ggplot(aes(x = .hora, y = .temperatura_
    retorno_  _c_agua_cl01_ventas, group=fecha))+geom_line(),
75                  p1<-g() %>% ggplot(aes(x = .hora, y = .temperatura_
    retorno_  _c_agua_cl02_ventas, group=fecha))+geom_line(),
76                  p1<-g() %>% ggplot(aes(x = .hora, y = .temperatura_
    retorno_  _c_agua_cl03_almacen, group=fecha))+geom_line(),
77                  p1<-g() %>% ggplot(aes(x = .hora, y = .temperatura_  _
    c_entrada_agua_general, group=fecha))+geom_line(),
78                  p1<-g() %>% ggplot(aes(x = .hora, y = .humedad_
    relativa_percent_ventas, group=fecha))+geom_line())
79        prettify(p1)
80    })
81    output$vecindad<-renderUI({
82        req(data())
83        sliderInput("nn", label = h3("Seleccione el n mero de
    vecindades"), min = 20,
84                    max = nrow(data()), value = 20)
85    })
86    observeEvent(input$ej, {
87        # Show a modal when the button is pressed
88        shinyalert("Espere", "Se est  ejecutando.", type = "warning")
89    })
90    dloci<-eventReactive(input$ej,{
91        req(input$nn)
92        num<-data()[,-c(1,2)]
93        start_time <- Sys.time()
```

```
94          res<-LOCI(num,alpha = 0.75,nn = input$nn,k = 2.0972)
95          end_time <- Sys.time()
96          duracion<<-end_time - start_time
97          dataloci<-data()%>% mutate(fecha=as.Date(fecha))
98          graph<-data.frame(dataloci %>% select(fecha,hora),puntaje=res$
    MDEF/res$norm_MDEF,limite=2.0972)
99          graph<-graph %>% mutate(clase=ifelse(puntaje>=limite,1,0))
100         out<-graph  %>% filter(clase==1) %>% select(fecha,clase) %>%
    unique()
101         graph2<-full_join(graph %>% select(-clase),out,by='fecha')
102         graph2<-graph2%>% mutate(clase=as.factor(ifelse(is.na(clase)
    ,0,clase)))
103         down<<-full_join(graph %>%mutate(marca=clase) %>% select(-
    clase),graph2 %>%
104                              select(fecha,clase) %>% unique(),by='
    fecha')
105         graph2
106      })
107      output$gloci<-renderPlot({
108         req(dloci())
109
110         dloci() %>% mutate(fecha=as.factor(fecha))%>%
111             ggplot() +geom_line(aes(x=hora,y=limite))+
112             geom_line(aes(x=hora,y=puntaje,color=clase))+
113             scale_color_manual(values=c("#27AE60", "#FF1800"))+facet_
    wrap(~fecha,ncol=8)
114      })
115      output$dr <- downloadHandler(
116         filename = function() {
117             paste0("resumen etiquetas.xlsx")
118         },
119         content = function(file) {
120             owd <- setwd(tempdir())
121             on.exit(setwd(owd))
122             write.xlsx(down,file=file,asTable = T)
123         }
124      )
125      output$tar<-renderText({
126         req(req(dloci()))
127         paste('Tiempo de ejecuci n: ',round(duracion,2))
128      })
129 })
```

Apéndice C

Códigos

C.1. Establecer los parámetros y generar muestras Bootstrap

```
library(readxl)
data_tesis <- read_excel("data tesis.xlsx",  sheet = "HORAS INPUTADO")
##################################librer as
    ######################################
library(tsibble)
library(tsibbledata)
library(lubridate)
library(tidyverse)
library(viridis)
library(sugrrants)
library(janitor)
library(DDoutlier)
####################grafico de la evolucion de la potencia en el
    tiempo#################
data_tesis<-data_tesis %>% clean_names() %>% mutate(x_u_feff_date=date
    ) %>% select(-date)
Chiller <- data_tesis %>% select(x_u_feff_date,time,analizador_chiller
    _potencia_activa_kw) %>%
  mutate(Fecha=as.Date(x_u_feff_date)) %>% mutate(mes=month(Fecha)) #
    %>% filter(mes%in%c(9,10,11))
str(Chiller)
class(Chiller$x_u_feff_date)
Chiller_fc<- Chiller %>% frame_calendar(x = time, y = vars(analizador_
    chiller_potencia_activa_kw), date =Fecha, calendar = "monthly")
str(Chiller_fc)
```

```
p1 <- Chiller_fc%>%  ggplot(aes(x = .time, y = .analizador_chiller_
    potencia_activa_kw, group = Fecha)) +  geom_line()
p1
prettify(p1)

##########################################PARTICIONAR LA DATA
    ##########################################################
#0<-inlier
#1<-outlier
##################Quito la ultima fecha pues tiene datos incompletos
    ######################
data<-data_tesis %>% filter(x_u_feff_date!=max(data_tesis$x_u_feff_
    date))
###################################Selecciono unicamente las fechas
    ##################
fechas_et<- read_excel("data tesis.xlsx",  sheet = "FECHAS_ETIQUETA")
fechas_et<-fechas_et %>% clean_names() %>% mutate(x_u_feff_date=date)
    %>% select(-date)
fechas_et<-fechas_et %>% filter(x_u_feff_date!=max(fechas_et$x_u_feff_
    date))
#######################################
# Marca info modelamiento / validaci n (60% / 40%)

library(caTools)
library(ROSE)
library(data.table)
set.seed(12345)
muestra <- sample.split(setDF(fechas_et), SplitRatio=0.5)
setDT(fechas_et)[,.N,by=marca]
mod <- setDT(subset(setDF(fechas_et), muestra == TRUE))
setDT(mod)[,.N,by=marca]
val <- setDT(subset(setDF(fechas_et), muestra == FALSE))
setDT(mod)[,.N,by=marca]
#####################################################################
test<-data %>% filter(x_u_feff_date%in%val$x_u_feff_date)
test<-merge(test,val,by="x_u_feff_date")

##########trabajando con el train###################################
library(DDoutlier)
#########################Establecer Alpha############################
alpha<-seq(0.1,1,0.05)
##############Aplico el algoritmo loci para saber con que alpha
    trabajar#################################33
dtnumeric<-data %>% filter(x_u_feff_date%in%mod$x_u_feff_date)   %>%
```

```
    select(-x_u_feff_date,-time)
#clase<-c()
# for(k in 1:length(alpha)){
#   print(k)
#   clase<-cbind(clase,LOCI(dtnumeric,alpha[k])$class)
# }
# colnames(clase)<-paste0("alpha ",alpha)
# clase<-data.frame(clase)
#save(clase,file="clase.RData")
#sum(LOCI(dtnumeric,0.75,1000)$class=="Outlier")
load("clase.RData")
apply(clase,2,function(x)sum(x=="Outlier"))

##################################establecemos la cantidad de radios
    ########################
#el numero de vecinos de cada vecindad
radios<-round(seq(20,nrow(dtnumeric),length.out = 100),0)
score<- c()
for (i in 45:length(radios) ){
  print(i)
  a<- LOCI(dtnumeric,0.75,radios[i])
  score<-cbind(score,a$MDEF/a$norm_MDEF)
}
score<-data.frame(score)
names(score)<-paste0("r_",radios)
##########################################Selecciono el maximo valor
    del mdef###############
save(score,file="matriz_score_train.Rdata")
##############################################Selecciono el radio con
    mejor desempe o#############
clase<-ifelse(score[,1]>3,1,0)
for (i in 2:ncol(score)) {
  clase<-cbind(clase,ifelse(score[,i]>3,1,0))
}
clase<-data.frame(clase)
names(clase)<-paste0('r_',radios)
#####################################################################
dttrain<-data %>% filter(x_u_feff_date%in%mod$x_u_feff_date)   %>%
    select(x_u_feff_date,time)
clase<-cbind(dttrain,clase)
#####################################################################
anomalias<-list()
library(tidyverse)
for(i in 3:ncol(clase)){
```

```
  anomalias[[i-2]]<-clase[,c(1,i)] [clase[,c(i)]==1,] %>% unique()
}
#####################################################################
mod2<-list()
for(i in 1:length(anomalias)){
  mod2[[i]]<- mod %>% mutate(evaluacion=ifelse(x_u_feff_date%in%
    anomalias[[i]]$x_u_feff_date,1,0))
}
names(mod2)<-radios
# save.image(file="DatosCOmpletos.RData")
# load("DatosCOmpletos.RData")
#############################################################
#mc <- table(res, val$var_dep)

mc<-lapply(mod2, function(x)table(x$evaluacion,x$marca))
mc[[1]][1,1] # Verdaderos positivos
mc[[1]][2,2] # Verdaderos negativos
mc[[1]][1,2] # Falsos positivos
mc[[1]][2,1] # Falsos negativos
############################exactitud#################################
mc$`3217`
(mc$`3217`[1,1]+mc$`3217`[2,2])/sum(mc$`3217`)
############################Presicion
(mc$`3217`[1,1] /(mc$`3217`[1,1] +mc$`3217`[1,2]))
#######################################exactitud en general
    ################################
acc<-lapply(mc,function(x)(x[1,1]+x[2,2])/sum(x))
pres<-lapply(mc, function(x)(x[1,1] /(x[1,1] +x[1,2])))
acc<-unlist(acc)
pres<-unlist(pres)
which(acc%in%max(acc))
which(pres%in%max(pres))
mc[[62]]
mc[[63]]
acc[62]
acc[63]
pres[62]
pres[63]
radios[which(acc%in%max(acc))]
mc
#################################################escojemos el radio m s
    peque o y que mejor clasifica a los datos######
radel<-radios[which(acc%in%max(acc))][1]
#################################Eliminamos los outliers
```

```
    ###################################
inlier<-mod%>% filter(marca==0)
##########################data sin outliers
    ####################################3
datatrain<-data%>% filter(x_u_feff_date%in%inlier$x_u_feff_date)
###################################generamos los datos de
    entrenamiento bootstrap
    ######################################################################
# train_list<-list()
# for(i in 1:500){
#   mf<-inlier[sample(1:nrow(inlier),replace = T),]
#   train<-c()
#   for(j in 1:nrow(mf)){
#     train<-rbind(train,datatrain %>% filter(x_u_feff_date%in%mf$x_u_
    feff_date[j]))
#   }
#
#   train_list[[i]]<-train
# }
save(train_list,file="datosentrenamiento.RData",version = 2)
load("datosentrenamiento.RData")
#####obtengo unicamente las filas numericas de mi train list
train_list_num<-lapply(train_list,select,-x_u_feff_date,-time)
lloci<-lapply(train_list_num,LOCI,alpha=0.75,nn=radel,k=3)
library(DDoutlier)
system.time(LOCI(train_list_num[[1]],0.75,radel))
```

C.2. Ejecutar el algoritmo

```
library(DDoutlier)
library(tidyverse)
load("datosentrenamiento.RData")
select<-dplyr::select
# train_list_num<-lapply(train_list,select,-x_u_feff_date,-time)
# scores<-list()
# for(i in 451:length(train_list_num)){
#   print(i)
#   scores[[i]]<-LOCI(train_list_num[[i]],0.75,3217)
#   save(scores,file="./tesis/resultados.RData")
# }
##########################obtengo la cantidad de inliers
    ##########################################
```

```
cres<-lapply(scores, function(x)x[["class"]])
clasres<-unlist(cres)
length(clasres)-sum(clasres=="Inlier")
############################calculo el score
    ############################
lsce<-lapply(scores, function(x)x[["MDEF"]]/x[["norm_MDEF"]])
sce<-unlist(lsce)
sum(sce>3)
hist(sce,main = "")
#####################################################################
# dataloci<-cbind(dttrain,score$r_3217)
# save(dataloci,file="resultado/lociscore.RData")
########################Evaluamos el entrenamiento
    ############################
quantile(sce, c(.50, .75, .90,.99))
q_99<-cbind(dataloci,limite=quantile(sce, c(.99)) )
#########################################
names(q_99)<-c("fecha","time","score","limite")
#########################################################
q_99<-q_99 %>% mutate(clase=ifelse(score>limite,1,0))
#########################################################
anom_q99<-q_99[q_99$clase==1,] %>% select(fecha,clase) %>% unique()
#########################################################
q_99<-q_99 %>% mutate(clase_res=as.factor(ifelse(fecha%in%anom_q99$
    fecha,1,0)))
sum(as.numeric(q_99$clase_res))
#########################################################
library(ggplot2)
q_99$fecha<-as.factor(q_99$fecha)
##############GRAFICA DE LA EVOLUCI N EN EL TIEMPO DE LAS ANOMALIAS
ggplot(q_99) +geom_line(aes(x=time,y=limite))+
  geom_line(aes(x=time,y=score,color=clase_res))+
  scale_color_manual(values=c("#27AE60", "#FF1800"))+facet_wrap(~fecha
    ,ncol=20)
#################################
anom_q99<-anom_q99 %>% mutate(fecha=as.Date(as.character(fecha)))
mod<-mod %>% mutate(marca_2=ifelse(as.Date(x_u_feff_date)%in%anom_q99$
    fecha,1,0))
sum(mod$marca_2)
confusion<-table(mod$marca,mod$marca_2)
confusion

#Elemento de grafica para calcular el  rea  bajo la curva AUC
library(plotROC)
```

```
p1 <- ggplot(mod, aes(d=marca,m=marca_2)) + geom_roc()+ style_roc() #
   Grafica Curva ROC basica para colocar el area bajo la curva
#Curva ROC
logroc<-ggplot(mod, aes(d=marca,m=marca_2)) +
  theme_bw()+
  geom_roc(n.cuts = 0, colour="#3AA717") +
  theme(axis.text = element_text(colour = "black"),
        plot.title = element_text(hjust = 0.5))+
  ggtitle("Curva roc LOCI 1 mite 1.5")+
  scale_x_continuous("\n1 - especificidad (FPF)", breaks = seq(0, 1,
   by = .2))+
  scale_y_continuous("Sensibilidad (TPF)\n", breaks = seq(0, 1, by =
   .2)) +
  geom_abline(intercept=0, slope=1, colour="blue", linetype="dashed")
   +
  annotate("text", x=0.6, y=0.45, parse=TRUE,
           label=paste0("AUC: ",round(calc_auc(p1)$AUC,3)), colour="
   blue")+
  ggExtra::removeGridX()+
  ggExtra::removeGridY()

logroc
############################PRIMER LIMITE PAPER###################
q_99<-cbind(dataloci,limite=1.5)
#########################################
names(q_99)<-c("fecha","time","score","limite")
#########################################################
q_99<-q_99 %>% mutate(clase=ifelse(score>limite,1,0))
#########################################################
anom_q99<-q_99[q_99$clase==1,] %>% select(fecha,clase) %>% unique()
#########################################################
q_99<-q_99 %>% mutate(clase_res=as.factor(ifelse(fecha%in%anom_q99$
   fecha,1,0)))
sum(as.numeric(q_99$clase_res))
#######GRAFICA EVOLUCION EN EL TIEMPO DLE SCORE
library(ggplot2)
q_99$fecha<-as.factor(q_99$fecha)
# ggplot(q_99) +geom_line(aes(x=time,y=limite))+
#   geom_line(aes(x=time,y=score,color=clase_res))+
#   scale_color_manual(values=c("#27AE60", "#FF1800"))+facet_wrap(~
   fecha,ncol=20)
#################################
anom_q99<-anom_q99 %>% mutate(fecha=as.Date(as.character(fecha)))
mod<-mod %>% mutate(marca_2=ifelse(as.Date(x_u_feff_date)%in%anom_q99$
```

```
   fecha,1,0))
sum(mod$marca_2)
confusion2<-table(mod$marca,mod$marca_2)
confusion2

#Elemento de grafica para calcular el  rea  bajo la curva AUC
library(plotROC)
p1 <- ggplot(mod, aes(d=marca,m=marca_2)) + geom_roc()+ style_roc() #
   Grafica Curva ROC basica para colocar el area bajo la curva
#Curva ROC
logroc2<-ggplot(mod, aes(d=marca,m=marca_2)) +
  theme_bw()+
  geom_roc(n.cuts = 0, colour="#3AA717") +
  theme(axis.text = element_text(colour = "black"),
        plot.title = element_text(hjust = 0.5))+
  ggtitle("Curva roc LOCI l mite 1.5")+
  scale_x_continuous("\n1 - especificidad (FPF)", breaks = seq(0, 1,
   by = .2))+
  scale_y_continuous("Sensibilidad (TPF)\n", breaks = seq(0, 1, by =
   .2)) +
  geom_abline(intercept=0, slope=1, colour="blue", linetype="dashed")
   +
  annotate("text", x=0.6, y=0.45, parse=TRUE,
           label=paste0("AUC: ",round(calc_auc(p1)$AUC,3)), colour="
   blue")+
  ggExtra::removeGridX()+
  ggExtra::removeGridY()

logroc2

################################################
##############################Segundo limite del paper
   2.5#########################
q_99<-cbind(dataloci,limite=2.5)
#########################################
names(q_99)<-c("fecha","time","score","limite")
#########################################################
q_99<-q_99 %>% mutate(clase=ifelse(score>limite,1,0))
#########################################################
anom_q99<-q_99[q_99$clase==1,] %>% select(fecha,clase) %>% unique()
#########################################################
q_99<-q_99 %>% mutate(clase_res=as.factor(ifelse(fecha%in%anom_q99$
   fecha,1,0)))
sum(as.numeric(q_99$clase_res))
```

```
########################GRAFICA EVOLUCION EN EL TIEMPO DEL SCORE
library(ggplot2)
q_99$fecha<-as.factor(q_99$fecha)
# ggplot(q_99) +geom_line(aes(x=time,y=limite))+
#   geom_line(aes(x=time,y=score,color=clase_res))+
#   scale_color_manual(values=c("#27AE60", "#FF1800"))+facet_wrap(~
   fecha,ncol=20)
################################
anom_q99<-anom_q99 %>% mutate(fecha=as.Date(as.character(fecha)))
mod<-mod %>% mutate(marca_2=ifelse(as.Date(x_u_feff_date)%in%anom_q99$
   fecha,1,0))
sum(mod$marca_2)
confusion3<-table(mod$marca,mod$marca_2)
confusion3

#Elemento de grafica para calcular el  rea  bajo la curva AUC
library(plotROC)
p1 <- ggplot(mod, aes(d=marca,m=marca_2)) + geom_roc()+ style_roc() #
   Grafica Curva ROC basica para colocar el area bajo la curva
#Curva ROC
logroc3<-ggplot(mod, aes(d=marca,m=marca_2)) +
  theme_bw()+
  geom_roc(n.cuts = 0, colour="#3AA717") +
  theme(axis.text = element_text(colour = "black"),
        plot.title = element_text(hjust = 0.5))+
  ggtitle("Curva roc LOCI 1 mite 1.5")+
  scale_x_continuous("\n1 - especificidad (FPF)", breaks = seq(0, 1,
   by = .2))+
  scale_y_continuous("Sensibilidad (TPF)\n", breaks = seq(0, 1, by =
   .2)) +
  geom_abline(intercept=0, slope=1, colour="blue", linetype="dashed")
   +
  annotate("text", x=0.6, y=0.45, parse=TRUE,
           label=paste0("AUC: ",round(calc_auc(p1)$AUC,3)), colour="
   blue")+
  ggExtra::removeGridX()+
  ggExtra::removeGridY()

logroc3

library(ROCit)
ROC <- rocit(score=mod$marca_2, class=mod$marca)
# plot(ROC)
# ksplot(ROC)
```

```
##############################Loci entrenamiento
    ################################33
names(mod2$`3217`)
str(mod2$`3217`)
confusion2<-table(mod2$`3217`$marca,mod2$`3217`$evaluacion)
confusion2

########################Curva roc#########################
#Elemento de grafica para calcular el  rea  bajo la curva AUC
p2<- ggplot(mod2$`3217`, aes(d=marca,m=evaluacion)) + geom_roc()+
    style_roc() #Grafica Curva ROC basica para colocar el area bajo la
     curva
#Curva ROC
logroc2<-ggplot(mod2$`3217`, aes(d=marca,m=evaluacion)) +
  theme_bw()+
  geom_roc(n.cuts = 0, colour="#3AA717") +
  theme(axis.text = element_text(colour = "black"),
        plot.title = element_text(hjust = 0.5))+
  ggtitle("Curva roc LOCI tradicional")+
  scale_x_continuous("\n1 - especificidad (FPF)", breaks = seq(0, 1,
    by = .2))+
  scale_y_continuous("Sensibilidad (TPF)\n", breaks = seq(0, 1, by =
    .2)) +
  geom_abline(intercept=0, slope=1, colour="blue", linetype="dashed")
    +
  annotate("text", x=0.6, y=0.45, parse=TRUE,
           label=paste0("AUC: ",round(calc_auc(p2)$AUC,3)), colour="
    blue")+
  ggExtra::removeGridX()+
  ggExtra::removeGridY()

logroc2
#########################################testeo con loci bootstrap
    ##############################
sc_test<-restest$MDEF/restest$norm_MDEF
dat_test<-data.frame(fecha=test$x_u_feff_date,time=test$time,score=sc_
    test,limite=quantile(sce, c(.99)))

dat_test<-dat_test %>% mutate(evaluacion=ifelse(score>limite,1,0))
######################
anom_test<-dat_test[dat_test$evaluacion==1,] %>% select(fecha,
    evaluacion) %>% unique()
#################
dat_test<-dat_test %>% mutate(clase_res=as.factor(ifelse(fecha%in%anom
```

```
    _test$fecha,1,0)))
196 val<-val %>% mutate(tet=ifelse(marca==marca_2,"Bien etiquetado","Mal
    etiquetado"))
197 dat_test<-merge(dat_test,val,by.x="fecha",by.y="x_u_feff_date")
198 unique(dat_test %>% select(fecha,marca))
199 ############################Grafico testeo#################
200 dat_test$fecha<-as.factor(dat_test$fecha)
201 dat_test$tipo_etiquetado<-as.factor(dat_test$tet)
202 ########GRAFICA DE LA EVOLUCION EN EL TIEMPO DEL SCORE
203 ggplot(dat_test) +geom_line(aes(x=time,y=limite))+
204   geom_line(aes(x=time,y=score,color=clase_res, linetype = tipo_
    etiquetado))+
205   scale_color_manual(values=c("#27AE60", "#FF1800"))+facet_wrap(~fecha
    ,ncol=20)
206 ####################construccion de la matriz de confusion
207 val<-val %>% mutate(marca_2=ifelse(x_u_feff_date%in%anom_test$fecha
    ,1,0))
208 confusion<-table(val$marca,val$marca_2)
209 confusion
210 ######################################
211 p3<- ggplot(val, aes(d=marca,m=marca_2)) + geom_roc()+ style_roc() #
    Grafica Curva ROC basica para colocar el area bajo la curva
212 #Curva ROC
213 logroc<-ggplot(val, aes(d=marca,m=marca_2)) +
214   theme_bw()+
215   geom_roc(n.cuts = 0, colour="#3AA717") +
216   theme(axis.text = element_text(colour = "black"),
217         plot.title = element_text(hjust = 0.5))+
218   ggtitle("Curva roc LOCI limite 2.5")+
219   scale_x_continuous("\n1 - especificidad (FPF)", breaks = seq(0, 1,
    by = .2))+
220   scale_y_continuous("Sensibilidad (TPF)\n", breaks = seq(0, 1, by =
    .2)) +
221   geom_abline(intercept=0, slope=1, colour="blue", linetype="dashed")
    +
222   annotate("text", x=0.6, y=0.45, parse=TRUE,
223            label=paste0("AUC: ",round(calc_auc(p3)$AUC,3)), colour="
    blue")+
224   ggExtra::removeGridX()+
225   ggExtra::removeGridY()
226
227 logroc
228 ############################LOCI limite
    1.5#############################
```

```
229 ########################################testeo con loci bootstrap
    ##############################
230 sc_test<-restest$MDEF/restest$norm_MDEF
231 dat_test<-data.frame(fecha=test$x_u_feff_date,time=test$time,score=sc_
    test,limite=1.5)
232
233 dat_test<-dat_test %>% mutate(evaluacion=ifelse(score>limite,1,0))
234 ########################
235 anom_test<-dat_test[dat_test$evaluacion==1,] %>% select(fecha,
    evaluacion) %>% unique()
236 ######################
237 dat_test<-dat_test %>% mutate(clase_res=as.factor(ifelse(fecha%in%anom
    _test$fecha,1,0)))
238 val<-val %>% mutate(tet=ifelse(marca==marca_2,"Bien etiquetado","Mal
    etiquetado"))
239 dat_test<-merge(dat_test,val,by.x="fecha",by.y="x_u_feff_date")
240 unique(dat_test %>% select(fecha,marca))
241 ############################Grafico testeo###################
242 dat_test$fecha<-as.factor(dat_test$fecha)
243 dat_test$tipo_etiquetado<-as.factor(dat_test$tet)
244 ###########GRAFICA EVOLUCI N EN EL TIEMPO DEL SCORE
245 # ggplot(dat_test) +geom_line(aes(x=time,y=limite))+
246 #   geom_line(aes(x=time,y=score,color=clase_res, linetype = tipo_
    etiquetado))+
247 #   scale_color_manual(values=c("#27AE60", "#FF1800"))+facet_wrap(~
    fecha,ncol=20)
248 ####################contruccion de la matriz de confusion
249 val<-val %>% mutate(marca_2=ifelse(x_u_feff_date%in%anom_test$fecha
    ,1,0))
250 confusion2<-table(val$marca,val$marca_2)
251 confusion2
252 ##########################
253 p4<- ggplot(val, aes(d=marca,m=marca_2)) + geom_roc()+ style_roc() #
    Grafica Curva ROC basica para colocar el area bajo la curva
254 #Curva ROC
255 logroc2<-ggplot(val, aes(d=marca,m=marca_2)) +
256   theme_bw()+
257   geom_roc(n.cuts = 0, colour="#3AA717") +
258   theme(axis.text = element_text(colour = "black"),
259         plot.title = element_text(hjust = 0.5))+
260   ggtitle("Curva roc LOCI limite 2.5")+
261   scale_x_continuous("\n1 - especificidad (FPF)", breaks = seq(0, 1,
     by = .2))+
262   scale_y_continuous("Sensibilidad (TPF)\n", breaks = seq(0, 1, by =
```

```
    .2)) +
263   geom_abline(intercept=0, slope=1, colour="blue", linetype="dashed")
      +
264   annotate("text", x=0.6, y=0.45, parse=TRUE,
265            label=paste0("AUC: ",round(calc_auc(p4)$AUC,3)), colour="
      blue")+
266   ggExtra::removeGridX()+
267   ggExtra::removeGridY()
268
269 logroc2
270 ######################################################################
271 ###################################LOCI limite
      2.5#############################
272 ########################################testeo con loci bootstrap
      ##############################
273 sc_test<-restest$MDEF/restest$norm_MDEF
274 dat_test<-data.frame(fecha=test$x_u_feff_date,time=test$time,score=sc_
      test,limite=2.5)
275 #dat_test<-data.frame(fecha=test$x_u_feff_date,time=test$time,score=sc
      _test,limite=2.5)
276
277 dat_test<-dat_test %>% mutate(evaluacion=ifelse(score>limite,1,0))
278 ##############################
279 anom_test<-dat_test[dat_test$evaluacion==1,] %>% select(fecha,
      evaluacion) %>% unique()
280 ##########################
281 dat_test<-dat_test %>% mutate(clase_res=as.factor(ifelse(fecha%in%anom
      _test$fecha,1,0)))
282 val<-val %>% mutate(tet=ifelse(marca==marca_2,"Bien etiquetado","Mal
      etiquetado"))
283 dat_test<-merge(dat_test,val,by.x="fecha",by.y="x_u_feff_date")
284 unique(dat_test %>% select(fecha,marca))
285 ##########################Grafico testeo
      ################################
286 dat_test$fecha<-as.factor(dat_test$fecha)
287 dat_test$tipo_etiquetado<-as.factor(dat_test$tet)
288 #####GRAFICA DE EVOLUCION EN EL TIEMPO DEL SCORE
289 # ggplot(dat_test) +geom_line(aes(x=time,y=limite))+
290 #   geom_line(aes(x=time,y=score,color=clase_res, linetype = tipo_
      etiquetado))+
291 #   scale_color_manual(values=c("#27AE60", "#FF1800"))+facet_wrap(~
      fecha,ncol=20)
292 #############construccion matriz de confusion###############
293 val<-val %>% mutate(marca_2=ifelse(x_u_feff_date%in%anom_test$fecha
```

```
   ,1,0))
confusion3<-table(val$marca,val$marca_2)
confusion3
###################################
p5<- ggplot(val, aes(d=marca,m=marca_2)) + geom_roc()+ style_roc() #
   Grafica Curva ROC basica para colocar el area bajo la curva
#Curva ROC
logroc3<-ggplot(val, aes(d=marca,m=marca_2)) +
  theme_bw()+
  geom_roc(n.cuts = 0, colour="#3AA717") +
  theme(axis.text = element_text(colour = "black"),
        plot.title = element_text(hjust = 0.5))+
  ggtitle("Curva roc LOCI limite 2.5")+
  scale_x_continuous("\n1 - especificidad (FPF)", breaks = seq(0, 1,
   by = .2))+
  scale_y_continuous("Sensibilidad (TPF)\n", breaks = seq(0, 1, by =
   .2)) +
  geom_abline(intercept=0, slope=1, colour="blue", linetype="dashed")
   +
  annotate("text", x=0.6, y=0.45, parse=TRUE,
           label=paste0("AUC: ",round(calc_auc(p5)$AUC,3)), colour="
   blue")+
  ggExtra::removeGridX()+
  ggExtra::removeGridY()

logroc3
#########################################testeo loci
   ##############################################
dat_test_loci<-data.frame(fecha=test$x_u_feff_date,time=test$time,
   score=sc_test)
dat_test_loci<-dat_test_loci %>% mutate(eval=ifelse(score>3,1,0))
###############################
anom_test_loci<-dat_test_loci[dat_test_loci$eval==1,] %>% select(fecha
   ,eval) %>% unique()
#################
dat_test_loci<-dat_test_loci%>% mutate(clase_res=as.factor(ifelse(
   fecha%in%anom_test_loci$fecha,1,0)))
############################Grafico testeo
   ###################################
dat_test_loci$fecha<-as.factor(dat_test_loci$fecha)
ggplot(dat_test_loci) +geom_line(aes(x=time,y=3))+
  geom_line(aes(x=time,y=score,color=clase_res))+
  scale_color_manual(values=c("#27AE60", "#FF1800"))+facet_wrap(~fecha
   ,ncol=20)
```

```
val2<-val %>% mutate(marca_2=ifelse(x_u_feff_date%in%anom_test_loci$
    fecha,1,0))
sum(val2$marca_2)
confusion4<-table(val2$marca,val2$marca_2)
confusion4
############################exactitud#########################
(confusion4[1,1]+confusion4[2,2])/sum(confusion4)
############################Presicion
(confusion4[1,1] /(confusion4[1,1] +confusion4[1,2]))
###################################
p4<- ggplot(val2, aes(d=marca,m=marca_2)) + geom_roc()+ style_roc() #
    Grafica Curva ROC basica para colocar el area bajo la curva
#Curva ROC
logroc4<-ggplot(val2, aes(d=marca,m=marca_2)) +
  theme_bw()+
  geom_roc(n.cuts = 0, colour="#3AA717") +
  theme(axis.text = element_text(colour = "black"),
        plot.title = element_text(hjust = 0.5))+
  ggtitle("Curva roc metodo LOCI")+
  scale_x_continuous("\n1 - especificidad (FPF)", breaks = seq(0, 1,
    by = .2))+
  scale_y_continuous("Sensibilidad (TPF)\n", breaks = seq(0, 1, by =
    .2)) +
  geom_abline(intercept=0, slope=1, colour="blue", linetype="dashed")
    +
  annotate("text", x=0.6, y=0.45, parse=TRUE,
           label=paste0("AUC: ",round(calc_auc(p4)$AUC,3)), colour="
    blue")+
  ggExtra::removeGridX()+
  ggExtra::removeGridY()

logroc4
```

C.3. Evaluar los modelos

```
##########################Entrenamiento loci-bootstrap##########\
confusion
#################tasa de acierto en positivos, sensibilidad
    ###############
tpr<-confusion[1,1]/(confusion[1,1]+confusion[1,2])
tpr
#################tasa de acierto en negativos, especificidad
```

```
   ###############
tnr<-confusion[2,2]/(confusion[2,2]+confusion[2,1])
tnr
###############precision global o tasa de aciertos, accuracy
   ###############
#se usa en caso que las clases no se encuentren balanceadas
acc<-(confusion[1,1]+confusion[2,2])/sum(confusion)
acc
############################presicion balanceada#######################
ba<-(tpr+tnr)/2
ba
######################### entrenamiento Loci
   ##############################
confusion2
#################tasa de acierto en positivos, sensibilidad
   ###############
tpr2<-confusion2[1,1]/(confusion2[1,1]+confusion2[1,2])
tpr2
#################tasa de acierto en negativos, especificidad
   ###############
tnr2<-confusion2[2,2]/(confusion2[2,2]+confusion2[2,1])
tnr2
###############precision global o tasa de aciertos, accuracy
   ###############
#se usa en caso que las clases no se encuentren balanceadas
acc2<-(confusion2[1,1]+confusion2[2,2])/sum(confusion2)
acc2
############################presicion balanceada#######################
ba2<-(tpr2+tnr2)/2
ba2
############################Testeo loci-bootstrap#############
confusion3
#################tasa de acierto en positivos, sensibilidad
   ###############
tpr3<-confusion3[1,1]/(confusion3[1,1]+confusion3[1,2])
tpr3
#################tasa de acierto en negativos, especificidad
   ###############
tnr3<-confusion3[2,2]/(confusion3[2,2]+confusion3[2,1])
tnr3
###############precision global o tasa de aciertos, accuracy
   ###############
#se usa en caso que las clases no se encuentren balanceadas
acc3<-(confusion3[1,1]+confusion3[2,2])/sum(confusion3)
```

```
acc3
##########################presicion balanceada#######################
ba3<-(tpr3+tnr3)/2
ba3
##########################Testeo Loci############
confusion4
##################tasa de acierto en positivos, sensibilidad
   ###############
tpr4<-confusion4[1,1]/(confusion4[1,1]+confusion4[1,2])
tpr4
##################tasa de acierto en negativos, especificidad
   ###############
tnr4<-confusion4[2,2]/(confusion4[2,2]+confusion4[2,1])
tnr4
###############precision global o tasa de aciertos, accuracy
   ###############
#se usa en caso que las clases no se encuentren balanceadas
acc4<-(confusion4[1,1]+confusion4[2,2])/sum(confusion4)
acc4
##########################presicion balanceada#######################
ba4<-(tpr4+tnr4)/2
ba4
```

C.4. Graficar las matrices de confusión

```
mconfusion <-rbind(data.frame(data.frame(confusion),tipo="Bootstrap
   LOCI"),
                   data.frame(data.frame(confusion2),tipo="LOCI Lim
   =1.5"),
                   data.frame(data.frame(confusion3),tipo="LOCI Lim
   =2.5"))
names(mconfusion)<-c("predicho","real","valores","tipo")
mconfusion$real<-factor(mconfusion$real,levels = c(0,1))
mconfusion$tipo<-as.factor(mconfusion$tipo)
mconfusion$predicho<-factor(mconfusion$predicho,levels = c(1,0))
ggplot(data = mconfusion,
       mapping = aes(x = real,
                     y = predicho)) +
  geom_tile(aes(fill = valores)) +
  geom_text(aes(label = sprintf("%1.0f", valores)), vjust = 1) +
  scale_fill_gradient(low = "#1E8449",
                      high = "#FDFEFE")+theme_bw()+theme(legend.
```

```
  position = "none")+
facet_wrap(~tipo,ncol=3)
```