Trabajo Fin de Máster

Medición de incertidumbre en los modelos de detección de anomalías

Celia Melendi Ortega

Máster en Técnicas Estadísticas

Curso 2024-2025

Propuesta de Trabajo Fin de Máster

Título en galego: Medición da incerteza en modelos de detección de anomalías
Título en español: Medición de incertidumbre en modelos de detección de anomalías
English title: Uncertainty quantification in anomaly detection models
Modalidad: Modalidad B
Autor/a: Celia Melendi Ortega, Universidad de Santiago de Compostela
Director/a: Marta Sestelo Pérez, Universidade de Vigo
Tutor/a: Pablo Cereijo García, Gradiant
Breve resumen del trabajo: Este Trabajo Fin de Máster presenta un estudio de la aplicación de metodologías de medición de incertidumbre sobre algoritmos de detección de anomalías en contextos no supervisados, aplicados a un caso de seguridad industrial. Los métodos de detección de anomalías no están exentos de errores, por lo que medir la incertidumbre de estos modelos y de sus predicciones es crucial, especialmente en contextos críticos como es la ciberseguridad. La metodología es evaluada en dos escenarios, uno sintético para facilitar la representación gráfica de su funcionamiento, y otro más realista con datos de tráfico de red generados en un laboratorio.
Recomendaciones:
Otras observaciones:

Doña Marta Sestelo Pérez, Profesora Titular de la Universidade de Vigo, y don Pablo Cereijo García, Ingeniero Investigador de Gradiant, informan que el Trabajo Fin de Máster titulado

Medición de incertidumbre en los modelos de detección de anomalías

fue realizado bajo su dirección por doña Celia Melendi Ortega para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal. Además, doña Marta Sestelo Pérez y doña Celia Melendi Ortega

√ sí ☐ no

autorizan a la publicación de la memoria en el repositorio de acceso público asociado al Máster en Técnicas Estadísticas.

En Vigo, a 22 de Julio de 2025.

La directora:
Doña Marta Sestelo Pérez
SESTELO PEREZ MARTA - 76998300G
Firmado digitalmente por SESTELO PEREZ MARTA - 76998300G
Fecha: 2025.07.22 16:46:45 +02'00'

La autora:
Doña Celia Melendi Ortega
MELENDI ORTEGA, CELIA (AUTENTICACIÓN)
Firmado digitalmente por MELENDI ORTEGA, CELIA (AUTENTICACIÓN)
Fecha: 2025.07.22 17:12:33 +02'00'

El tutor:
Don Pablo Cereijo García
Firmado por CEREIJO GARCIA PABLO - ***6259** el día 22/07/2025

Declaración responsable. Para dar cumplimiento a la Ley 3/2022, de 24 de febrero, de convivencia universitaria, referente al plagio en el Trabajo Fin de Máster (Artículo 11, Disposición 2978 del BOE núm. 48 de 2022), **el/la autor/a declara** que el Trabajo Fin de Máster presentado es un documento original en el que se han tenido en cuenta las siguientes consideraciones relativas al uso de material de apoyo desarrollado por otros/as autores/as:

- Todas las fuentes usadas para la elaboración de este trabajo han sido citadas convenientemente (libros, artículos, apuntes de profesorado, páginas web, programas,...)
- Cualquier contenido copiado o traducido textualmente se ha puesto entre comillas, citando su procedencia.
- Se ha hecho constar explícitamente cuando un capítulo, sección, demostración,...sea una adaptación casi literal de alguna fuente existente.

Y, acepta que, si se demostrara lo contrario, se le apliquen las medidas disciplinarias que correspondan.

Agradecimientos

Quiero comenzar agradeciendo a mis padres y mi hermano, sin vuestro constante apoyo no hubiese podido llegar a donde estoy ahora. También al resto de mi familia, mi roca y lo más bonito que tengo. No podría faltar la mención especial a Inés, mi ejemplo a seguir todos estos años. A continuación, a mis tutores, Marta y Pablo. Muchísimas gracias por vuestro apoyo constante, las ganas de enseñarme y todo el tiempo que me habéis dedicado. Por último, a todo el equipo de Gradiant, principalmente al área tractora. Gracias por acogerme como una más desde el primer día y por ayudarme en todo lo que habéis podido.

Índice general

Capítulo 1

Introducción

El aprendizaje automático ha experimentado un crecimiento exponencial en los últimos años. Modelos de aprendizaje automático son utilizados en una gran cantidad de sectores, como la salud, las finanzas, la automatización industrial o la ciberseguridad. Estos modelos son capaces de aprender patrones a partir de grandes volúmenes de datos para realizar inferencia, hacer predicciones o tomar decisiones informadas, permitiendo mejoras en la eficiencia y la automatización de numerosos procesos. En particular, en el ámbito de la ciberseguridad, el aprendizaje automático es ampliamente utilizado para la detección de intrusiones en sistemas informáticos, entornos IT y OT o redes industriales. Algunos ejemplos de detección son el análisis de comportamientos anómalos en redes de comunicaciones, o la detección de correos electrónicos maliciosos.

A pesar de su eficacia, estos sistemas no están exentos de limitaciones. Tanto su proceso de aprendizaje como sus predicciones pueden verse afectadas por ruido en los datos, sesgos durante el entrenamiento o problemas de generalización debido a condiciones no observadas previamente. Todos estos factores comprometen la fiabilidad del modelo; en entornos críticos como la ciberseguridad de redes, sistemas u organizaciones, un error en la predicción del modelo puede tener consecuencias graves, por lo que resulta imprescindible cuantificar la incertidumbre asociada a las predicciones. Cuantificar esta incertidumbre, por ejemplo, mediante estimaciones probabilísticas o intervalos de confianza, permite evaluar la fiabilidad del modelo y de sus predicciones, y por tanto, tomar decisiones más informadas y robustas.

Este trabajo se enmarca en un caso de uso de ciberseguridad, concretamente en el ámbito de la detección de anomalías como mecanismo para identificar comportamientos inusuales que puedan estar relacionados con posibles amenazas o ataques. La creciente digitalización e interconexión de dispositivos ha incrementado tanto el volumen como la complejidad de los riesgos, además de la sofisticación de los ataques a los que se enfrentan las organizaciones y la industria. Es por tanto necesario desarrollar métodos que permitan identificar, prevenir y mitigar posibles amenazas de forma rápida y eficiente, pero que a la vez sean capaces de determinar el grado de confianza en sus predicciones, mejorando así la capacidad de toma de decisiones y respuesta ante incidentes de seguridad.

Para poder implementar estas metodologías es necesario conocer la estructura del proceso de detección de anomalías en sistemas de monitorización e identificar en qué fases deben incluirse. El flujo

de trabajo de un sistema de anomalías se desarrolla en las siguientes etapas:

1. Recopilación de datos: los sistemas recopilan una amplia gama de datos de diversas fuentes.

2. Procesamiento, normalización y enriquecimiento de datos: la información obtenida durante la recopilación debe limpiarse para asegurar la calidad de los datos (por ejemplo, evitando duplicados) y es necesario un procesado de estos para convertirlos a un formato consistente para los modelos de aprendizaje.

3. Extracción de características: es de interés identificar características relevantes que permitan diferenciar entre comportamientos normales y anómalos.

4. Creación de una línea base: mediante el uso de algoritmos de aprendizaje automático se analizan los datos recopilados y se establece la línea base de comportamiento normal.

5. Detección de anomalía: tras establecer una línea base de comportamiento normal los sistemas identifican las desviaciones o anomalías. La identificación se basa en la predicción del valor de anomalía y su comparación con respecto al valor umbral de anomalía de la línea base.

6. Alerta y puntuación del riesgo: para ayudar a priorizar las alertas, el sistema asigna una puntuación de riesgo a cada anomalía detectada basándose en la gravedad de la desviación. Así el sistema acaba de perfilar la alerta generada y se procede a la notificación.

En un contexto de detección de anomalías, trabajar con datos reales provenientes de entornos complejos y en tiempo real plantea importantes desafíos. Uno de los principales problemas es obtener una gran cantidad de falsos positivos, es decir, que el sistema clasifique erróneamente observaciones normales como anómalas. Este problema no implica únicamente un rendimiento subóptimo del algoritmo, sino que también genera una distribución ineficiente de recursos y desconfianza en el sistema, ya que conlleva invertir esfuerzos en investigar alertas que, en realidad, no representan amenazas. Habitualmente, los sistemas de detección de anomalías proporcionan para cada observación sobre la que se realiza inferencia una puntuación de anomalía, además de la clasificación binaria correspondiente. Esta puntuación se suele emplear como medida de criticidad o urgencia asociada a cada alerta. No obstante, esta información puede resultar insuficiente, especialmente en contextos críticos, por lo que es conveniente incorporar mecanismos que permitan valorar la fiabilidad de las clasificaciones.

Una posible solución es aplicar medición de incertidumbre en el proceso de detección de anomalías. Estas técnicas permiten cuantificar el grado de confianza en las predicciones obtenidas mediante métricas adicionales que complementan a la puntuación de anomalía. Conocer esta información permite priorizar el análisis de las predicciones que tienen una mayor confianza asociada, o descartar las más inciertas, consiguiendo una gestión más eficiente de los recursos disponibles. Por otra parte, la medición de incertidumbre también puede aportar información sobre la robustez del modelo y la incertidumbre asociada a este. Por ejemplo, que un alto porcentaje de las predicciones presente incertidumbre elevada podría indicar la necesidad de revisar el modelo, ajustar los parámetros empleados o modificar el conjunto de datos empleado para entrenar.

El presente trabajo analiza distintas técnicas de medición de incertidumbre y su aplicación práctica en un proyecto real, cuyo objetivo es la implementación de mecanismos de protección en la plataforma tecnológica de una entidad.

1.1. Objetivos

Los principales objetivos de este Trabajo Fin de Máster son los siguientes:

1. Realizar una revisión del estado del arte de las distintas técnicas de medición de incertidumbre en aprendizaje automático, centrándose en aquellas que están diseñadas para detección de anomalías basado en aprendizaje no supervisado.

2. Estudio y evaluación de diferentes algoritmos de detección de anomalías en un contexto no supervisado en el marco de un proyecto real en Gradiant, en concreto, realizar pruebas con algoritmos de aprendizaje automático basados en árboles de decisión, como son los modelos *Isolation Forest* y *Extended Isolation Forest*.

3. Aplicar los métodos de detección de anomalías ya entrenados sobre un conjunto de datos para realizar inferencia y clasificar en observaciones anómalas o normales.

4. Implementar las técnicas más relevantes de medición de incertidumbre (o comprender y adaptar implementaciones ya desarrolladas) y aplicarlas en el paso 6 del flujo de trabajo del sistema de detección de anomalías. Posteriormente, estudiar el efecto de estos métodos en los modelos y en las predicciones obtenidas.

5. Discutir las diferencias y similitudes de los distintos métodos aplicados.

1.2. Gradiant

Este trabajo se realiza en Gradiant (Fundación Centro Tecnológico de Telecomunicaciones de Galicia). Gradiant es una fundación privada sin ánimo de lucro, fundada en 2007 y con sede en Vigo (Galicia) que se centra en la transferencia de conocimiento y tecnología en los ámbitos de la conectividad, inteligencia y seguridad con el objetivo de mejorar la competitividad de las empresas. Cuenta con más de 200 trabajadores y 14 patentes solicitadas, además de haber desarrollado más de 800 proyectos de investigación y desarrollo. Se conforma a partir de un patronato que engloba a distintas organizaciones tanto públicas como privadas, formado por las Universidades de A Coruña, Santiago de Compostela y Vigo y las empresas Altia, Arteixo Telecom, Egatel, Indra, Plexus, R, Telefónica, Televés, y la Asociación empresarial INEO. Desde sus inicios, el compromiso del centro con la calidad es una constante. Gradiant cuenta con los siguientes certificados: Sistema de Gestión de Calidad UNE-EN ISO 9001:2015, Sistema de Gestión de Proyectos de I+D+i UNE 166002:2014, Sistemas de Gestión de la Seguridad de la Información UNE-EN ISO/IEC 7001:2013; y forma parte del registro estatal de Centros de Innovación Tecnológica (sello CIT).

Concretamente, este Trabajo Fin de Máster se desarrolla dentro del área de Seguridad y Privacidad, en la línea de *Security and Privacy Analytics*, dentro del proyecto GIC-TEL (Gestión de identidad y ciberseguridad en procesos telemáticos, Councilbox (2024)). El proyecto cuenta con dos líneas de investigación principales: la gestión de la identidad y la ciberseguridad. La línea de ciberseguridad está enfocada en la investigación y aplicación de tecnologías destinadas a implantar mecanismos de protección durante el procesamiento de datos, técnicas de detección de anomalías de ciberseguridad y métodos de desarrollo seguro de sistemas para la empresa Councilbox, específicamente para su aplicación OVAC (Oficina Virtual de Atención Ciudadana), una solución tecnológica diseñada para replicar

la atención presencial de administraciones públicas y empresas en entornos digitales. Esta ofrece servicios remotos con plena validez legal, garantizando la trazabilidad, la auditabilidad y el cumplimiento normativo.

1.3. Estructura del documento

El resto de la memoria se organiza como sigue: en el capítulo 2 se presenta el estado del arte en medición de incertidumbre, enfocado al contexto de detección de anomalías no supervisado. Asimismo, se presenta una revisión de la literatura relacionada con técnicas de detección de anomalías. En el capítulo 3 se desarrollan las bases teóricas de los métodos y algoritmos que serán aplicados posteriormente al análisis de los datos reales. Se realiza un estudio en profundidad de los algoritmos de detección de anomalías que serán empleados, así como de las técnicas de medición de incertidumbre aplicadas sobre dichos algoritmos. Los resultados obtenidos tras aplicar estas técnicas al conjunto de datos seleccionado se exponen en el capítulo 4, junto con una descripción detallada de dicho conjunto de datos. Además, se incluyen resultados para un conjunto de datos sintético sencillo, acompañado de representaciones gráficas. Por último, en el capítulo 5 se presentan las conclusiones del trabajo realizado y se proponen posibles líneas de trabajo futuro.

Capítulo 2

Estado del arte

En este capítulo se presenta la revisión del estado del arte llevada a cabo sobre distintas metodologías de detección de anomalías y de medición de incertidumbre en modelos de aprendizaje automático. Concretamente, esta revisión bibliográfica se ha centrado en las técnicas de medición de incertidumbre enfocadas a la detección de anomalías en contextos no supervisados.

2.1. Detección de anomalías

La detección de anomalías es el proceso de identificar patrones que no se corresponden con el comportamiento esperado en un conjunto de datos. Para ello se emplean técnicas de aprendizaje automático, que pueden ser supervisadas, no supervisadas o semisupervisadas. Esta clasificación depende de si los datos de entrenamiento están etiquetados o no, es decir, si se conoce la categoría a la que pertenecen (normal o anómalo) o no.

En contextos de aprendizaje supervisado, los algoritmos son entrenados empleando conjuntos de datos etiquetados, asociando cada dato de entrada con la correspondiente etiqueta de salida. Así, el algoritmo aprende a predecir la clase a la que pertenecen nuevos datos (Ortega-Fernandez, 2024). Algunos ejemplos de algoritmos de detección de anomalías basados en aprendizaje supervisado son *Support Vector Machines* (Cortes y Vapnik, 1995) o k-vecinos más cercanos (Cover y Hart, 1967).

Por otra parte, en problemas de aprendizaje no supervisado, los datos no están etiquetados, por lo que el modelo descubre de forma autónoma patrones o estructuras en los datos que permiten realizar inferencia sin etiquetas de entrenamiento que guíen el proceso de aprendizaje (Ortega-Fernandez, 2024). Existen multitud de métodos no supervisados de detección de anomalías, algunos de los cuales se exponen en la siguiente sección.

Por último, el aprendizaje semisupervisado es una combinación de los anteriores, en el que el modelo cuenta con una pequeña cantidad de datos etiquetados y un mayor volumen de datos sin etiquetar (van Engelen y Hoos, 2020). Algunos métodos de detección de anomalías semisupervisados son los *Support Vector Machines* semisupervisados (Bennett y Demiriz, 1998) o los *autoencoders* semisupervisados (Khaire y Kumar, 2022).

No obstante, la detección de valores anómalos constituye, en gran medida, un problema no supervisado, ya que habitualmente no se dispone de ejemplos de anomalías que permitan entrenar el mejor modelo para un conjunto de datos específico (Aggarwal, 2017). Por esta razón, el presente trabajo se centra en la aplicación de métodos no supervisados.

2.1.1. Detección de anomalías mediante aprendizaje no supervisado

A continuación, se describen algunos de los métodos más utilizados para la detección de anomalías mediante aprendizaje no supervisado.

Por una parte, existen algoritmos que se basan en el uso de densidades, utilizando tanto estimadores paramétricos como no paramétricos de la función de densidad de los datos, para después comparar cada punto con los de su entorno. Este es el caso de *Local Outlier Factor* (Breunig et al., 2000) o *Density-Based Spatial Clustering of Applications with Noise* (Ester et al., 1996). Entre estos enfoques se incluyen también los modelos de mezclas gaussianas (Dempster et al., 1977), que clasifican las observaciones como anómalas si se encuentran en zonas de baja densidad.

Otras metodologías están enfocadas a espacios de datos de alta dimensión, como es el caso de la detección de anomalías basada en Clasificadores de Componentes Principales (Shyu et al., 2003), que comprimen la dimensión de los datos para no sufrir problemas derivados de la alta dimensionalidad. Además, existen métodos que emplean la varianza entre un conjunto de puntos y sus vecinos más próximos para identificar las anomalías, como *Angle-Based Outlier Detection* (Kriegel et al., 2008). Uno de los métodos más empleados es *Isolation Forest* (Liu et al., 2008). Este modelo se basa en construir árboles de decisión que aíslan cada dato de la muestra, de forma que las anomalías, más susceptibles de ser separadas de los datos normales, sean separadas en nodos cercanos a la raíz de los árboles. Por otro lado, el *Extended Isolation Forest* (Hariri et al., 2021) es una generalización del *Isolation Forest*. Este método también emplea árboles de decisión, pero modificando los cortes que se realizan de forma que se evita la generación de sesgos.

Por último, dentro de los métodos de detección de anomalías basados en aprendizaje profundo, cabe destacar los *autoencoders* (Sakurada y Yairi, 2014). Estos modelos están formados por dos redes neuronales profundas, el codificador (*encoder*) y el decodificador (*decoder*), y una capa intermedia, el espacio latente, que se entrenan para aprender el comportamiento normal de los datos. Para llevar a cabo este proceso, se proyectan inicialmente las observaciones al espacio latente, de dimensión reducida, mediante el *encoder*, procurando conservar la mayor cantidad de información posible. A continuación, dichas representaciones latentes se reconstruyen en el espacio original mediante el *decoder*, con el objetivo de minimizar el error de reconstrucción. Dado que el modelo se entrena principalmente utilizando datos normales, las observaciones anómalas son reconstruidas de forma menos precisa, lo que da lugar a errores de reconstrucción más elevados.

2.2. Medición de incertidumbre

En esta sección se presentan las distintas metodologías existentes en la literatura para medir incertidumbre sobre modelos de aprendizaje automático. Las técnicas de aprendizaje automático establecen

relaciones entre variables observables y no observables, y la medición de incertidumbre caracteriza la variabilidad en esas relaciones (Stracuzzi et al., 2017). Las técnicas de medición de incertidumbre permiten aumentar la confianza en los modelos de aprendizaje automático, ya que aportan una métrica adicional que refleja el grado de certeza de las predicciones o proporciona garantías estadísticas sobre ellas. Además, esta métrica también puede ser un indicador de la adecuación del modelo y de los datos de entrenamiento.

Habitualmente, la incertidumbre se divide en dos categorías: incertidumbre aleatoria e incertidumbre epistémica. La incertidumbre aleatoria es inherente a los datos, debido a su aleatoriedad y estocasticidad, y se considera irreducible, ya que no disminuye al aumentar el tamaño muestral. En cambio, la incertidumbre epistémica se debe al modelo y a las imperfecciones que puedan producirse durante el proceso de entrenamiento de este, pero su reducción es posible al ampliar el conjunto de entrenamiento(Gawlikowski et al., 2022; Wenchong et al., 2024).

En la literatura existen numerosas técnicas de medición de incertidumbre para modelos de aprendizaje automático, siendo una de las más extendidas la predicción conformal (*conformal prediction*), una metodología relativamente novedosa para cuantificar la incertidumbre presente en las estimaciones obtenidas a partir de algoritmos de predicción, que además es agnóstica al modelo. La técnica fue introducida originalmente en Vovk et al. (2005), en el marco de la teoría de la predicción bajo condiciones de aleatoriedad. No obstante, con el avance del aprendizaje automático, ha experimentado un notable incremento en su popularidad y aplicación. Esta metodología permite transformar una noción heurística de incertidumbre que aporta la puntuación de anomalía de cualquier modelo en incertidumbre rigurosa (Angelopoulos y Bates, 2022). Se parte de un algoritmo de aprendizaje supervisado ya entrenado y de un conjunto de datos de calibración distinto al conjunto de entrenamiento. Ambos conjuntos de datos están formado por observaciones etiquetadas, es decir, cada observación pertenece a una entre varias clases posibles. El objetivo es construir para cada nueva observación un conjunto compuesto por distintas posibles clases, tal que, para un nivel de significación prefijado α, la verdadera clase de dicha observación pertenezca a ese conjunto con probabilidad $1 - \alpha$. Esta técnica es aplicable a cualquier modelo de aprendizaje automático, ya que es agnóstica al algoritmo empleado y refleja tanto la incertidumbre aleatoria como la epistémica (Wenchong et al., 2024).

Por otra parte, en Stracuzzi et al. (2017) se emplea un modelo de mezclas gaussianas junto a remuestras bootstrap para cuantificar la incertidumbre en un problema de clasificación de imágenes no supervisada. En Tyralis y Papacharalampous (2024) presentan una revisión de distintos métodos para capturar la incertidumbre que se propaga a las predicciones de los modelos de aprendizaje automático. Para capturarla, los modelos generan distribuciones de probabilidad en lugar de estimaciones puntuales.

La medición de incertidumbre resulta especialmente relevante en contextos de aprendizaje profundo, ya que la mayor parte de los modelos son de tipo "caja negra". Además, las redes neuronales tienden a ofrecer predicciones con un nivel de confianza excesivo y presentan limitaciones a la hora de identificar situaciones en las que su conocimiento es insuficiente o incierto. Por ello, existen multitud de metodologías destinadas a medir la incertidumbre en modelos de este tipo. Por una parte, las técnicas bayesianas se basan en considerar los parámetros del modelo como variables aleatorias, lo que permite incorporar la incertidumbre de manera explícita. En este enfoque, se utilizan modelos como las redes neuronales bayesianas y el aprendizaje profundo bayesiano. Estos métodos capturan la incer-

tidumbre epistémica y engloban métodos como la inferencia variacional (Graves, 2011), la técnica de Monte-Carlo *Dropout* (Gal y Ghahramani, 2016), las cadenas de Markov Monte-Carlo (Salakhutdinov y Mnih, 2008) o las aproximaciones de Laplace (MacKay, 2003). Por otra parte, existen otras técnicas muy extendidas para la estimación de incertidumbre en modelos de aprendizaje profundo, entre las que destacan los métodos basados en ensamblados de modelos (Lakshminarayanan et al., 2017). Estos producen una estimación basada en las predicciones obtenidas a partir de múltiples modelos. Además, permiten cuantificar la incertidumbre producida por el modelo midiendo la variabilidad que existe en las distintas estimaciones obtenidas.

No obstante, las metodologías mencionadas hasta el momento están enfocadas en aprendizaje supervisado, por lo que no son aplicables en el contexto de este proyecto, ya que no se dispone de una muestra de datos etiquetados para entrenar.Por lo tanto, son necesarios métodos de medición de incertidumbre para contextos de aprendizaje no supervisado.

2.3. Medición de incertidumbre en detección de anomalías

Tal como se ha expuesto en la sección anterior, existe una amplia literatura sobre técnicas de medición de la incertidumbre aplicadas a modelos de aprendizaje automático y aprendizaje profundo, especialmente enfocadas al aprendizaje supervisado y problemas multiclase. Sin embargo, la investigación en torno a modelos de detección de anomalías basados en aprendizaje no supervisado es mucho más limitada. Además, no se han identificado referencias específicas que aborden esta problemática en contextos concretos de ciberseguridad.

Si bien las técnicas de predicción conformal han sido tradicionalmente concebidas para problemas multiclase y aprendizaje supervisado, se han desarrollado diversas adaptaciones que permiten su aplicación en tareas de detección de anomalías en contextos no supervisados, como es el caso del presente trabajo. Estas también son agnósticas al modelo empleado y fueron introducidas en Laxhammar (2014), donde se presentan tanto la predicción de anomalías conformal como la predicción de anomalías inductiva conformal (también conocida como *split-conformal*), una adaptación del método anterior que es mucho más eficiente computacionalmente. Ambas técnicas consisten en calcular p-valores que permitan resolver el contraste de hipótesis

$$\begin{cases} H_0: & x \text{ es normal} \\ H_1: & x \text{ es una anomalía} \end{cases}$$

para cualquier nueva observación x, a partir de las puntuaciones de anomalía del algoritmo empleado. Así, es posible controlar el error tipo I mediante el nivel de significación α y con ello, la cantidad de observaciones normales que son clasificadas como anómalas, pudiendo limitar un problema importante en contextos de ciberseguridad. Por lo tanto, estos métodos calibran la incertidumbre del modelo con garantías estadísticas. No obstante, esta técnica presenta distintas limitaciones. Una de las más importantes es la necesidad de emplear un conjunto de calibración formado por observaciones normales, distintas a las de entrenamiento, por lo que se precisa de una gran cantidad de datos. Además, son sensibles a este conjunto de entrenamiento y pueden estar sujetos a sobreajuste.

A raíz de estos problemas, Hennhöfer y Preisach (2024) proponen métodos de detección de anomalías *cross-conformal* a partir de la adaptación de métodos de predicción *cross-conformal* para pro-

blemas multiclase. Estas metodologías también se basan en resolver el contraste de hipótesis anterior, pero sin la necesidad de un conjunto de calibración, ya que combinan resultados obtenidos a partir de distintas particiones del conjunto de entrenamiento, de forma similar a las técnicas de validación cruzada.

Por otra parte, en Bates et al. (2023) se estudian los p-valores obtenidos mediante las técnicas *split-conformal* desde una perspectiva de contrastes múltiples. Se analizan las dependencias presentes en los p-valores conformales y se estudia la validez de distintos métodos de corrección de contrastes múltiples, ya que algunos pueden no ser apropiados debido a estas dependencias. Adicionalmente, se propone una nueva metodología para generar otro tipo de p-valores conformales, los cuales satisfacen una garantía estadística más fuerte que la mencionada anteriormente.

Otra técnica de medición de incertidumbre en contextos de detección de anomalías es *ExCeeD* (Perini et al., 2021). El método se basa en disminuir la dependencia del conjunto de entrenamiento en el proceso de predicción, dado que a partir de este conjunto de datos se suele construir un umbral de decisión utilizado para clasificar nuevas observaciones como normales o anómalas. Para ello, se estima la probabilidad de que una observación nueva reciba la misma clasificación si se hubiera empleado un conjunto de entrenamiento distinto.

Las metodologías de medición de incertidumbre aplicadas a la detección de anomalías presentadas hasta el momento se caracterizan por ser agnósticas al modelo de detección de anomalías empleado, ya que únicamente requieren que dicho modelo proporcione una función de anomalía que asigne una puntuación a cada nueva observación sobre la que se desea realizar inferencia, de tal manera que observaciones anómalas obtengan puntuaciones más altas y normales más bajas (o viceversa). Sin embargo, también existen métodos de medición de incertidumbre específicos para modelos basados en *autoencoders*. Estos modelos de aprendizaje profundo han demostrado una mayor eficacia en tareas de detección de anomalías en entornos de alta dimensionalidad, en comparación con los métodos tradicionales de aprendizaje automático. Sin embargo, presentan una limitación importante: no informan sobre el nivel de confianza asociado a sus predicciones. A raíz de esta limitación, han surgido variantes como los *autoencoders* variacionales (Kingma y Welling, 2014) o los *autoencoders* bayesianos (Yong et al., 2020), que incorporan mecanismos probabilísticos para estimar la incertidumbre.

Los *autoencoders* variacionales son el método más empleado para medir la incertidumbre en este tipo de modelos (Saetta et al., 2024). En ellos, los parámetros del espacio latente se consideran variables aleatorias (habitualmente variables aleatorias i.i.d. con distribución normal), mientras que el resto de parámetros son deterministas y el modelo es entrenado empleando inferencia variacional. Los *autoencoders* bayesianos surgen al emplear un marco bayesiano, al igual que se realiza para las redes neuronales profundas, considerando distribuciones de probabilidad para todos los parámetros de la red. Por ejemplo, en Yong y Brintrup (2022) se presenta una formulación que permite capturar la incertidumbre de las anomalías, basada en convertir las estimaciones de la log-verosimilitud negativa que proporciona el *autoencoder* en incertidumbre.

Capítulo 3

Metodología

En este capítulo se presentan y desarrollan las técnicas que se emplearán a lo largo del presente trabajo, tanto para la detección de anomalías como para la estimación o medición de incertidumbre asociada a dichas detecciones.

Tal como se ha expuesto en el capítulo anterior, existen numerosas técnicas de detección de anomalías basadas en aprendizaje no supervisado. En diversos proyectos llevados a cabo por Gradiant se emplean tanto métodos de aprendizaje automático basados en árboles de decisión (*Isolation Forest* y *Extended Isolation Forest*) como modelos de aprendizaje profundo basados en redes neuronales profundas (*autoencoder*). En este trabajo se ha optado por emplear exclusivamente los algoritmos basados en árboles de decisión. Por un lado, esta elección responde al interés de la empresa por incorporar técnicas de medición de incertidumbre que permitan mejorar la fiabilidad de los modelos empleados. Por otro lado, pese a que los *autoencoders* han demostrado un rendimiento muy satisfactorio en multitud de escenarios, son modelos altamente complejos que conllevan un elevado coste computacional. Además, en el contexto específico del proyecto citado, el algoritmo *Isolation Forest* ha mostrado un desempeño limitado, lo que motiva la exploración de técnicas de medición de incertidumbre como posible vía de mejora.

3.1. Detección de anomalías

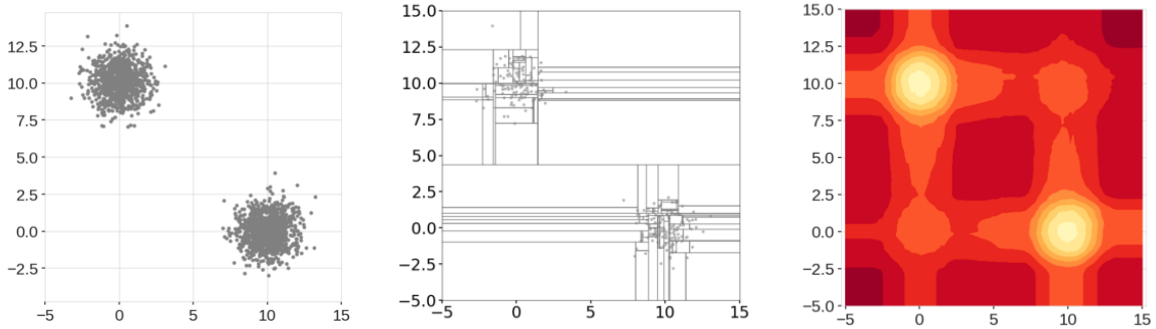
Las anomalías se describen como patrones de datos que presentan características distintas a las observaciones normales (Liu et al., 2008). A continuación, se describen en detalle los métodos de detección que que serán empleados en este trabajo.

3.1.1. *Isolation Forest*

El *Isolation Forest*, propuesto en Liu et al. (2008), es el modelo de aprendizaje automático más utilizado en problemas de detección de anomalías en contextos no supervisados. Este método puede manejar grandes conjuntos de datos y de alta dimensionalidad, y su capacidad para aislar rápidamente las anomalías lo hace útil en aplicaciones donde se necesita identificar observaciones atípicas en conjuntos de datos complejos y de gran escala.

El enfoque de este método se basa en la idea de que las anomalías son poco frecuentes y de que para ellas, las variables toman valores que difieren considerablemente de aquellos que toman para las observaciones normales, por lo que son más susceptibles de ser aisladas mediante la partición aleatoria e iterativa del dominio de los datos que las observaciones normales. Para ajustar este modelo se construye un conjunto de árboles de decisión aleatorios (*iTrees* o *Isolation Tree*) que permiten aislar cada observación a partir de un conjunto de entrenamiento $\mathcal{D}_{train} = \{x_1, \ldots, x_n : x_i \in \mathbb{R}^d\}$ formado por n observaciones de la variable aleatoria X, d-dimensional. De esta forma, en general, las observaciones normales serán aisladas en el final del árbol, mientras que las anomalías serán separadas más cerca del nodo raíz. El método requiere establecer dos parámetros: el número de árboles de decisión que se van a construir y el tamaño de los subconjuntos con los que se va a entrenar cada árbol (n_{obs}). Tras establecer ambos, se procede al ajuste del modelo, proceso que se divide en dos fases: una primera en la que se construyen los árboles de decisión binarios, cada uno a partir de un subconjunto del conjunto de entrenamiento $\mathcal{D}_{train}$, y una segunda en la que se obtienen puntuaciones de anomalía para la observación sobre la que se desea realizar inferencia a partir de la función de puntuación de anomalía, $\hat{s}(x, n_{obs})$ (observaciones normales tendrán puntuaciones de anomalía más bajas y observaciones anómalas más altas).

Cada árbol de decisión se construye de la siguiente manera: se parte de un subconjunto aleatorio del conjunto de datos de entrenamiento $D_k \subset \mathcal{D}_{train}$ formado por n_{obs} observaciones. Durante el proceso de ramificación del árbol se divide de forma recursiva el espacio de características, seleccionando de forma aleatoria en cada iteración una variable q y un punto de corte p, el cual se encuentra entre los valores máximo y mínimo de la variable q de los datos que se hallan en ese nodo. Entonces, aquellas observaciones para las que la variable q es menor que p, son enviadas a la rama de la izquierda, y en caso contrario, a la de la derecha. Esto se repite hasta que cada observación está aislada en un nodo terminal (hoja) o hasta que se alcanza el número máximo de cortes, si ha sido previamente establecido. Por lo tanto, cada nodo del árbol es, o un nodo terminal o un nodo interno con un test y dos nodos hijos. Este proceso se repite hasta construir los n_{trees} árboles de decisión.



(a) Datos normales de dos distribuciones gaussianas bivariantes que se emplean para entrenar el algoritmo.

(b) Cortes que realiza el algoritmo para las ramificaciones de un árbol de decisión.

(c) Mapa de calor de la función de puntuación de anomalía $\hat{s}$. Áreas más claras representan puntuaciones de anomalía más bajas.

Figura 3.1: Representación del funcionamiento del algoritmo *Isolation Forest*. Imágenes extraídas de Hariri et al. (2021).

Una vez se han construido los árboles de decisión, se obtiene la función de puntuación de anomalía $\hat{s}$, que se basa en la esperanza de la profundidad, $h(x)$, en la que se encuentra la instancia en los árboles de decisión, ($\mathbb{E}[h(x)]$), ya que se espera que las anomalías se encuentren en nodos menos profundos debido a que son más propensas a ser separadas en pocos pasos:

$$\hat{s}(x, n_{obs}) = 2^{-\mathbb{E}[h(x)]/c(n_{obs})}.$$

También se añade un factor de normalización $c(n_{obs})$, definido como

$$c(n_{obs}) = 2H(n_{obs} - 1) - 2\frac{n_{obs} - 1}{n_{obs}},$$

donde $H(i)$ es el número armónico (que se estima mediante $\ln(i) + \lambda$, con λ la constante de Euler-Mascheroni). Después se calcula la puntuación de anomalía asociada a cada observación x, $\hat{s}(x, n_{obs})$[1].

En la Figura 3.1 se muestra gráficamente el funcionamiento del método. Se parte del conjunto de puntos (que tienen un comportamiento normal) representado en la Figura 3.1a. Después, a partir de subconjuntos de estos, se crea cada árbol de decisión, que divide el plano de forma recursiva realizando cortes horizontales y verticales, como se observa en la Figura 3.1b. Por último, se asignan las puntuaciones de anomalía a las observaciones sobre las que se desean realizar predicciones, representadas en la Figura 3.1c.

No obstante, para poder realizar inferencia sobre instancias nuevas, clasificándolas como anómalas o normales, es necesario un umbral γ. Este se calcula a partir de las puntuaciones de anomalía del conjunto de entrenamiento y la proporción de observaciones anómalas que hay en este (contaminación), δ, y se corresponde con la estimación del cuantil $1 - \delta$ de las puntuaciones de anomalía del conjunto de entrenamiento. Si una observación x tiene una puntuación de anomalía mayor que γ, $\hat{s}(x) > \gamma$, x será clasificada como anómala. En caso contrario, será clasificada como normal. Otra posibilidad es no especificar ninguna proporción de contaminación y, en este caso, se emplea por defecto un umbral de 0,5.

Para su implementación se utiliza el software Python (Van Rossum y De Boer, 1991) y el paquete *scikit-learn* (Pedregosa et al., 2011), concretamente la clase/función *IsolationForest*.

3.1.2. *Extended Isolation Forest*

El método *Extended Isolation Forest* es una generalización del *Isolation Forest*. Propuesto en Hariri et al. (2021), surge para mitigar algunas de las limitaciones del método anterior. Una de las más importantes es la existencia de sesgos en las puntuaciones de anomalía, debido a cómo se realizan los cortes en la construcción de los árboles de ramificación. Al realizar los cortes teniendo en cuenta una única variable y un punto dentro del intervalo definido por los valores mínimo y máximo de dicha variable en esa rama, a medida que avanza la ramificación, el rango de posibles puntos de corte es cada vez más acotado, de forma que tiende a haber más cortes en donde hay una mayor concentración de observaciones. Por lo tanto, se crean regiones que obtienen puntuaciones de anomalía más bajas o más altas en función de si están alineadas con el conjunto de datos de entrenamiento para esa variable. Esto se ve reflejado de forma clara en las Figuras 3.1. Observando la Figura 3.1b y atendiendo

[1] A partir de este momento se referirá a la puntuación de anomalía de una observación como $\hat{s}(x)$ el lugar de $\hat{s}(x, n_{obs})$, ya que para un *Isolation Forest* dado, n_{obs} es constante.

a la variable representada en el eje vertical, los puntos anómalos que tengan un valor similar a los normales, serán objeto de un mayor número de cortes, y por lo tanto será muy posible que acaben en nodos terminales más lejanos de la raíz del árbol, frente a aquellos puntos que también sean anómalos pero que no estén alineados con el conjunto de entrenamiento. Efectivamente, en el mapa de calor expuesto en la Figura 3.1c se observa cómo hay regiones en las bandas horizontales y verticales de los clúster de datos normales que se corresponden con puntos anómalos pero que presentan puntuaciones de anomalía considerablemente más bajas. Si se consideran los puntos $(0, 0)$ y $(5, 5)$, es evidente que ambos son anómalos. No obstante, el punto $(0, 0)$ va a estar sujeto a un mayor número de cortes que el punto $(5, 5)$ y así obtienen puntuaciones de anomalía muy distintas, siendo menor la del $(0, 0)$. Por lo tanto, dependiendo del umbral γ, estas observaciones pueden ser clasificadas de forma distinta pese a ser ambas claramente anómalas.

Para solucionar este problema, se plantea una modificación en el proceso de ramificación de los árboles de decisión de forma que, en vez de realizar los cortes en una sola variable q, se empleen hiperplanos. Estos se definen a partir de un vector normal unitario $\vec{n}$ d-dimensional y un intercepto p, obtenido de una distribución uniforme d-dimensional cuyos extremos son los máximos y mínimos de cada variable en las observaciones que se encuentran en el nodo correspondiente. Entonces, el criterio empleado para realizar la ramificación es el siguiente: para una observación x, si se verifica

$$\overrightarrow{xp} \cdot \vec{n} \leq 0,$$

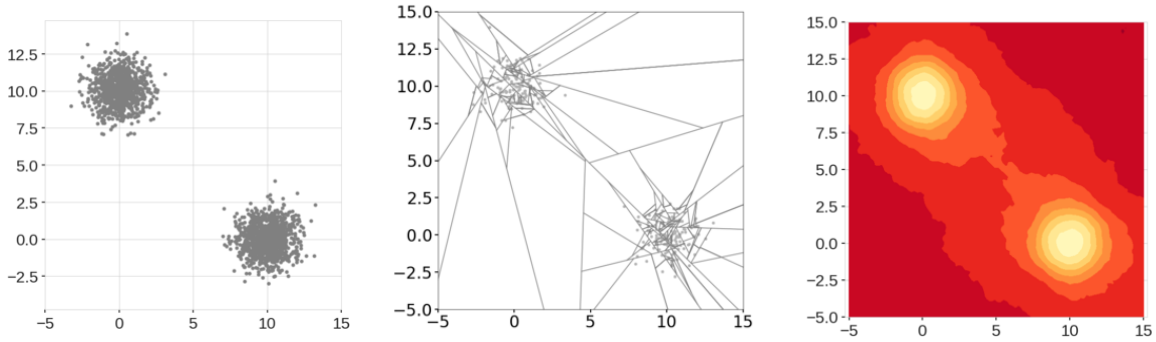
la observación se envía a la rama de la derecha y en caso contrario, a la de la izquierda. El vector $\overrightarrow{xp}$ es aquel que une los puntos x y p. En la Figura 3.2 se puede observar cómo esta modificación alivia las limitaciones del *Isolation Forest*. Se emplea el mismo conjunto de datos para entrenamiento, pero en este caso, como se aprecia en 3.2b, ya no existen regiones sujetas a un mayor número de cortes de forma artificial pese a que por construcción los interceptos p sí que se acumulen en las regiones con mayor concentración de datos. En la Figura 3.2c se representa el mapa de calor de las puntuaciones de anomalía, donde queda reflejado que este método no genera regiones "fantasma" con puntuaciones demasiado bajas.

El resto del método funciona de forma análoga al *Isolation Forest*: se establece el número de árboles de decisión que se van a construir, n_{trees}, y el tamaño de los subconjuntos con los que se va a entrenar cada uno, n_{obs}. A continuación, para construir cada árbol se realizan cortes de forma recursiva, hasta que se aíslan todas las observaciones empleadas para entrenamiento o hasta que se alcanza el máximo número de cortes, si ha sido establecido. Finalmente y de manera análoga, se calculan las puntuaciones de anomalía para nuevas observaciones y se realiza inferencia sobre estas a partir de un umbral γ (obtenido mediante un procedimiento idéntico).

Para la implementación del *Extended Isolation Forest* se ha empleado la librería `eif`, presentada en Hariri et al. (2021). Uno de los parámetros clave en el entrenamiento del modelo es la extensión, la cual determina la cantidad de variables involucradas en cada división del espacio. En otras palabras, este parámetro determina el número de variables que se ven afectadas por cada corte.

3.1.3. Métricas de evaluación

Una vez entrenado un modelo de aprendizaje automático, es imprescindible evaluar su calidad. Para ello, se emplea una muestra de datos etiquetados que no ha sido utilizada en el entrenamiento



(a) Datos normales de dos distribuciones gaussianas bivariantes que se emplean para entrenar el algoritmo.

(b) Cortes que realiza el algoritmo durante las ramificaciones de un árbol de decisión.

(c) Mapa de calor de la función de puntuación de anomalía $\hat{s}$. Áreas más claras representan puntuaciones de anomalía más bajas.

Figura 3.2: Representación del funcionamiento del algoritmo *Extended Isolation Forest*. Imágenes extraídas de Hariri et al. (2021).

del modelo y se mide la capacidad de este de generalizar sobre observaciones nuevas.

Para llevar a cabo dicha evaluación, se emplean diferentes métricas. En el contexto de la detección de anomalías, se lleva a cabo una clasificación binaria, en la que se distinguen dos clases: observaciones normales y observaciones anómalas. En consecuencia, las métricas de evaluación básicas serían las siguientes:

- Verdaderos negativos (VN): Observaciones normales que el modelo identifica correctamente como normales.
- Verdaderos positivos (VP): Observaciones anómalas que el modelo identifica correctamente como anómalas.
- Falsos negativos (FN): Observaciones anómalas que el modelo identifica incorrectamente como normales.
- Falsos positivos (FP): Observaciones normales que el modelo identifica incorrectamente como anómalas.

Es frecuente representar estas cantidades mediante una matriz de confusión, la cual permite visualizar el rendimiento de un algoritmo de clasificación. Dicha matriz está compuesta de cuatro cuadrantes, que contienen los verdaderos positivos, los falsos positivos, los verdaderos negativos y los falsos negativos.

A partir de las métricas anteriores, se pueden derivar otras métricas convenientes para evaluar el comportamiento del modelo:

- Exactitud (*accuracy*): proporción de predicciones correctas en relación con todas las predicciones realizadas

$$ACC = \frac{VP + VN}{VP + VN + FP + FN}.$$

- Precisión (*precision*): proporción de predicciones anómalas (positivas) que fueron correctas

$$PREC = \frac{VP}{VP + FP}.$$

- Exhaustividad o sensibilidad (*recall*): proporción de anomalías (reales) que fueron clasificadas correctamente

$$REC = \frac{VP}{VP + FN}.$$

- Especificidad (*specificity*): proporción de observaciones normales (reales) que fueron clasificadas correctamente

$$ESP = \frac{VN}{VN + FP}.$$

- Puntuación F1 (F1 *score*): es una medida que combina la precisión y la exhaustividad

$$F_1 = 2\frac{PREC \cdot REC}{PREC + REC} = \frac{2TP}{2TP + FP + FN}.$$

- AUC ROC (*Area Under the Receiver Operating Characteristic Curve*): métrica que cuantifica el rendimiento global de un modelo de clasificación binaria. Se define como el área bajo la curva ROC, la cual traza la tasa de verdaderos positivos (exhaustividad) frente a la tasa de falsos positivos en todos los umbrales de decisión. Un valor cercano a 1 indica que el modelo presenta una buena capacidad de discriminación.

Estas métricas serán utilizadas en el siguiente capítulo para comparar y seleccionar los modelos de detección de anomalías sobre los que se aplicará la medición de incertidumbre.

3.2. Medición de incertidumbre en detección de anomalías

Tras revisar el estado del arte y analizar las principales técnicas utilizadas en medición de incertidumbre aplicada a detección de anomalías, se realiza un estudio en profundidad de aquellas que se consideran más relevantes para este trabajo. Debido a que en el proyecto GIC-TEL se trabaja con modelos de detección de anomalías basados en aprendizaje no supervisado, concretamente *Isolation Forest*, *Extended Isolation Forest* y *Autoencoders*, se seleccionan métodos de medición de incertidumbre que se puedan aplicar sobre ellos. En concreto, se emplearán distintas técnicas de predicción conformal, específicamente *split-conformal* y *cross-conformal*, y el método *ExCeeD* (Perini et al., 2021). Adicionalmente, se ha desarrollado una implementación propia mediante el uso del *bootstrap* uniforme (Efron, 1979) para estimar la probabilidad de que una observación esté correctamente clasificada.

3.2.1. Detección de anomalías *split-conformal*

Tal como se introdujo en el capítulo anterior, esta técnica proporciona para cada nueva observación x, sobre la que se desea realizar inferencia, un p-valor que permite resolver el siguiente contraste de hipótesis

$$\begin{cases} H_0: & x \text{ es normal} \\ H_1: & x \text{ es anómala,} \end{cases}$$

aumentando así la fiabilidad del modelo.

Para su aplicación, se parte de un algoritmo de detección de anomalías, $\mathcal{A}$, y de un conjunto de datos normales $\mathcal{D} = \{x_1, \ldots, x_n : x_i \in \mathbb{R}^d\}$ formado por observaciones de la variable aleatoria X d-dimensional, que se divide en conjuntos de entrenamiento y calibración, $\mathcal{D}_{train}$ y $\mathcal{D}_{cal}$ respectivamente, tales que $\mathcal{D}_{train} \cap \mathcal{D}_{cal} = \emptyset$. A continuación, se ajusta el algoritmo $\mathcal{A}$ en $\mathcal{D}_{train}$ y se obtiene una función de puntuación de anomalía $\hat{s} : \mathbb{R}^d \longrightarrow \mathbb{R}$, que toma valores mayores para observaciones anómalas y menores para normales. Posteriormente, se obtiene el conjunto de las puntuaciones de calibración de $\hat{s}$, $\hat{s}(\mathcal{D}_{cal}) := \{\hat{s}(x_i) : x_i \in \mathcal{D}_{cal}\}$, que consiste en evaluar la función de puntuación de anomalía en el conjunto $\mathcal{D}_{cal}$.

Por consiguiente, para realizar inferencia sobre una nueva observación, se calcula su puntuación de anomalía $\hat{s}(x)$ y se obtiene el p-valor *split-conformal* relativo a la observación x de la siguiente manera:

$$p(x) = \frac{|\{i \in \mathcal{D}_{cal} : \hat{s}(x_i) \geq \hat{s}(x)\}| + 1}{|\mathcal{D}_{cal}| + 1}.$$

La hipótesis nula de que x es una observación normal es rechazada si existen evidencias significativas de que la observación es anómala, es decir, si el p-valor obtenido es menor que un nivel de significación α previamente fijado.

Otra forma análoga de proceder es, dado dicho nivel de significación α, calcular el cuantil empírico de orden $1-\alpha$, $Q_{1-\alpha}[\hat{s}(\mathcal{D}_{cal})]$. Si se verifica que $Q_{1-\alpha}[\hat{s}(\mathcal{D}_{cal})] \leq \hat{s}(x)$, se concluye que existen evidencias significativas de que x es una observación anómala y, por lo tanto, es clasificada como tal.

Por lo tanto, una observación será clasificada como anómala únicamente si existen evidencias significativas de que lo es, y en caso contrario, será clasificada como normal, proporcionando garantías estadísticas a la clasificación. Además, esta metodología permite limitar la cantidad de falsos positivos que obtiene el algoritmo de detección de anomalías, controlando el error tipo I en el contraste anterior mediante α para cada observación sobre la que se desea realizar inferencia, por lo que tiene un gran impacto en la detección de anomalías en contextos de ciberseguridad.

No obstante, los p-valores obtenidos dependen claramente del conjunto de calibración $\mathcal{D}_{cal}$, así como de las fluctuaciones aleatorias asociadas a su composición. Para intentar solventar este problema y, al mismo tiempo, eliminar la necesidad de reservar un subconjunto de calibración y poder emplear de forma más eficiente los datos disponibles, surgen las extensiones denominadas métodos *cross-conformal*.

3.2.2. Detección de anomalías *cross-conformal*

Para mitigar las limitaciones del método previamente descrito, en Hennhöfer y Preisach (2024) se proponen los métodos presentados en esta sección. Estos se corresponden con adaptaciones de procedimientos *cross-conformal* originalmente diseñados para realizar predicciones en contextos de aprendizaje supervisado, adaptados aquí para su aplicación en detección de anomalías en escenarios no supervisados.

En general, se parte de un conjunto de datos $\mathcal{D}$ compuesto por n observaciones normales de la variable aleatoria d-dimensional X y un algoritmo de detección de anomalías $\mathcal{A}$ y, de forma análoga a la detección de anomalías *split-conformal*, dada una observación nueva x, se calcula un p-valor a partir

del siguiente contraste de hipótesis

$$\begin{cases} H_0: \ x \text{ es normal} \\ H_1: \ x \text{ es anómala.} \end{cases}$$

A continuación, se describen las distintas metodologías:

Jackknife$_{\text{AD}}$

El método Jackknife$_{\text{AD}}$ se basa en el método de remuestreo Jackknife (Quenouille, 1956; Tukey, 1958). Consiste en calcular n funciones de puntuación de anomalías $\hat{s}_{-i}$, obteniendo cada una entrenando el algoritmo $\mathcal{A}$ con $\mathcal{D}\backslash\{x_i\}$, es decir, excluyendo el i-ésimo elemento de $\mathcal{D}$. Para obtener la puntuación de anomalía de las observaciones se emplea $\hat{s}$, ajustando $\mathcal{A}$ con el conjunto $\mathcal{D}$ completo. El p-valor relativo a la observación x se calcula de la siguiente forma:

$$p(x) = \frac{|\{x_i \in \mathcal{D} : \hat{s}_{-i}(x_i) \geq \hat{s}(x)\}| + 1}{|\mathcal{D}| + 1}.$$

Jackknife+$_{\text{AD}}$

Jackknife+$_{\text{AD}}$ consiste en una modificación del procedimiento anterior. La diferencia radica en que ahora, la puntuación de anomalía de las observaciones sobre las que se desea realizar inferencia no vendrá dada por $\hat{s}(x)$, sino por la mediana de las puntuaciones obtenidas al evaluar $\hat{s}_{-i}(x)$ para cada i,

$$\text{Median}(\{\hat{s}_{-i}(x)\}) = \text{Median}(\{\hat{s}_1(x), \hat{s}_2(x), \ldots, \hat{s}_n(x)\}).$$

El p-valor relativo a la observación x se calcula de la siguiente forma:

$$p(x) = \frac{|\{x_i \in \mathcal{D} : \hat{s}_{-i}(x_i) \geq \text{Median}(\{\hat{s}_{-i}(x)\})\}| + 1}{|\mathcal{D}| + 1}.$$

No obstante, tanto este método como el anterior se vuelven excesivamente costosos computacionalmente para conjuntos $\mathcal{D}$ muy grandes, ya que conllevan ejecutar el algoritmo $\mathcal{A}$ n veces. Para solventar este problema, se han propuesto dos metodologías alternativas, que se exponen a continuación.

CV$_{\text{AD}}$

Esta propuesta utiliza *k-fold cross validation* para obtener el conjunto de calibración. En este caso, el conjunto $\mathcal{D}$ se divide en k subconjuntos disjuntos, $S_1, \ldots, S_k$, de igual tamaño $m = \lfloor \frac{n}{k} \rfloor$. A continuación, para cada $j \in \{1, \ldots, k\}$ se obtiene la función de puntuación de anomalía $\hat{s}_{-S_j}$, entrenando el algoritmo mediante $\mathcal{D}\backslash S_j$, es decir, sin los elementos del conjunto S_j. Para obtener el conjunto de calibración de $\hat{s}$, los elementos de cada S_j serán evaluados mediante la función $\hat{s}_{-S_j}$ correspondiente. Así, el cálculo del p-valor relativo a la observación x se realiza como sigue:

$$p(x) = \frac{\left|\bigcup_{j=1}^{k} \{x_i \in S_j : \hat{s}_{-S_j}(x_i) \geq \hat{s}(x)\}\right| + 1}{|\bigcup_{j=1}^{k} S_j| + 1}.$$

CV+$_{\mathbf{AD}}$

De forma análoga a Jackknife+$_{\text{AD}}$, este método evalúa la puntuación de anomalía de cada observación x mediante la mediana de las puntuaciones obtenidas al evaluar $\hat{s}_{-S_j}(x)$ en cada $j \in \{1, \ldots, k\}$,

$$\text{Median}(\{\hat{s}_{-S_j}(x)\}) = \text{Median}(\{\hat{s}_{-S_1}(x), \hat{s}_{-S_2}(x), \ldots, \hat{s}_{-S_k}(x)\}).$$

Por lo tanto, el p-valor relativo a la observación x calculado mediante CV+$_{\text{AD}}$ viene dado por la siguiente fórmula:

$$p(x) = \frac{\left|\bigcup_{j=1}^{k}\{x_i \in S_j : \hat{s}_{-S_j}(x_i) \geq \text{Median}(\{\hat{s}_{-S_j}(x)\})\}\right| + 1}{|\bigcup_{j=1}^{k} S_j| + 1}.$$

En resumen, estos métodos ofrecen una solución eficaz frente a una importante limitación de la metodología *split-conformal*, al permitir el aprovechamiento completo del conjunto de datos disponible sin la necesidad de reservar una parte exclusiva para la calibración.

3.2.3. *ExCeeD*

En Perini et al. (2021) se propone un método que permite medir la fiabilidad de las predicciones obtenidas a partir de un detector de anomalías. El método se basa en la idea de que, habitualmente, al emplear modelos de detección de anomalías es necesario calcular un umbral γ a partir de las puntuaciones de anomalías de los datos de entrenamiento (como ocurre en *Isolation Forest* y *Extended Isolation Forest*). Por lo tanto, modificaciones en los datos de entrenamiento pueden producir distintos umbrales y como consecuencia, diferentes predicciones. Para estimar esta incertidumbre (que incluye tanto la incertidumbre epistémica como la aleatoria), se propone una medida de fiabilidad, $\mathcal{C}(\hat{y})_x$, que calcula la probabilidad de que, dada una nueva observación x, el modelo de detección de anomalías mantenga la misma predicción $\hat{y} \in \{0, 1\}$ (donde 1 indica anomalía y 0 indica normalidad) si se emplease un conjunto de entrenamiento distinto:

$$\mathcal{C}(\hat{y})_x = \begin{cases} \mathbb{P}(\hat{y} = 1|s, n, \delta, \hat{p}_s) & \text{si} \quad \hat{y} = 1 \\ 1 - \mathbb{P}(\hat{y} = 1|s, n, \delta, \hat{p}_s) & \text{si} \quad \hat{y} = 0, \end{cases}$$

donde s es la puntuación de anomalía de x proporcionada por el algoritmo de detección de anomalías, n el tamaño del conjunto de entrenamiento, δ la contaminación en este conjunto (proporción de datos anómalos) y $\hat{p}_s$ la probabilidad de que x sea anómala.

Sea X una variable aleatoria real d-dimensional, e Y una variable aleatoria que toma valores en $\{0, 1\}$. Sea $\mathcal{D}_{train}$ un conjunto de datos de tamaño n, formado por una muestra independiente e idénticamente distribuida de X y $\hat{s}(\mathcal{D}_{train}) = \{\hat{s}(x_i) : x_i \in \mathcal{D}_{train}\} = \{s_1, \ldots, s_n\}$. Dado que la función de puntuación de anomalías $\hat{s}$ se asume medible, $S = \hat{s}(X)$ es una variable aleatoria real. Así, se define la probabilidad de anomalía de x como sigue

$$\mathbb{P}(Y = 1|S = \hat{s}(x)) := \mathbb{P}(S \leq s).$$

Se considera la variable aleatoria condicional $Y|S = s$, y

$$P_s := \mathbb{P}((Y|S) = 1|s) = \mathbb{P}(Y = 1|S = s) = \mathbb{P}(S \leq s).$$

La variable aleatoria condicional $Y|S = s$ toma valores en el conjunto $\{0, 1\}$, por lo que sigue una distribución Bernoulli de parámetro P_s, $Y|S = s \sim Bern(P_s)$. Dado que no se conoce la distribución de S, se emplea un enfoque Bayesiano, asumiendo que P_s es una variable aleatoria que sigue una distribución uniforme en el intervalo $[0, 1]$. Así, se estima la probabilidad de anomalía a partir de la esperanza de P_s:

$$\hat{p}_s := \mathbb{E}[P_s] = \frac{1+t}{2+n},$$

donde $t = |\{i \in \{1 \dots n\} : s_i \leq s\}|$.

Ahora que $\hat{p}_s$ es conocido, es posible calcular $\mathcal{C}(\hat{Y})_x$. Existen dos escenarios dependiendo de la contaminación δ en el conjunto de entrenamiento $\mathcal{D}_{train}$ y ambos se basan en calcular la probabilidad de que la puntuación de anomalía de la observación sobre la que se desea realizar inferencia, s, sea menor que el umbral que se obtendría con un conjunto de datos distinto.

- $\delta \in (0, 1)$, es decir, existe contaminación. Entonces se utiliza como umbral el cuantil empírico $1 - \delta$ del conjunto de puntuaciones de anomalías de los datos de entrenamiento, $\hat{s}(\mathcal{D}_{train})$, que se corresponde con el elemento $s_{(\lfloor n-\delta n \rfloor + 1)}$ en $\hat{s}(\mathcal{D}_{train})$ ordenado.

  Por lo tanto, para que la observación sea clasificada de nuevo como anomalía tiene que haber en el nuevo conjunto de entrenamiento por lo menos $\lfloor n - \delta n \rfloor + 1$ datos con puntuación de anomalía menor que s. Esta probabilidad se calcula mediante una suma de probabilidades de una distribución binomial con n intentos y probabilidad $\hat{p}_s$:

$$P(\hat{Y} = 1 \mid s, n, \delta, \hat{p}_s) = \sum_{i=\lfloor n(1-\delta) \rfloor + 1}^{n} \binom{n}{i} \hat{p}_s^i (1 - \hat{p}_s)^{n-i}. \tag{3.1}$$

- $\delta = 0$ y, por lo tanto, únicamente hay datos normales en el conjunto de entrenamiento. Entonces se emplea como umbral la mayor puntuación de anomalía que se obtiene a partir del conjunto de entrenamiento $\mathcal{D}_{train}$, $\max(\hat{s}(\mathcal{D}_{train}))$.

  Ahora, para clasificar de nuevo x como anómala, todas las puntuaciones de anomalía del nuevo conjunto de entrenamiento deben ser menores que s (y así también lo será el nuevo umbral), por lo que

$$P(\hat{Y} = 1 | s, n, 0, \hat{p}_s) = \hat{p}_s^n. \tag{3.2}$$

3.2.4. Implementación *bootstrap*

Partiendo de la idea en la que se basa el método *ExCeeD* —modificaciones en los datos de entrenamiento podrían modificar las predicciones del modelo—, en este trabajo se ha desarrollado una implementación propia basada en el *bootstrap* uniforme (Efron, 1979) para obtener estimaciones de la probabilidad de que una observación esté correctamente clasificada empleando un enfoque frecuentista.

El método consiste en entrenar B modelos de detección de anomalías a partir de B muestras *bootstrap* del conjunto de datos original. A continuación, cada observación sobre la que se desea realizar inferencia es clasificada empleando cada uno de los B modelos, obteniendo por lo tanto B clasificaciones. Finalmente, se estima la probabilidad de que la observación sea anómala calculando la proporción de veces que la observación ha sido clasificada como tal. Esta metodología se clasificaría dentro de los métodos ensambladores, ya que combina los resultados de múltiples modelos para proporcionar las

estimaciones de incertidumbre.

Se comienza entrenando el modelo sobre el que se desea medir incertidumbre con un conjunto de datos de entrenamiento de tamaño n, $\mathcal{D}_{train} = \{x_1, \dots x_n\}$, completo, y clasificando como anómalas o no las observaciones sobre las cuales se desea realizar inferencia y obtener medidas de incertidumbre. Para ello se aplica el algoritmo *bootstrap* que se muestra a continuación:

1. Para cada $i = 1, \dots n$ arrojar $U_i \sim \mathcal{U}(0, 1)$ y hacer $x_i^* = x_{\lfloor nU_i \rfloor + 1}$.
2. Obtener $x^* = (x_1^*, \dots, x_n^*)$.
3. Entrenar el modelo de detección de anomalías y obtener la función de puntuación de anomalías $\hat{s}^*$ y el umbral γ^*.
4. Calcular la puntuación de anomalía de la observación x sobre la que se desea realizar inferencia, $\hat{s}^*(x)$ y clasificar en anomalía o no, $\hat{y}^* = I(\hat{s}^*(x) > \gamma^*)$.
5. Repetir B veces los pasos $1 - 4$ para obtener las réplicas *bootstrap* $\hat{y}^{*(1)}, \dots, \hat{y}^{*(B)}$.
6. Emplear las réplicas para estimar la probabilidad de que la observación x sea anómala:

$$\hat{p}_{anom}(x) = \frac{\sum_{j=1}^{B} \hat{y}^{*(j)}}{B}$$

Posteriormente, a partir de esta estimación, es posible obtener la probabilidad de que la observación x esté correctamente clasificada como $\hat{y}$,

$$\hat{p}_{correcto}(x) = \begin{cases} \hat{p}_{anom}(x) & \text{si } \hat{y} = 1 \\ 1 - \hat{p}_{anom}(x) & \text{si } \hat{y} = 0. \end{cases}$$

La obtención de probabilidades altas indica un bajo nivel de incertidumbre en el modelo y en sus predicciones, ya que, pese a modificar el conjunto de entrenamiento, se obtienen las mismas predicciones. En caso contrario, la obtención de probabilidades bajas indica una alta incertidumbre en el modelo y en las predicciones que genera, puesto que modificaciones en el conjunto de entrenamiento causan que el modelo clasifique de forma distinta una misma observación.

Al comparar este método con ExCeeD, se observa que la estimación de probabilidades mediante *bootstrap* implica un mayor coste computacional. Sin embargo, este resulta inferior al requerido por el cálculo de los p-valores mediante el método *Jackknife*, ya que para su obtención es necesario entrenar n modelos, frente a los B modelos que se emplean en el caso de la implementación *bootstrap*.

3.3. Contrastes múltiples

Tal como se ha explicado previamente, en el contexto de la predicción conformal, la incertidumbre se cuantifica mediante contrastes de hipótesis, lo que también permite controlar la tasa de falsos positivos. Sin embargo, dado que se realizan una gran cantidad de contrastes de forma simultánea —uno por cada nueva observación sobre la que se desea realizar inferencia—, es necesario adoptar una perspectiva de contrastes múltiples. Para ello, pueden emplearse procedimientos de test globales o estrategias que

permitan controlar la tasa de falsos descubrimientos (FDR, por sus siglas en inglés), entendida como el valor esperado de la proporción de observaciones normales entre aquellas clasificadas como anómalas.

Esta tasa se define como

$$FDR = \mathbb{E}\left(\frac{V}{\text{máx}\{R, 1\}} I(R > 0)\right),$$

donde V denota la variable aleatoria correspondiente al número de errores tipo 1 (falsos descubrimientos) y R representa el número total de hipótesis rechazadas.

No obstante, un aspecto fundamental en este contexto es la dependencia entre los p-valores obtenidos. En el caso de la predicción conformal, los p-valores obtenidos para las observaciones pertenecientes a un conjunto sobre el que se desea realizar inferencia no son independientes entre sí, ya que todos dependen, o bien del conjunto de calibración (predicción *split-conformal*) o del conjunto $\mathcal{D}$ (métodos *cross-conformal*). Esta dependencia debe tenerse en cuenta, ya que multitud de métodos clásicos de contrastes múltiples pueden no ser válidos en el caso de que los p-valores sean dependientes entre sí. En Bates et al. (2023) se estudia la validez de distintos métodos de contraste globales y de control de FDR bajo la dependencia que presentan los p-valores obtenidos mediante predicción *split-conformal*. Se concluye que el método de Benjamini–Hochberg (Benjamini y Hochberg, 1995) permite controlar la FDR en media, al igual que el método de Benjamini–Hochberg con la corrección de Storey (J. Storey, 2002). Estos métodos son robustos frente al tipo de dependencia que presentan los p-valores conformales y *cross-conformal* (Hennhöfer y Preisach, 2024), por lo que son herramientas adecuadas para emplear en este trabajo. A continuación se explican en detalle ambos procedimientos.

El test de Benjamini-Hochberg es un método que permite controlar la FDR al realizar multitud de contrastes de hipótesis, $H_1, \ldots, H_m$, cuyos respectivos p-valores son $p_1, \ldots, p_m$. El método se desarrolla de la siguiente manera:

1. Se ordenan los p-valores de menor a mayor, $p_{(1)} \leq \cdots \leq p_{(m)}$
2. Se fija el nivel $\alpha \in (0, 1)$ al que se quiere controlar la FDR.
3. Se obtiene $k = \text{máx}_i\{i : p_{(i)} \leq \frac{i}{m}\alpha\}$
4. Se rechazan todas las hipótesis nulas relativas a los p-valores $p_{(1)}, \ldots, p_{(k)}$.

Así, este método permite controlar la FDR a un nivel α. No obstante, si el porcentaje de hipótesis nulas es mucho menor que uno, este método es conservador.

Por otra parte, el método propuesto por Storey proporciona una técnica alternativa para controlar la FDR. Para su aplicación es necesario estimar la proporción de hipótesis nulas verdaderas, π_0, como

$$\hat{\pi}_0 = \frac{1 + \sum_{i=1}^{m} I(p_i > \lambda)}{m(1 - \lambda)},$$

donde λ es un parámetro de ajuste que puede estimarse mediante *bootstrap* (J. Storey, 2002) o mediante técnicas de suavizado (J. D. Storey y Tibshirani, 2003). A continuación, para controlar la FDR a un nivel α, se aplica el método de Benjamini-Hocherberg empleando como nivel $\alpha/\hat{\pi}_0$. Esta corrección presenta una potencia mayor al controlar las mismas tasas de error, ya que incorpora información sobre

la cantidad de hipótesis nulas verdaderas. En el presente trabajo, para obtener la estimación de π_0 se ha empleado la función `pval.estimate.eta0` del paquete `fdrtools` (Klaus y Strimmer, 2024) de R (R Core Team, 2021), estimando λ mediante el método basado en *bootstap*.

Capítulo 4

Resultados

En este capítulo se detallan los resultados obtenidos tras aplicar los métodos de detección de anomalías expuestos en el capítulo anterior a dos conjuntos de datos. En la sección 4.1, se incluyen resultados para un conjunto sencillo de datos sintéticos (con únicamente dos variables), con el objetivo de mostrar representaciones gráficas y facilitar el entendimiento los métodos utilizados. Posteriormente, en la sección 4.2, se presentan los resultados obtenidos al aplicarlos a un conjunto de datos de tráfico industrial, de mayor complejidad, con el fin de estudiar el rendimiento de las técnicas de medición de incertidumbre en un contexto real. En esta última sección se incluye una descripción detallada del conjunto de datos y recoge los resultados de aplicar las técnicas de medición de incertidumbre sobre los algoritmos *Isolation Forest* y *Extended Isolation Forest*.

4.1. Aplicación a datos simulados

En primer lugar, se realizan pruebas para un conjunto de datos sintético que consta únicamente de dos variables, con el objetivo de mostrar representaciones gráficas del efecto de los métodos de medición de incertidumbre. Este conjunto de datos es ampliamente empleado en aprendizaje automático, en contextos de clasificación binaria, y consiste en observaciones que forman dos lunas (semicírculos) no superpuestas en el plano. Los datos que pertenecen a las lunas se consideran normales y aquellos que no, anómalos. Se comienza generando un conjunto de entrenamiento, formado por 5000 observaciones normales, mediante la función `make_moons()` del paquete *sklearn* en Python. Asimismo, se genera un conjunto test para evaluar los algoritmos, formado por 370 observaciones anómalas y 370 observaciones normales. Las observaciones anómalas son generadas a partir de una distribución uniforme en $[-6, 10] \times [-6, 10]$, eliminando aquellas que caen sobre las dos lunas, y las observaciones normales son generadas mediante la función `make_moons()`.

Cabe mencionar que, dado que principalmente se desea ilustrar gráficamente el comportamiento de los métodos de detección de incertidumbre pero no estudiar los resultados obtenidos, y estos se aplican de igual manera en el *Isolation Forest* y en el *Extended Isolation Forest*, en esta sección únicamente se van a realizar pruebas para el primer algoritmo, ya que funciona de forma correcta y es más sencillo.

4.1.1. *Isolation Forest*

El proceso comienza con el entrenamiento de un modelo de *Isolation Forest* a partir del conjunto de datos de entrenamiento previamente generado, el cual se denominará modelo de referencia a lo largo del presente documento. La configuración de parámetros empleada para llevar a cabo el proceso de aprendizaje de este modelo es

- `n_estimators`: 50
- `max_samples`: `auto`
- `contamination`: 0.05.

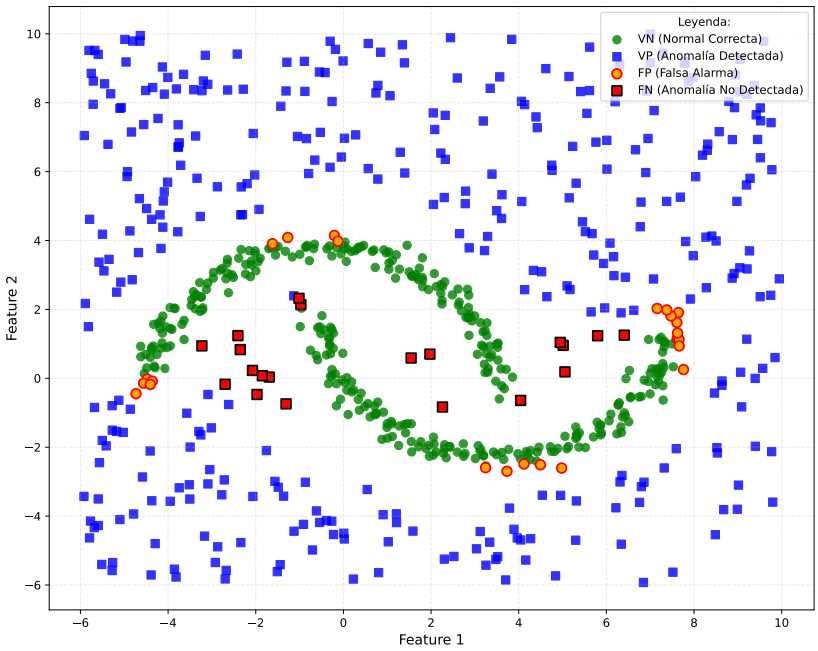


Figura 4.1: Clasificación en observaciones normales o anómalas del conjunto test de las dos lunas. Se muestra también si cada observación es un falso positivo, un falso negativo, un verdadero positivo o un verdadero negativo.

Tras evaluar el modelo mediante el conjunto test, se realiza una representación gráfica de este conjunto junto a la clasificación que proporciona el modelo para cada observación, la cual se muestra en la Figura 4.1. En dicha representación se aprecia que el modelo presenta dificultades para clasificar correctamente tanto las observaciones que se encuentran en el interior de las dos estructuras en forma de lunas, como aquellas que se encuentran en la periferia inmediata de estas. Adicionalmente, se calcula la matriz de confusión, expuesta en la Figura 4.2, y las métricas de evaluación correspondientes:

- Exactitud: 0,9405
- Precisión: 0,9358
- Exhaustividad: 0,9459

- Especificidad: 0,9351
- Puntuación F1: 0,9409
- AUC-ROC: 0,9405

Atendiendo a estas, se concluye que el modelo presenta un comportamiento satisfactorio, con 24 falsos positivos y 20 falsos negativos. Por lo tanto, se procede a aplicar sobre este las técnicas de medición de incertidumbre presentadas en el Capítulo 3.

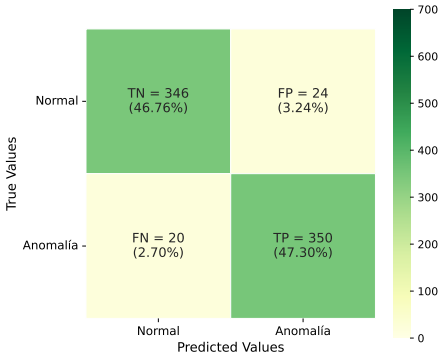


Figura 4.2: Matriz de confusión relativa a las predicciones obtenidas mediante el algoritmo *Isolation Forest*.

4.1.2. Medición de incertidumbre

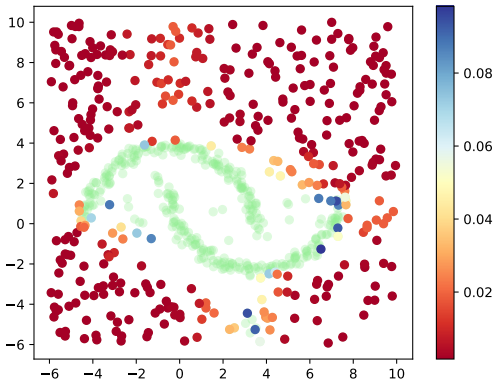
El análisis comienza con la aplicación de los métodos de predicción conformal, tanto el de *split-conformal* como los *cross-conformal*.

Split-conformal y *cross-conformal*

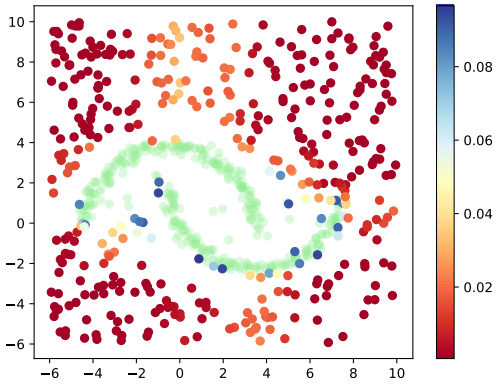
En el caso de la predicción *split-conformal* se emplea un conjunto de calibración formado por 1000 observaciones, por lo que se destinan las restantes 4000 para entrenar el modelo, utilizando la misma configuración de parámetros que para el modelo de referencia. Para los métodos *cross-conformal*, y conforme a lo expuesto previamente, se elimina la necesidad del conjunto de calibración, lo que permite emplear la totalidad de los 5000 datos disponibles para entrenamiento. Estos se dividen en 20 subconjuntos con el fin de llevar a cabo la validación cruzada, obteniendo así 20 modelos, excluyendo para cada uno de ellos 250 datos del conjunto de entrenamiento. Todos los *Isolation Forest* que son empleados durante el proceso también utilizan la misma configuración de parámetros que el *Isolation Forest* de referencia. Una vez completado el proceso de aprendizaje de los modelos necesarios para cada técnica, se calculan los p-valores correspondientes para realizar el contraste de hipótesis

$$\begin{cases} H_{0,i} : x_i \text{ es normal} \\ H_{1,i} : x_i \text{ es anómala,} \end{cases}$$

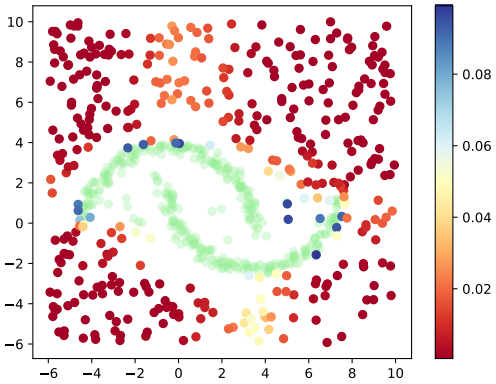
donde x_i representa la observación i-ésima del conjunto test. Tal y como se ha expuesto, es necesario tomar una perspectiva de test múltiples, de forma que se calculan los p-valores corregidos tras aplicar el método de Storey a partir de la función `p.adjust()` del paquete de R `stats`.



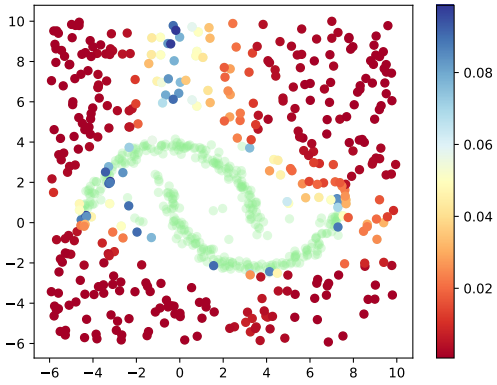
(a) p-valores corregidos obtenidos a partir del método de predicción *split-conformal*, tras aplicar el método de Storey.



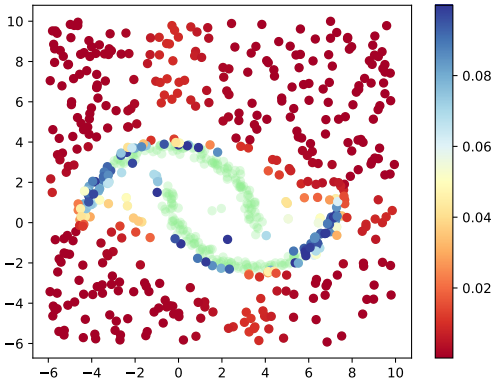
(b) p-valores corregidos obtenidos a partir del método de predicción *cross-conformal*, CV, tras aplicar el método de Storey.



(c) p-valores corregidos obtenidos a partir del método de predicción *cross-conformal*, CV+, tras aplicar el método de Storey.



(d) p-valores corregidos obtenidos a partir del método de predicción *cross-conformal*, Jackknife, tras aplicar el método de Storey.



(e) p-valores corregidos obtenidos a partir del método de predicción *cross-conformal*, Jackknife+, tras aplicar el método de Storey.

Figura 4.3: Representación gráfica de los p-valores obtenidos mediante los distintos métodos de medición de incertidumbre aplicados, tras aplicar el método de Storey. La escala de calor indica el valor que toman, con rojo oscuro los más bajos y azul oscuro los próximos a 0,1. En verde claro, translucido, se muestran aquellos puntos del conjunto test que no presentan evidencias significativas de ser anómalos, ya que sus p-valores son mayores de 0,1.

A continuación, se realizan representaciones gráficas en las que se visualiza el conjunto test en el plano, acompañado de los p-valores obtenidos mediante la metodología correspondiente. La escala de colores utilizada refleja los valores que adoptan dichos p-valores, tal y como se indica en la leyenda la Figura 4.3. En verde claro se muestran aquellos puntos cuyos p-valores son mayores que 0,1, es decir, aquellos puntos para los cuales no existen evidencias significativas de ser anómalos. El p-valor asociado al resto de puntos está representado mediante la escala de color, donde el color rojo oscuro indica la existencia de fuertes evidencias significativas de que la observación es anómala, mientras que, a medida que el color cambia de forma gradual hacia el azul oscuro, las evidencias de que la observación se trata de una anomalía disminuyen. Por lo tanto, estos p-valores cuantifican la incertidumbre asociada a las observaciones, ya que los p-valores muy bajos indican que apenas existe incertidumbre sobre si esa observación es anómala o no, mientras que p-valores en torno a 0,1 señalan que sí existe incertidumbre.

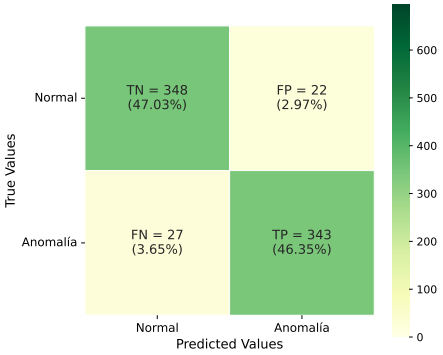
En general, en las gráficas se puede apreciar cómo, en las regiones alineadas horizontal y verticalmente con las nubes de puntos de las lunas, los p-valores son menores, incluso en zonas alejadas de ellas, especialmente para el método Jackknife. Esto es debido a los sesgos que presenta el *Isolation Forest*, descritos en el Capítulo 3. Por otra parte, coincidiendo con lo esperado, los puntos situados en regiones más alejadas de las dos lunas presentan p-valores más bajos, con un color rojo oscuro, mientras que a medida que se aproximan a las lunas, comienzan a ser más amarillos y azules, indicando menores evidencias significativas de que se tratan de anomalías y, por lo tanto, mayor incertidumbre asociada.

Partiendo de los p-valores previamente calculados se procede a realizar una clasificación de las observaciones del conjunto test, de forma que se consideran como anómalas únicamente aquellas para las que existen evidencias estadísticas significativas de que lo son. Como se ha detallado, se adopta una perspectiva de tests múltiples aplicando el método de Storey para controlar la tasa de descubrimientos falsos (FDR), y se controla a un nivel de 0,10. Tras llevar esto a cabo, se calculan las matrices de confusión resultantes al rechazar las hipótesis nulas correspondientes. Estas se muestran en la Figura 4.4, que contiene los resultados para los métodos de medición de incertidumbre *split-conformal* (4.4a), CV_{AD} (4.4b), $CV_{AD}+$ (4.4c), Jackknife$_{AD}$ (4.4d) y Jackknife$_{AD}+$ (4.4e). También se calculan las mismas métricas que fueron previamente calculadas para modelo de referencia, las cuales se muestran en la Tabla 4.1.

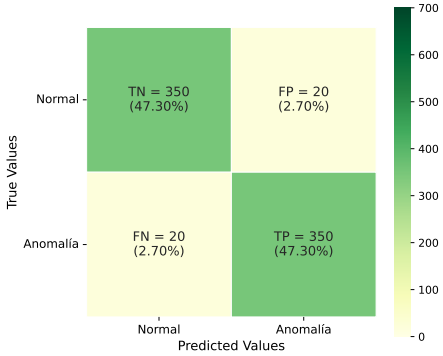
Al analizar tanto las matrices de confusión como las métricas calculadas, se aprecia que la clasificación que obtienen los métodos a partir de los p-valores es comparable a la obtenida mediante el modelo de referencia, siendo el método *cross-conformal* Jackknife$+_{AD}$ el que presenta una diferenciación mayor. Atendiendo a las métricas, en general, estas nuevas clasificaciones no deterioran el desempeño del modelo y, adicionalmente, cuentan con garantías estadísticas.

ExCeeD

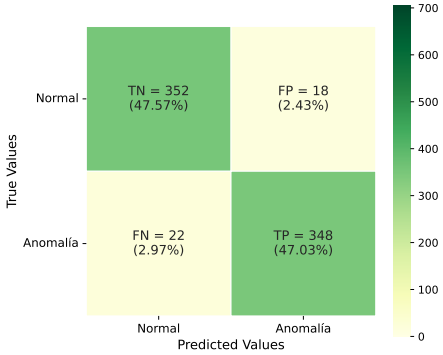
El método *ExCeeD* parte del modelo de referencia previamente entrenado, y se establece de nuevo (al igual que en la configuración de parámetros del modelo de referencia) que la proporción de elementos anómalos en la muestra de entrenamiento es de $\delta = 0{,}05$. Por lo tanto, la probabilidad de anomalía para cada observación es calculada a partir de la suma de probabilidades de una distribución binomial (Ecuación 3.1).



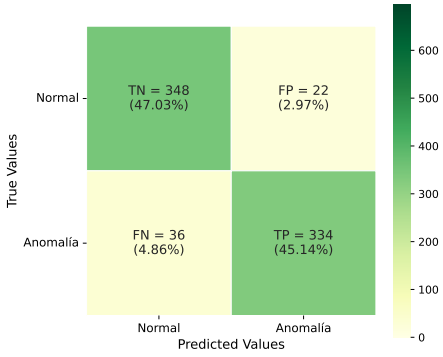
(a) *Split-conformal*



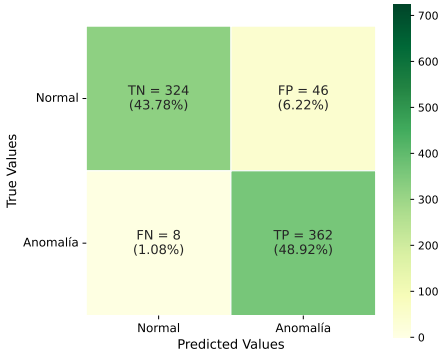
(b) *Cross-conformal* CV_{AD}



(c) *Cross-conformal* CV_{AD}^{+}



(d) *Cross-conformal* Jackknife$_{AD}$



(e) *Cross-conformal* Jackknife$_{AD}^{+}$

Figura 4.4: Matrices de confusión obtenidas para el conjunto test tras aplicar distintos métodos de inferencia conformal y *cross* conformal al modelo *Isolation Forest*.

Método	Exactitud	Precisión	Sensibilidad	Especificidad	Puntuación F1	AUC_ROC
Modelo de referencia	0.9405	0.9358	0.9459	0.9351	0,9409	0.9405
split-conformal	0.9338	0.9397	0.9270	0.9405	0.9333	0.9338
CV_{AD}	0.9459	0.9459	0.9459	0.9459	0.9459	0.9459
$CV_{AD}+$	0.9459	0.9508	0.9405	0.9514	0.9457	0.9459
Jackknife_{AD}	0.9216	0.9382	0.9027	0.9405	0.9201	0.9216
$\text{Jackknife}_{AD}+$	0.9270	0.8873	0.9784	0.8757	0.9306	0.9270

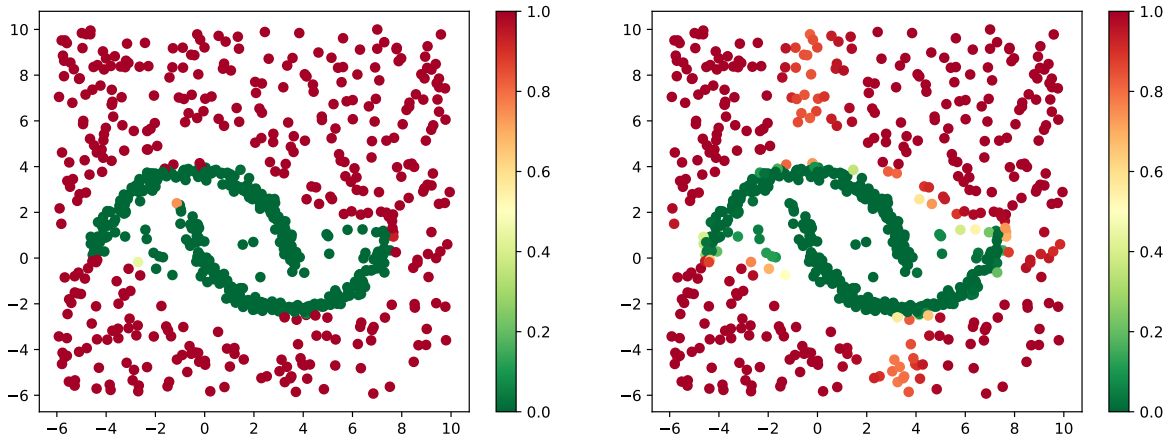
Tabla 4.1: Comparación de métricas por método de predicción.

Tras la aplicación del método, se obtienen las probabilidades de anomalía correspondientes para cada observación del conjunto test, que se representan en la Figura 4.5a, dónde el color rojo oscuro representa una alta probabilidad de anomalía, mientras que el verde oscuro se corresponde con una baja probabilidad. De forma generalizada, se observa una presencia reducida de puntos representados con colores claros, lo cual indica que, tras aplicar este método de medición de incertidumbre, apenas un número muy limitado de observaciones (en concreto, dos) presenta cierta incertidumbre asociada. Prácticamente la totalidad de los datos del conjunto test están representados por rojo o verde oscuro, lo que sugiere que los resultados proporcionados por *ExCeeD* indican una aparente certeza absoluta de que están correctamente clasificadas. Esto pone de manifiesto que, en este contexto, el método no es capaz de capturar adecuadamente la incertidumbre, ni sobre el modelo ni sobre sus predicciones.

Bootstrap

Finalmente, se procede a aplicar la implementación *bootstrap* desarrollada en el presente trabajo a este conjunto de datos. Para ello, se generan un total de 1000 remuestras *bootstrap*, cada una de las cuales se utiliza para entrenar un modelo *Isolation Forest* con la misma configuración de parámetros que la empleada en el modelo de referencia. Posteriormente, a partir de estos modelos, se calcula la probabilidad de anomalía para cada observación del conjunto test. En la Figura 4.5b se presentan las observaciones del conjunto test, junto con la probabilidad de que cada una de ellas sea considerada anómala, representada mediante una escala de color. Esta visualización sigue el mismo esquema empleado previamente para representar las probabilidades obtenidas con el método anterior.

Indicar que, en este caso, sí existe una mayor variabilidad en las probabilidades obtenidas. De acuerdo con lo esperado, y de forma similar a lo que ocurría para los p-valores conformales, las observaciones que obtienen probabilidades intermedias son aquellas próximas o que limitan con las estructuras en forma de luna. También obtienen probabilidades intermedias aquellas observaciones que se encuentran en regiones alineadas con estas estructuras (debido a los sesgos del algoritmo), indicando que presentan una mayor incertidumbre. En cambio, a medida que las observaciones se alejan de la región de los



(a) Probabilidad de anomalía (*ExCeeD*).

(b) Probabilidad de anomalía (*bootstrap*).

Figura 4.5: Probabilidad de anomalía calculada mediante métodos de medición de incertidumbre para cada observación. Esta está representada a partir de la escala de color, indicando el color rojo una alta probabilidad de anomalía y el verde una baja probabilidad.

semicírculos, o en su defecto, se adentran en ellas, se vuelven más próximas a 1 y a 0, respectivamente, como indican los colores rojo y verde oscuros, lo que indica una mayor confianza en la clasificación realizada.

Tras este ejemplo sencillo, se procede a presentar el conjunto de datos de tráfico de red sobre el que se aplican estos mismos métodos de medición de incertidumbre.

4.2. Aplicación a datos de tráfico de red

En contextos de ciberseguridad, la disponibilidad de alertas etiquetadas como verdaderas es muy limitada debido a la reducida presencia de ataques realmente detectados. En el marco del proyecto GIC-TEL, al tratarse de un proyecto en desarrollo, no se dispone de un conjunto de datos con anomalías etiquetadas suficientemente extenso como para realizar una evaluación precisa de los modelos que se van a aplicar. Por lo tanto, con el objetivo de evaluar el rendimiento de los distintos métodos de medición de incertidumbre expuestos en el Capítulo 3, se ha empleado el conjunto de datos de ciberseguridad para sistemas de control industrial (ICS) de la Universidad de Coimbra (Frazão et al., 2019). La elección de este conjunto de datos se justifica tanto por el conocimiento experto disponible en Gradiant de este tipo de datos como por su relevancia y aplicabilidad en proyectos dentro de la línea.

4.2.1. Descripción del conjunto de datos

En esta sección se expone una descripción detallada del conjunto de datos, incluyendo su proceso de generación, el preprocesamiento realizado y el análisis exploratorio llevado a cabo sobre el mismo.

El conjunto de datos utilizado para la validación de los métodos propuestos se ha generado en un laboratorio de ciberseguridad industrial a pequeña escala. En este, se simula la comunicación entre

diferentes dispositivos industriales, denominados PLC (*Programmable Logic Controler*, (National Institute of Standards and Technology), que utilizan el protocolo MODBUS/TCP (Modbus Organization, 2012) para comunicarse. El conjunto de datos está especialmente diseñado para su uso en la validación de algoritmos de detección de anomalías, y contiene datos etiquetados (tanto normales como anómalos) de tráfico de red, es decir, datos que se transmiten a través de una red de comunicaciones, encapsulados en paquetes de información que circulan entre los distintos dispositivos conectados.

De esta manera, sobre el laboratorio se ejecutan una serie de ataques de Denegación de Servicio, cuyo objetivo es interrumpir o degradar la capacidad de comunicación entre los dispositivos de la red. En este caso, el ataque ha sido ejecutado desde múltiples dispositivos al mismo tiempo (para maximizar su eficacia), por lo que se considera un ataque de Denegación de Servicio Distribuido (DDoS, Horak et al. (2021)). Dependiendo de la capacidad de la red y las características de los dispositivos atacados, el ataque puede tener diferentes efectos en un sistema industrial, por ejemplo, interrumpir o retrasar líneas de producción, afectando el comportamiento habitual de los dispositivos que están bajo ataque. En este conjunto de datos hay tres tipos diferentes de ataques DDoS:

- Inundación de *queries* MODBUS: El atacante consigue inyectar peticiones fraudulentas, específicamente paquetes de *queries* MODBUS falsos en la red, con el objetivo de tomar control del sistema mediante este flujo masivo de mensajes, y de este modo anular efectivamente las comunicaciones legítimas (Bhatia et al., 2014).

- Inundación de paquetes TCP SYN: Este tipo de ataque es el segundo más común que utiliza el protocolo clave de internet TCP, específicamente una de sus características, que las conexiones TCP se componen de tres etapas (*3-way handshake*). El comportamiento normal consiste en el envío de un paquete SYN, seguido de una respuesta con un paquete SYN/ACK (indica que la petición SYN es válida) y finalmente, otro envío de un ACK, completando la negociación en tres pasos. Sin embargo, durante un ataque de inundación de paquetes TCP SYN, el servidor que recibe el paquete SYN y responde con otro SYN/ACK, no recibe de vuelta el mensaje de confirmación ACK. Esto produce que el servidor no conozca el estado de la comunicación y se quede esperando el mensaje ACK, por lo que la conexión se mantiene semi-abierta durante un periodo de tiempo, lo cual provoca el agotamiento de sus recursos. Esta situación puede causar el agotamiento de los recursos del servidor, causando su mal funcionamiento (Horak et al., 2021).

- Inundación de paquetes Ping: Se trata de uno de los ataques DoS más comunes. El atacante inunda el objetivo con solicitudes Ping del protocolo ICMP (*Internet Control Message Protocol*), de forma similar a la inundación de *queries* MODBUS. Este ataque fuerza a la red a responder a la totalidad de las peticiones que llegan, causando la saturación e indisponibilidad de la red (Horak et al., 2021).

Como ya se ha mencionado, el conjunto de datos incluye tanto tráfico de red normal (sin la presencia de ataques), como tráfico de red anómalo (el tráfico generado por el propio atacante y el tráfico de red con comportamiento alterado de los dispositivos legítimos). Para generar el tráfico de red anómalo, los autores realizaron ataques de los tres tipos (en simulaciones diferenciadas) durante distintos períodos de tiempo (1, 5 y 15 minutos) sobre capturas de tráfico de 30 minutos y una hora, obteniendo así 18 escenarios diferentes. En este trabajo únicamente se emplean los datos de los ataques de un minuto de duración sobre capturas de una hora, para todos los tipos de ataques.

Tal y como se ha expuesto en el Capítulo 3, los modelos *Isolation Forest* y *Extended Isolation Forest* se entrenan de forma no supervisada, empleando exclusivamente datos normales, pero sin etiquetas. Así, el conjunto de datos empleado para el entrenamiento se corresponde con 30409 observaciones de tráfico de red normal. Una vez entrenados, se evalúan los algoritmos y los métodos de medición de incertidumbre aplicados sobre ellos mediante un conjunto test, que engloba a los tres ataques. Estos están formados por las capturas de tráfico de red, tanto anómalas como no, recopiladas durante una hora, en la que se producen los ataques de un minuto de duración, y en conjunto comprenden 7634 observaciones normales y 8795 anómalas. Las variables seleccionadas para el entrenamiento y la validación de los modelos se muestran en la Tabla 4.2.

Variable	**Descripción**	**Tipo de dato**
srcip	Dirección IP del origen de la conexión	IP
smac	Dirección MAC del origen de la conexión	MAC
dstip	Dirección IP del destino de la conexión	IP
dmac	Dirección MAC del destino de la conexión	MAC
proto	Protocolo empleado en la comunicación	Integer
dport	Puerto de destino de la conexión	Integer
dpkts	Paquetes enviados de destino a origen	Integer
dbytes	Bytes enviados de destino a origen	Integer
spkts	Paquetes enviados de origen a destino	Integer
sbytes	Bytes enviados de origen a destino	Integer
pkts	Total de paquetes enviados	Integer
bytes	Total de bytes enviados	Integer
dur	Duración de la conexión (en segundos)	Float
sintpkt	Tiempo entre paquetes (origen a destino, en milisegundos)	Float
dintpkt	Tiempo entre paquetes (destino a origen, en milisegundos)	Float

Tabla 4.2: Descripción de las variables seleccionadas del conjunto de datos empleado.

Preprocesamiento

Antes de realizar el entrenamiento de los algoritmos de detección de anomalías, es necesario llevar a cabo un preprocesamiento de los datos, tanto debido al formato de los datos, como a necesidades específicas de los propios algoritmos.

Atendiendo al formato de los datos, en la Tabla 4.2, se indica que `srcip` y `dstip` son direcciones IP (formadas por cuatro números enteros entre 0 y 255, separados por puntos, por ejemplo 192.168.1.1), y `smac` y `dmac` son direcciones MAC (compuestas por seis bloques de caracteres hexadecimales separados por dos puntos, por ejemplo 00:1A:2B:3C:4D:5E). No obstante, los métodos de detección de anomalías que van a ser empleados requieren que todas las variables de los datos tomen valores numéricos, por lo que es necesario realizar una transformación a estas variables si se desean incluir. Por una parte, en el caso de las direcciones IP, se emplea la clase `IPAdress` de *Python*, ya que el parámetro *ip* de un objeto de esta proporciona el equivalente numérico de una dirección IP. Por otra parte, para las direcciones MAC, se eliminan los caracteres no alfanuméricos y se convierte esa cadena hexadecimal (sin separadores) a un entero base 16.

Una vez realizada la transformación del formato de los datos, se procede al tratamiento específico de los datos requerido por los algoritmos de detección de anomalías empleados. El *Isolation Forest* divide el espacio de variables realizando cortes recursivos en una única variable cada iteración, por lo que no es sensible a la escala en la que se encuentran los datos. Esto elimina la necesidad de aplicar técnicas de normalización o estandarización. En cambio, el *Extended Isolation Forest* realiza las divisiones del espacio de características mediante hiperplanos aleatorios, que involucran más de una variable (la cantidad que determine el parámetro extensión), por lo que, si estas se encuentran en escalas muy distintas, se pueden producir sesgos y resultados incorrectos. Por ello, se lleva a cabo una normalización Min-Max. Esto se realiza mediante la función `MinMaxScaler()` de la librería `scikit-learn` de Python, y consiste en reescalar los datos en cada una de las variables de forma independiente, de forma que todas tomen valores en el intervalo $[0, 1]$. Dado un conjunto de datos $\mathcal{D}$, formado por n observaciones de d variables, el valor normalizado de la i-ésima observación para la j-ésima variable se obtiene mediante la siguiente fórmula:

$$x'_{ij} = \frac{x_{ij} - x_{min,j}}{x_{max,j} - x_{min,j}},$$

donde $x_{min,j}$ y $x_{max,j}$ son, respectivamente, el menor y mayor valor que toma la variable j-ésima en $\mathcal{D}$ y x_{ij} es el valor original de la variable para la i-ésima observación.

Análisis exploratorio

Para comenzar el análisis exploratorio de los datos, se calculan una serie de estadísticos descriptivos para las variables previamente seleccionadas. Específicamente, se calculan la media, la desviación típica, el primer, segundo (mediana) y tercer cuartil, el mínimo y el máximo, que permiten estudiar la estructura y comportamiento del conjunto de datos.

En la Tabla 4.3 se recogen dichas medidas para las variables numéricas en el caso de observaciones normales. Se puede observar que las variables `spkts` (paquetes enviados de origen a destino), `dpkts` (de destino a origen) y `pkts` (cantidad total de paquetes) presentan valores bajos, tanto en la mediana

como en el tercer cuartil, por lo que la mayoría de las comunicaciones legítimas involucran el intercambio de cantidades pequeñas de paquetes, pese a la existencia de alguna conexión con un mayor volumen, como indican los máximos. También se aprecia que se suelen recibir y enviar cantidades similares de paquetes, siendo superior la cantidad de paquetes enviados. La cantidad de bytes exhibe un comportamiento similar.

Atendiendo a la duración de las conexiones (`dur`), se observa que, en general, estas son breves, ya que presentan una mediana de 0.01 y un máximo de 0.1 segundos. El tiempo entre paquetes presenta un primer cuartil de 0, tanto para los que van de origen a destino (`sintpkt`) como para los que van desde el destino al origen (`dintpkt`). No obstante, atendiendo a la media, mediana y tercer cuartil, los tiempos entre paquetes enviados al destino son considerablemente más altos. En particular, `sintpkt` alcanza valores superiores a 190, lo que refleja la existencia de conexiones legítimas con pausas prolongadas entre paquetes, aunque no sean las más frecuentes. Mencionar que el valor de estas dos variables no es computable cuando solo se ha enviado un paquete (`spkts` = 1), por lo que se considera en su defecto 0 ms, pero su distribución puede incluir valores residuales. En resumen, el tráfico normal se caracteriza por conexiones breves, en las que se envían pocos paquetes de poco tamaño en ambas direcciones.

Variable	spkts	dpkts	pkts	sbytes	dbytes	bytes	dur	sintpkt	dintpkt
Media	1.813	0.901	2.714	114.520	65.945	180.465	0.039	44.527	0.520
Desv. tip.	2.853	1.126	3.628	173.205	70.480	216.663	0.044	48.183	1.699
Mínimo	0.0	0.0	1.0	0.0	0.0	60.0	0.0	0.0	0.0
Q_1	1.0	0.0	1.0	60.0	0.0	85.0	0.0	0.0	0.0
Q_2	1.0	1.0	2.0	66.0	85.0	126.0	0.010	1.466	0.0
Q_3	2.0	1.0	2.0	126.0	85.0	151.0	0.094	94.246	0.0
Máximo	14.0	10.0	18.0	1059.0	850.0	1318.0	0.100	195.933	7.860

Tabla 4.3: Medidas descriptivas para las variables numéricas del conjunto de datos normales, destinados para el entrenamiento de los modelos.

En la Tabla 4.4 se exponen las mismas medidas descriptivas que se han calculado anteriormente para el conjunto de datos normales pero, en este caso, se calculan para las observaciones anómalas del conjunto test, que engloban los ataques de inundación de *queries* MODBUS, de paquetes TCP SYN y de paquetes Ping y los datos de tráfico modificados como consecuencia de los ataques. Si se comparan estas métricas con las de los datos normales, se observa que la cantidad de paquetes enviados, tanto los totales como los que son enviados desde el origen al destino (`spkrs` y `pkts`), es menor a la que

se tiene en los datos normales, ya que para al menos un 75 % de las conexiones se intercambia un único paquete. Además, los paquetes enviados desde el destino al origen, `dpkts`, se reducen de forma clara (al menos en el 50 % de las comunicaciones esta variable es 0), quedando reflejado el carácter unidireccional de los ataques de denegación de servicio. Este comportamiento también se manifiesta en los bytes enviados y recibidos durante las comunicaciones.

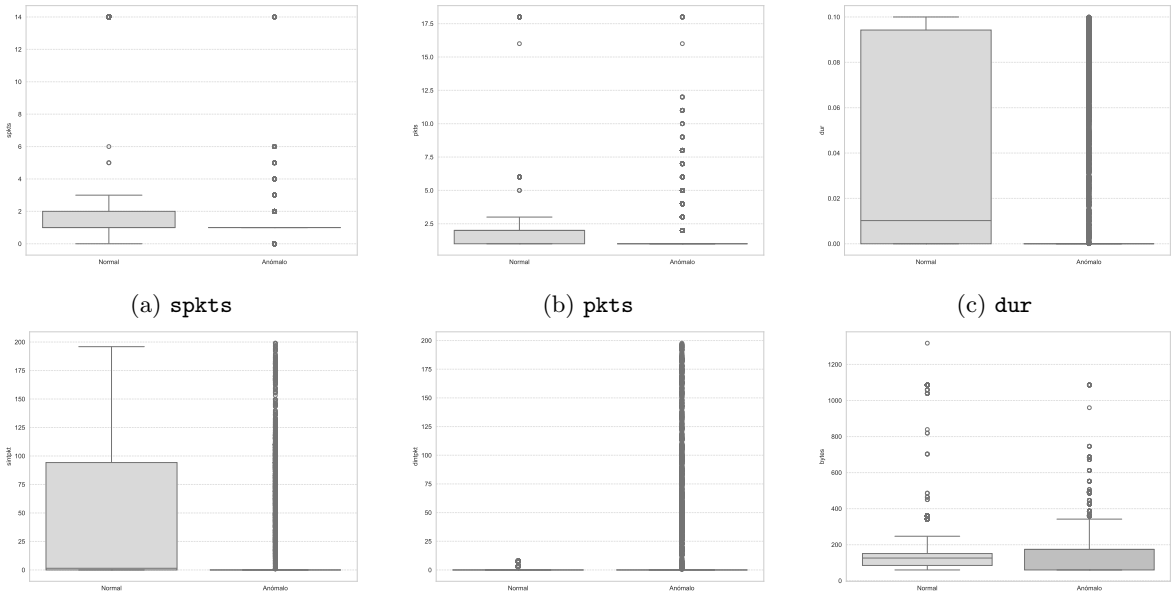
Además, los ataques también afectan a la duración de las conexiones y al tiempo que transcurre entre el envío de paquetes. Por una parte, la duración de las conexiones, `dur`, es notablemente menor en el tráfico anómalo, con una media de 0.011 y una mediana de 0.0, frente a 0.039 y 0.010, respectivamente, en el caso del tráfico normal, siendo la gran mayoría de las conexiones anómalas de 0 milisegundos de duración. En cuanto a los intervalos de tiempo entre paquetes, se observa que el tiempo entre paquetes enviados desde el origen, `sinpkt`, disminuye considerablemente frente al comportamiento normal. Esta reducción se explica por el hecho de que en la mayoría de las conexiones anómalas se envía un único paquete, por lo que la variable toma el valor 0. En cambio, en el caso de los tiempos entre paquetes enviados desde el destino, `dinpkt`, la media es superior en los datos anómalos (11,8 frente a 0,52), pese a que, al igual que en los datos normales, el tercer cuartil sea 0. Esto indica la presencia de tiempos muy elevados en el conjunto de las anomalías, que elevan la media. Por lo tanto, el tráfico anómalo se caracteriza por conexiones en las que se realizan envíos únicos unidireccionales y muy breves.

Variable	spkts	dpkts	pkts	sbytes	dbytes	bytes	dur	sintpkt	dintpkt
Media	1.189	0.446	1.635	120.02	27.209	147.229	0.011	14.883	11.797
Desv. Tip.	0.797	0.964	1.518	66.279	58.585	90.199	0.027	34.807	34.634
Mínimo	0.0	0.0	1.0	0.0	0.0	60.0	0.0	0.0	0.0
Q_1	1.0	0.0	1.0	60.0	0.0	60.0	0.0	0.0	0.0
Q_2	1.0	0.0	1.0	134.0	0.0	174.0	0.0	0.0	0.0
Q_3	1.0	1.0	1.0	174.0	60.0	174.0	0.0	0.0	0.0
Máximo	14.0	6.0	18.0	840.0	366.0	1086.0	0.100	199.003	197.701

Tabla 4.4: Medidas descriptivas para las variables numéricas del conjunto de datos anómalos correspondientes a los ataques DDoS y al tráfico alterado que estos causan.

En la Figura 4.6 se muestran diagramas de caja correspondientes a algunas de las variables del conjunto de datos. Estas representaciones permiten visualizar de forma clara las discrepancias ya mencionadas entre el comportamiento del tráfico normal y el anómalo. Estas diferencias serán explotadas por los métodos de detección de anomalías para conseguir discernir las conexiones reales de las fraudulentas. Cabe señalar que no se ha llevado a cabo un análisis de correlación entre variables, ya que,

en el caso de los modelos empleados, la presencia de correlaciones no resulta relevante para su funcionamiento.



(a) `spkts` (b) `pkts` (c) `dur`

(d) `sintpkt` (e) `dintpkt` (f) `bytes`

Figura 4.6: Diagramas de caja considerando tráfico normal y anómalo de algunas de las variables del conjunto de datos.

4.2.2. *Isolation Forest*

En esta sección, se muestran los resultados obtenidos tras aplicar las técnicas de medición de incertidumbre al algoritmo *Isolation Forest* utilizando el conjunto de datos descrito en la sección anterior.

El proceso se inicia con el entrenamiento de un modelo *Isolation Forest* utilizando el conjunto de entrenamiento, formado por un total de 30409 observaciones normales. Para determinar la configuración de parámetros más adecuada, se realiza previamente un *gridsearch* que abarca todas las posibles combinaciones de los siguientes parámetros:

- `n_estimators`: número total de árboles que son entrenados. Se consideran 50, 100, 150 y 200.
- `max_samples`: número de observaciones utilizadas para entrenar cada árbol. Se evalúan 4000, 5000, 6000, 7000, 8000 y 9000.

La combinación de parámetros seleccionada fue aquella que obtuvo un mejor AUC-ROC, que se corresponde con `n_estimators` = 50 y `max_samples` = 8000, por lo que es la empleada para entrenar el modelo de referencia. Por otra parte, la contaminación de la muestra no fue especificada, de forma que el algoritmo utiliza por defecto un umbral de 0,5.

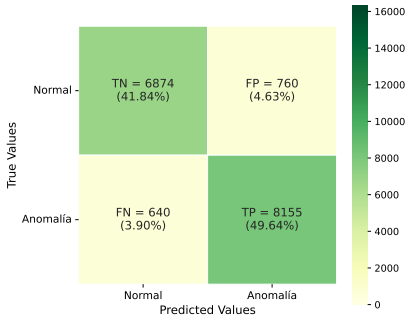


Figura 4.7: Matriz de confusión relativa a las predicciones obtenidas mediante el algoritmo *Isolation Forest* para el conjunto test.

A continuación, se procede a la evaluación de este modelo mediante el conjunto test, compuesto por observaciones anómalas y normales. En la Figura 4.7 se expone la matriz de confusión obtenida y, a partir de esta, se calculan las distintas métricas de evaluación:

- Exactitud: 0,9148
- Precisión: 0,9148
- Exhaustividad: 0,9272
- Especificidad: 0,9004
- Puntuación F1: 0,9209
- AUC-ROC: 0,9138

Dado que el conjunto test está balanceado, es posible interpretar las métricas directamente sin necesidad de ajustes. Atendiendo a ellas, se comprueba que el modelo (de referencia, a partir de ahora) presenta un rendimiento global satisfactorio, con alta capacidad de detección (exhaustividad = 0,9272) y un control satisfactorio de los falsos positivos (especificidad = 0,9004). La puntuación F1 y el AUC-ROC, ambos superiores a 0,91, reafirman su eficacia global, por lo que se procede a medir la incertidumbre del modelo y la de sus predicciones.

4.2.3. Medición de incertidumbre para *Isolation Forest*

Split-conformal y *cross-conformal*

En primer lugar, se procede a aplicar los métodos de medición de incertidumbre basados en los p-valores conformales. En el caso de los métodos *cross-conformal*, únicamente se emplean las técnicas CV_{AD} y CV+_{AD}, no aquellos basados en Jackknife. Esto se debe a la elevada complejidad computacional y espacial que conllevan, ya que se cuenta con una muestra de entrenamiento formada por 30409 observaciones y para aplicar estos métodos es necesario entrenar, y en el caso de Jackknife+$_{AD}$, almacenar, 30409 modelos. De esta forma, su aplicación resulta inviable, por lo que se aplican exclusivamente los métodos *split-conformal*, CV_{AD} y CV+_{AD}. La decisión de no incluir esta metodología también responde a su comportamiento deficiente para el conjunto de datos empleado anteriormente, de forma que no ofrece beneficios suficientes como para justificar el alto coste computacional.

Para la estimación de incertidumbre mediante el enfoque *split-conformal* se emplea un conjunto de calibración formado por 4000 observaciones. En el caso de los métodos *cross-conformal*, se prescinde del conjunto de calibración y se utilizan 20 subconjuntos para realizar la validación cruzada, excluyendo en cada caso la proporción correspondiente de datos. Todos los modelos *Isolation Forest* entrenados en este proceso emplean la misma configuración de parámetros que el *Isolation Forest* de referencia, garantizando así la coherencia metodológica entre técnicas. Una vez se lleva a cabo el entrenamiento de los modelos necesarios para cada método, se procede al cálculo de los p-valores asociados a cada observación x_i del conjunto test, para resolver el siguiente contraste de hipótesis

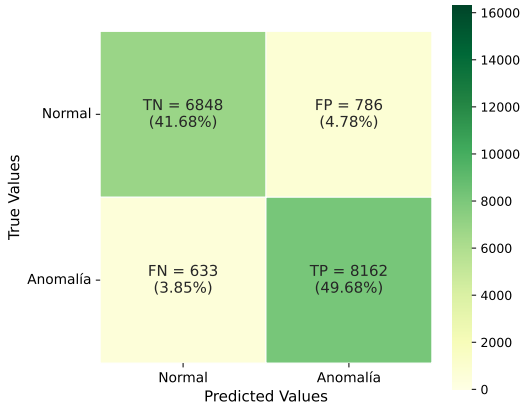
$$\begin{cases} H_{0,i} : x_i \text{ es una observación normal} \\ H_{1,i} : x_i \text{ es una observación anómala.} \end{cases}$$

Dado que este proceso implica la evaluación simultánea de un gran volumen de contrastes de hipótesis (16429), es necesario adoptar una perspectiva de test múltiples. De manera análoga a lo realizado con el conjunto de datos *Moon*, se emplea el método de Storey para controlar la tasa de falsos descubrimientos, y se lleva a cabo la corrección de los p-valores a partir de la función `p.adjust()` del paquete `stats` de R.

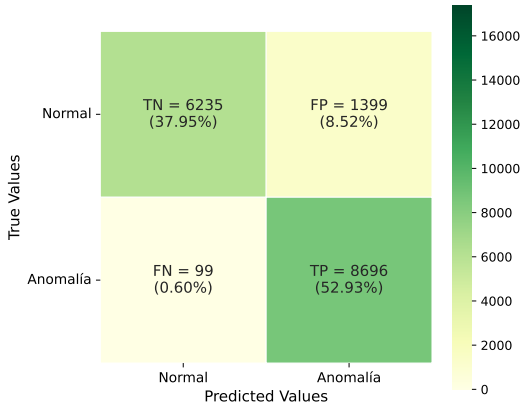
Se realiza una clasificación de todas las observaciones del conjunto test en normales o anómalas, mediante la resolución del contraste de hipótesis anterior, controlando la tasa de falsos descubrimientos para los niveles 0,10 y 0,15 y, por lo tanto, manteniendo garantías estadísticas. En la Figura 4.8 se presentan las seis matrices de confusión correspondientes a los tres métodos de predicción conformal y a ambas FDRs. Atendiendo a estas, y como era de esperar, se detecta que permitir una mayor tasa de falsos descubrimientos aumenta considerablemente los falsos positivos y reduce los falsos negativos, es decir, se clasifican más observaciones como anómalas. No obstante, este comportamiento no se produce para CV_{AD} (4.8d), donde sí que aumenta la cantidad de falsos positivos pero no se reducen los falsos negativos, reflejando un peor rendimiento.

Dado que se desea analizar el impacto de los distintos métodos de medición de incertidumbre y del nivel de FDR sobre el rendimiento en la detección de observaciones anómalas, se calculan de nuevo las métricas de evaluación para estas clasificaciones. Estas se muestran en la Tabla 4.5, junto a la FDR observada, que es menor que la controlada para los métodos *split-conformal* y $CV+_{AD}$, pero mayor para CV_{AD}. Analizando los valores expuestos, se observa que aumentar la FDR controlada tiene un efecto negativo en la mayoría de las métricas, especialmente en la especificidad y en la precisión, debido al aumento de los falsos positivos. En contraposición, se obtiene un incremento en la sensibilidad, ya que se detectan un mayor número de anomalías, especialmente en el caso de los métodos *split-conformal* y $CV+_{AD}$. Comparando los distintos métodos, en términos generales, todos ellos producen resultados similares, siendo los de $CV+_{AD}$ ligeramente inferiores.

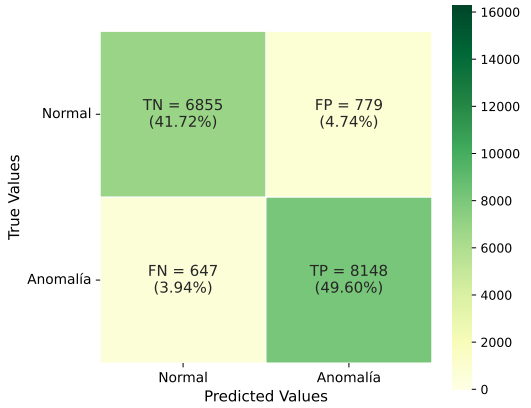
A continuación, se analizan los p-valores corregidos obtenidos para el conjunto test, en función de si la observación asociada a este se trata de un verdadero positivo, verdadero negativo, falso positivo o falso negativo en la clasificación obtenida a partir del *Isolation Forest* de referencia. Esto se lleva a cabo mediante los histogramas presentados en la Figura 4.9. Cada barra representa la densidad de observaciones en función del p-valor corregido, diferenciados según la etiqueta real y la predicción del modelo. Debido a que los p-valores están corregidos, estos no se distribuyen de forma uniforme en el intervalo $[0, 1]$ bajo la hipótesis nula. Por ello, en las gráficas únicamente se muestra en el eje X el



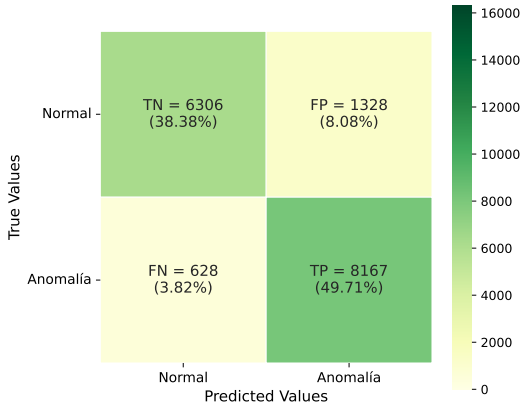
(a) *split-conformal* controlando la FDR al nivel 0.10



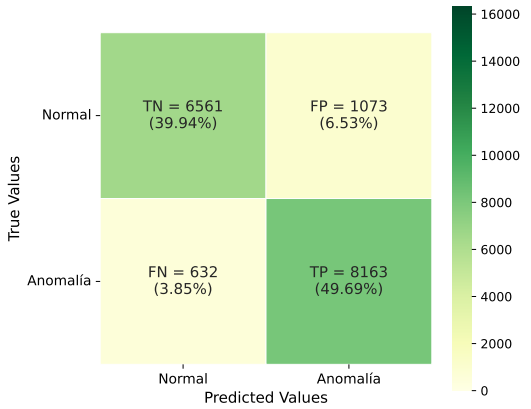
(b) *split-conformal* controlando la FDR al nivel 0.15



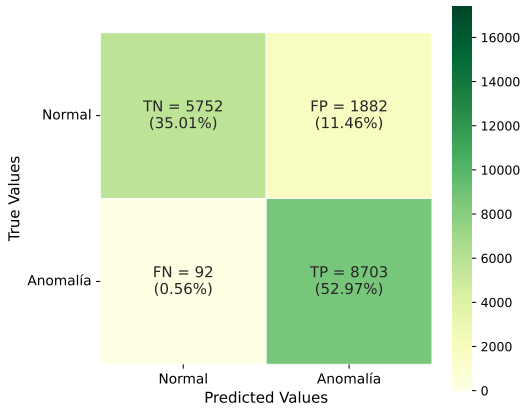
(c) CV_{AD} controlando la FDR al nivel 0.10



(d) CV_{AD} controlando la FDR al nivel 0.15



(e) $CV_{AD}+$ controlando la FDR al nivel 0.10



(f) $CV_{AD}+$ controlando la FDR al nivel 0.15

Figura 4.8: Matrices de confusión obtenidas a partir de los p-valores obtenidos al aplicar técnicas de predicción conformal al modelo *Isolation Forest*. Se muestran los resultados que resultan de controlar la FDR a dos niveles: 0.10 (columna izquierda) y 0.15 (columna derecha).

Método	FDR	FDR_{obs}	Exactitud	Precisión	Sensibilidad	Especificidad	Puntuación F1	AUC ROC
Modelo de referencia	–	–	0.9148	0.9148	0.9272	0.9004	0.9209	0.9138
split-conformal	0.10	0.0878	0.9136	0.9122	0.9280	0.8970	0.9200	0.9125
	0.15	0.1386	0.9088	0.8614	0.9887	0.8167	0.9207	0.9027
CV_{AD}	0.10	0.0868	0.9132	0.9127	0.9264	0.8980	0.9195	0.9122
	0.15	0.1399	0.8809	0.8601	0.9286	0.8260	0.8931	0.8773
$CV_{AD}+$	0.10	0.1162	0.8962	0.8838	0.9281	0.8594	0.9054	0.8938
	0.15	0.1778	0.8798	0.8222	0.9895	0.7535	0.8981	0.8715

Tabla 4.5: Comparación de métricas de evaluación para los distintos métodos de predicción conformal de detección de anomalías aplicados al modelo *Isolation Forest*, bajo dos niveles de control de la tasa de falsos descubrimientos (FDR). Se incluye tanto la FDR controlada como la FDR observada tras aplicar cada técnica.

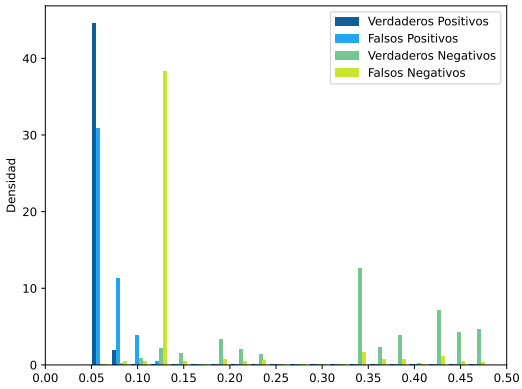
intervalo $[0, 0{,}5]$ que contiene a la totalidad de los p -valores, con el objetivo de facilitar la interpretabilidad de las mismas.

En términos generales, se observa que los verdaderos positivos se concentran en regiones caracterizadas por p -valores bajos, cercanos a 0, lo cual indica que existen fuertes evidencias significativas que respaldan la clasificación de estas observaciones como anomalías. No obstante, esta misma concentración de valores-p bajos también se observa en los falsos positivos, lo que indica que estas metodologías no permiten distinguir entre ambos. Esta limitación es especialmente evidente en el caso del método CV_{AD} (Figura 4.9b). En cambio, las técnicas de *split-conformal* (Figura 4.9a) y $CV+_{AD}$ (Figura 4.9c), muestran una mayor capacidad para diferenciar ambas categorías, lo que se traduce en un mejor rendimiento. En relación a los verdaderos positivos, estos se distribuyen de manera uniforme en el intervalo para los tres métodos, de forma que no existen evidencias significativas de que estas observaciones sean anómalas. Por último, se observa que los falsos negativos tienden a agruparse en intervalos específicoss: en torno a 0.13 para *split-conformal*, 0.20 para CV_{AD} y 0.15 para $CV+_{AD}$. Por tanto, en este caso particular, la clasificación obtenida puede variar de forma significativa en función del nivel de la tasa de falsos descubrimientos que se desee controlar, lo cual dependerá de los requisitos específicos de cada proyecto. Esta variabilidad se refleja claramente en las matrices de confusión mostradas en la Figura 4.8.

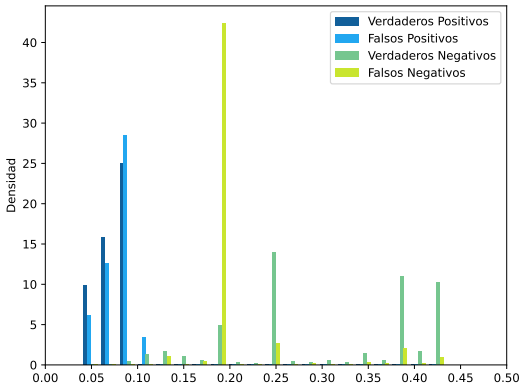
ExCeeD

Tras aplicar los métodos de predicción conformal, se procede a aplicar el método *ExCeeD* sobre el modelo de referencia. Este método permite estimar, para cada observación, la probabilidad de que esta esté correctamente clasificada, cuantificando así la incertidumbre en las predicciones del modelo.

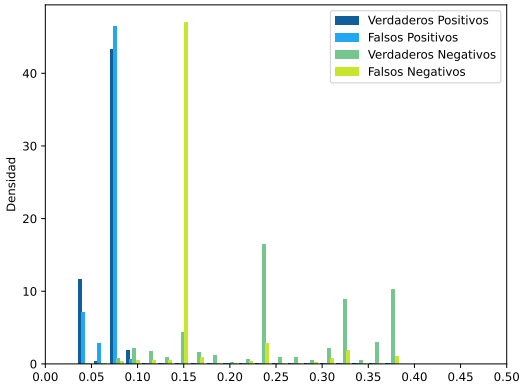
En este caso, únicamente se calcula la probabilidad de que una observación del conjunto test sea anómala si ha sido clasificada como tal por el *Isolation Forest* de referencia. Esta decisión responde a



(a) Histograma de los p-valores corregidos obtenidos mediante el método *split-conformal*.



(b) Histograma de los p-valores corregidos obtenidos mediante el método CV_{AD}.



(c) Histograma de los p-valores corregidos obtenidos mediante el método $CV+_{AD}$.

Figura 4.9: Distribución de los p-valores corregidos para cada una de las técnicas de predicción conformal, diferenciando observaciones según la clasificación obtenida en el modelo de referencia y su categoría real.

una consideración práctica habitual en sistemas reales de detección de anomalías: el análisis se realiza exclusivamente sobre los casos que el modelo clasifica como anómalos, ya que son los que activan procesos de revisión, alerta o mitigación. Las observaciones clasificadas como normales no generan una respuesta automática en el sistema y, por tanto, no requieren una estimación adicional de su probabilidad de anomalía. Así, cada alerta es acompañada de una probabilidad de anomalía, aportando al investigador información adicional, que complementa el aviso y que puede ser empleada para priorizar, filtrar o automatizar decisiones.

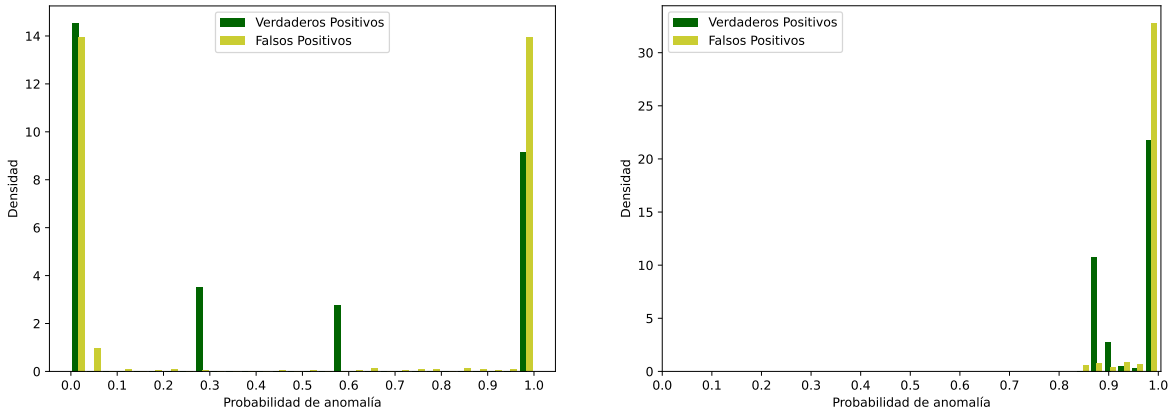
A continuación, se procede a aplicar el método. En principio, dado que el conjunto de entrenamiento empleado está compuesto exclusivamente de observaciones normales, la tasa de contaminación esperada sería cero. No obstante, con el fin de considerar posibles errores de etiquetado, ruido o la presencia de observaciones anómalas, se establece una tasa de contaminación de $\delta = 0{,}05$. Sin embargo, cabe destacar que, pese a que esta metodología tenga una formulación específica para la situación en la que la tasa de contaminación es cero (Ecuación 3.2), esta no es apropiada si el conjunto de entrenamiento es grande. En este caso, la probabilidad es estimada mediante una potencia, en la que la base es estrictamente menor que uno, y el exponente es el número de observaciones del conjunto de entrenamiento (más de 30000 en este contexto). Por lo tanto, se generan de forma sistemática valores cercanos al 0, y no se cuantifican de manera correcta las probabilidades correspondientes.

Por lo tanto, se emplea la Ecuación 3.1 para calcular las probabilidades de anomalía de las observaciones que han sido clasificadas como anómalas por el modelo de referencia. En la Figura 4.10a se exponen las probabilidades de anomalía para las observaciones clasificadas como tales por el *Isolation Forest* de referencia, diferenciando entre falsos y verdaderos positivos, representados mediante los colores amarillo y verde, respectivamente.

Se espera que el método asigne probabilidades de anomalía elevadas a los verdaderos positivos y probabilidades más bajas a los falsos positivos. No obstante, la gráfica muestra que una gran cantidad de verdaderos positivos recibe una probabilidad de anomalía igual a 0, lo cual es contrario a lo esperado. Esto muestra cómo el método no es capaz de asignar probabilidades de anomalía elevadas a una gran proporción de los verdaderos positivos. Al mismo tiempo, se observa que también asigna probabilidades altas a falsos positivos, transmitiendo una falsa sensación de seguridad. Aunque en algunos casos el método sí asigna probabilidades altas a verdaderos positivos y bajas a falsos positivos, lo hace en la misma proporción que en el caso anterior, sin mejorar la discriminación entre ambos grupos. Por lo tanto, esto sugiere que este método no solo no captura correctamente la incertidumbre ya que únicamente proporciona valores extremos, sino que además compromete la fiabilidad y utilidad del sistema de detección.

Bootstrap

También se emplea en este conjunto de datos la implementación *bootstrap* presentada en el Capítulo 3, que permite estimar probabilidades de anomalía. Se generan 1000 remuestras *bootstrap* a partir del conjunto de entrenamiento original, y con ellas se entrenan 1000 modelos *Isolation Forest*. Cada uno de estos modelos ha sido entrenado empleando la configuración de parámetros utilizada para el modelo de referencia. Posteriormente, se ha evaluado cada observación del conjunto test clasificada como anómala por el modelo de referencia mediante los 1000 modelos *bootstrap*, y se ha estimado su



(a) Probabilidad de anomalía calculada mediante *ExCeeD*.

(b) Probabilidad de anomalía calculada mediante *bootstrap*.

Figura 4.10: Probabilidad de anomalía calculada mediante métodos de medición de incertidumbre para cada observación detectada como anómala del conjunto test. Se distingue entre los falsos positivos (amarillo) y los verdaderos positivos (verde).

probabilidad de ser clasificada como anómala a partir de la proporción de veces que fue identificada como tal, de forma análoga al método anterior.

Los resultados obtenidos se muestran en la Figura 4.10b, donde, de nuevo, las probabilidades de los verdaderos positivos están representadas en verde y las probabilidades de los falsos positivos en amarillo. Para este método, todas las probabilidades toman valores elevados, mayores que 0,8. Por una parte, se aprecia que las probabilidades de los verdaderos positivos están concentradas en valores cercanos al 1, indicando que el sistema sí que es capaz de otorgar alta confianza a las observaciones que realmente son anómalas. Sin embargo, esto también ocurre para los falsos positivos, incluso en mayor medida. Esto refleja que el modelo presenta una mayor incertidumbre para más anomalías correctamente clasificadas que para incorrectas, lo cual es contrario a lo deseado, que el modelo esté seguro de las anomalías reales pero dude en las observaciones normales. No obstante, es también informativo y puede indicar que es necesario modificar el proceso de entrenamiento o el algoritmo empleado.

Tras haber aplicado las técnicas de medición de incertidumbre al *Isolation Forest*, se procede a realizar lo análogo para el *Extended Isolation Forest*, el otro modelo objeto de estudio.

4.2.4. *Extended Isolation Forest*

En esta sección se presentan los resultados obtenidos tras aplicar las técnicas de medición de incertidumbre estudiadas sobre el modelo *Extended Isolation Forest*.

Al igual que en el caso anterior, el modelo es entrenado empleando únicamente observaciones normales, siguiendo el enfoque de aprendizaje no supervisado habitual en detección de anomalías. Para garantizar un buen rendimiento del modelo, se realiza un *gridsearch*, evaluando distintas configuraciones de parámetros y conservando aquella que obtenga un mayor AUC-ROC. Los parámetros explorados

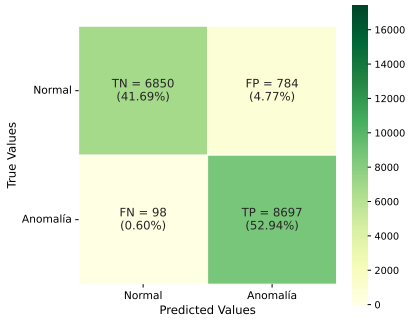


Figura 4.11: Matriz de confusión relativa a las predicciones obtenidas mediante el algoritmo *Extended Isolation Forest* para el conjunto test.

fueron:

- `sample_size`: tamaño de muestra empleado para entrenar cada árbol. Se probaron los valores 500, 1000, 1500, 2000, 2500 y 3000.
- `ntrees`: número de árboles que se entrenan. Se consideraron 50, 100 y 150 árboles.
- `ExtensionLevel`: controla la cantidad de variables que se consideran en la división del espacio de los datos. Un valor de cero es análogo a los cortes que realiza el algoritmo *Isolation Forest*. Se evaluan los niveles 1, 2, 3 y 4.

Tras evaluar el rendimiento de cada combinación de los valores presentados, se selecciona como óptima la siguiente configuración de parámetros: `sample_size`= 2000, `ntrees`= 150 y `ExtensionLevel`= 1.

A continuación, para llevar a cabo la clasificación de las observaciones del conjunto test, es necesario establecer una tasa de contaminación δ. Siguiendo un razonamiento análogo al empleado en la aplicación del método *ExCeeD*, y manteniendo la coherencia metodológica, se emplea $\delta = 0{,}05$. Tras calcular el umbral correspondiente y llevar a cabo la clasificación, se calcula la matriz de confusión sobre el conjunto test, mostrada en la Figura 4.11, y a partir de ella se obtienen las métricas de evaluación:

- Exactitud: 0,9463
- Precisión: 0,9173
- Exhaustividad: 0,9889
- Especificidad: 0,8973
- Puntuación F1: 0,9517
- AUC-ROC: 0,9431

Estas reflejan el muy buen rendimiento del modelo para la detección de anomalías. Al comparar este modelo con su análogo *Isolation Forest*, resulta evidente que el *Extended Isolation Forest* supera al *Isolation Forest* en todas las métricas evaluadas. Esto se corresponde con lo esperado, ya que este método es una mejora del anterior.

Una vez se ha validado el comportamiento del modelo de referencia, se aplican las técnicas de estimación de incertidumbre previamente descritas sobre las predicciones del *Extended Isolation Forest* de referencia, con el objetivo de evaluar su capacidad para cuantificar la confianza en las detecciones de anomalías, comenzando por los métodos de predicción conformal.

4.2.5. Medición de incertidumbre para *Extended Isolation Forest*

Predicción *split-conformal* y *cross-conformal*

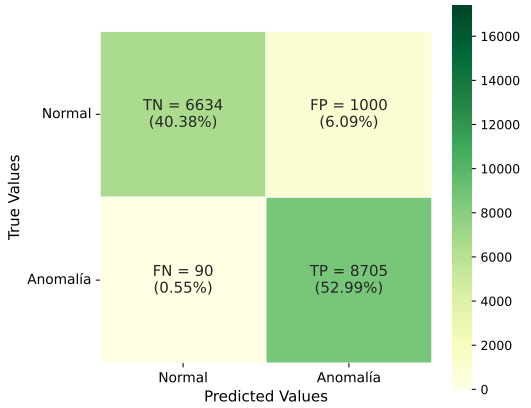
Al igual que ocurre con el *Isolation Forest*, no es viable aplicar los métodos basados en Jackknife debido a su alta complejidad computacional y espacial, por lo que emplean exclusivamente los métodos *split-conformal*, CV_{AD} y CV+_{AD}.

La configuración empleada para estas metodologías es la misma a la que se empleó para el algoritmo anterior, con 4000 observaciones para el conjunto de calibración necesario para la técnica *split-conformal* y 20 particiones para la validación cruzada de CV_{AD} y CV+_{AD}. La configuración de parámetros de los *Extended Isolation Forest* entrenados es la misma que para el modelo de referencia. Tras el entrenamiento de los modelos requeridos para cada método, se procede a calcular los p -valores asociados a cada observación x_i del conjunto test, y a la resolución del siguiente contraste de hipótesis:

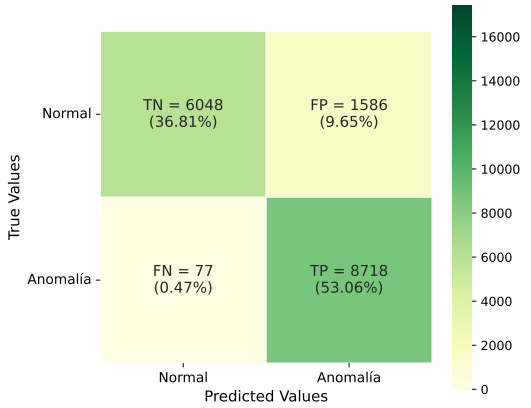
$$\begin{cases} H_{0,i} : x_i \text{ es una observación normal} \\ H_{1,i} : x_i \text{ es una observación anómala.} \end{cases}$$

Siguiendo el procedimiento aplicado para el *Isolation Forest*, se aplica el método de Storey y se lleva a cabo la corrección de los p -valores de acuerdo a este. A continuación, se lleva a cabo una clasificación de las observaciones del conjunto test en normales o anómalos, resolviendo el contraste anterior mediante los p -valores obtenidos. De nuevo, se controla la tasa de falsos descubrimientos para los niveles 0,10 y 0,15. En la Figura 4.12 se exponen seis matrices de confusión, correspondientes a los tres métodos de medición de incertidumbre y a ambas FDRs. Al igual que ocurre en la sección anterior, permitir una FDR mayor tiene como consecuencia un aumento considerable de los falsos positivos, y, para este algoritmo, la mejora en la detección de comportamientos anómalos es mucho más sutil, sugiriendo que este incremento no es adecuado. En la Tabla 4.6 se muestran las métricas de evaluación calculadas para estas clasificaciones. También se presenta la FDR observada, estando ligeramente mejor controlada en este caso que para el *Isolation Forest*. Se observa que todas las variantes mantienen niveles de sensibilidad muy altos, con valores próximos o superiores a 0.99 en todos los casos, lo que sugiere que las técnicas conservan la capacidad del modelo para identificar comportamientos anómalos. Sin embargo, al igual que ocurría con el modelo *Isolation Forest*, este aumento en sensibilidad conlleva una reducción de la precisión y la especificidad, especialmente cuando se eleva el nivel de FDR controlado.

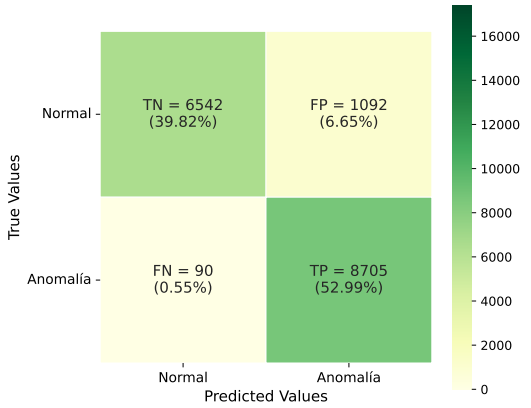
Entre los métodos considerados, *cross-conformal* $CV_{AD}+$ destaca por mantener un buen equilibrio entre sensibilidad y precisión, con una puntuación F1 de 0,9439 y AUC ROC de 0,9331 al controlar la FDR al nivel 0,10. Por otra parte, *split-conformal* y *cross-conformal* CV_{AD} también muestran un buen comportamiento, aunque su precisión y especificidad se ven más afectadas al aumentar la FDR a 0,15. En términos generales, puede concluirse que el modelo *Extended Isolation Forest* se comporta de forma estable al incorporar técnicas de medición de incertidumbre de predicción conformal, manteniendo su capacidad de detección de anomalías. La posibilidad de emplear distintas configuraciones permite adaptar el comportamiento del sistema a los requisitos específicos del proyecto en el que



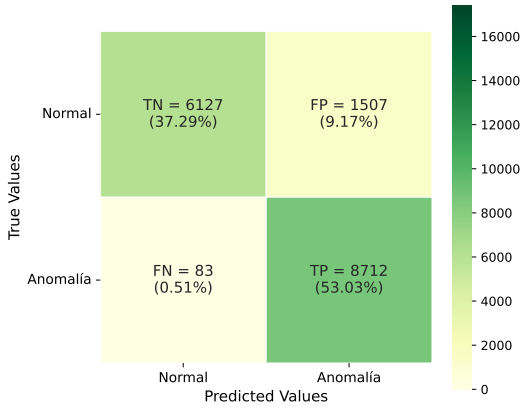
(a) *split-conformal* controlando la FDR al nivel 0.10



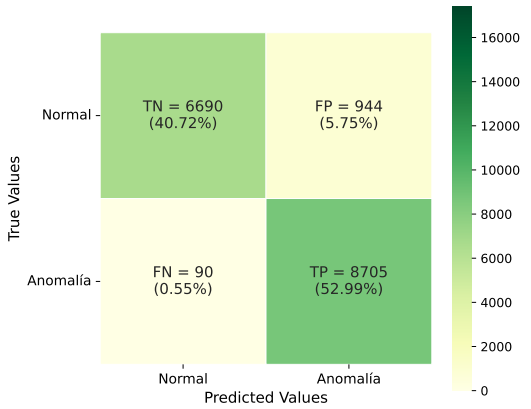
(b) *split-conformal* controlando la FDR al nivel 0.15



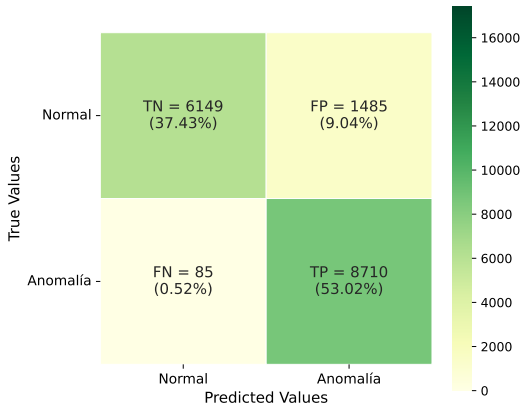
(c) CV_{AD} controlando la FDR al nivel 0.10



(d) CV_{AD} controlando la FDR al nivel 0.15



(e) $CV_{AD}+$ controlando la FDR al nivel 0.10



(f) $CV_{AD}+$ controlando la FDR al nivel 0.15

Figura 4.12: Matrices de confusión obtenidas a partir de los p-valores obtenidos al aplicar técnicas de predicción conformal al modelo *Extended Isolation Forest*. Se muestran los resultados que resultan de controlar la FDR a dos niveles: 0.10 (columna izquierda) y 0.15 (columna derecha).

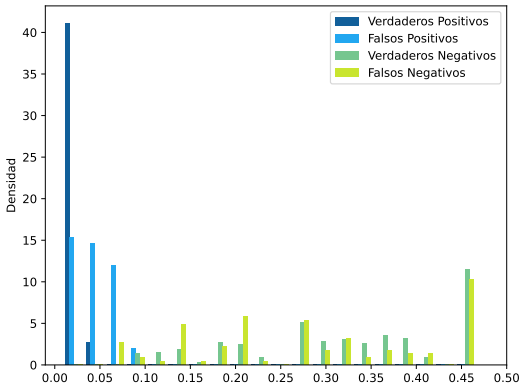
se despliegue. Por ejemplo, priorizando una alta sensibilidad en escenarios donde es crucial detectar la mayoría de las amenazas, o teniendo un mayor control sobre la tasa de falsos positivos cuando se desea minimizar la carga de revisiones manuales o evitar falsas alarmas (controlando una menor FDR). En cualquier caso, todas las configuraciones consideradas ofrecen garantías estadísticas sobre el control de la tasa de falsos descubrimiento, lo que refuerza su aplicabilidad práctica en contextos reales.

Comparando estos resultados con los obtenidos para el *Isolation Forest*, para cada uno de los métodos (*split-conformal*, CV_{AD} y $CV_{AD}+$), la combinación de estos con el *Extended Isolation Forest* permite preservar o mejorar las distintas métricas calculadas respecto al mismo método aplicado sobre el *Isolation Forest*. Además, el *Extended Isolation Forest* presenta una mayor estabilidad ante cambios en la FDR que el *Isolation Forest*, lo cual se refleja, por ejemplo, en la precisión de CV_{AD}, que pasa de 0,8885 a 0,8525 en el *Extended Isolation Forest* (una caída moderada), mientras que para el *Isolation Forest* lo hace de 0,9127 a 0,8601, una bajada más pronunciada en términos relativos.

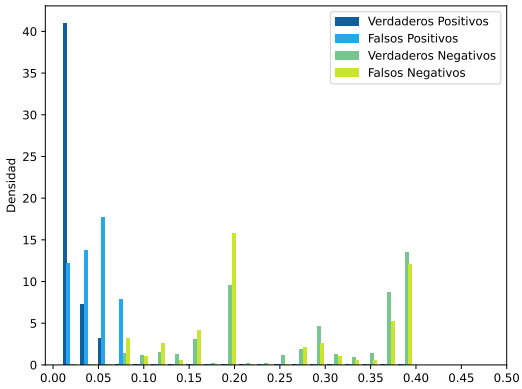
Método	**FDR**	**FDR$_{obs}$**	**Exactitud**	**Precisión**	**Sensibilidad**	**Especificidad**	**Puntuación F1**	**AUC ROC**
Modelo de referencia	–	–	0.9463	0.9173	0.9889	0.8973	0.9517	0.9431
split-conformal	0.10	0.1030	0.9337	0.8970	0.9898	0.8690	0.9411	0.9294
	0.15	0.1539	0.8988	0.8461	0.9912	0.7922	0.9129	0.8917
CV_{AD}	0.10	0.1115	0.9281	0.8885	0.9898	0.8570	0.9364	0.9234
	0.15	0.1475	0.9032	0.8525	0.9906	0.8026	0.9164	0.8966
$CV_{AD}+$	0.10	0.0978	0.9371	0.9022	0.9898	0.8763	0.9439	0.9331
	0.15	0.1457	0.9044	0.8543	0.9903	0.8055	0.9173	0.8979

Tabla 4.6: Comparación de métricas de evaluación para los distintos métodos de predicción conformal de detección de anomalías aplicados al modelo *Extended Isolation Forest*, bajo dos niveles de control de la tasa de falsos descubrimientos (FDR). Se incluye tanto la FDR controlada como la FDR observada tras aplicar cada técnica.

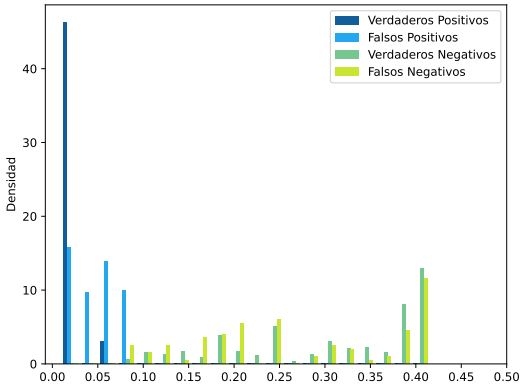
Al igual que se hizo para el *Isolation Forest*, se lleva a cabo un análisis de los p-valores corregidos obtenidos para el conjunto test, en función de si las observaciones asociadas a estos son verdaderos positivos, falsos positivos, verdaderos negativos o falsos negativos. Para ello, se estudian los histogramas presentados en la Figura 4.13, que muestran la distribución de los p-valores corregidos obtenidos tras aplicar las técnicas *split-conformal* (4.13a), CV_{AD} (4.13b) y $CV_{AD}+$ (4.13c). Para los tres métodos se observa la evidente concentración de verdaderos positivos (azul oscuro) en el rango bajo de p-valores, indicando la presencia de fuertes evidencias significativas de que estas observaciones son anómalas. Por otra parte, los falsos positivos también se encuentran en un rango de p-valores bajos, pero ligeramente más elevados que el de los verdaderos positivos, pudiéndose apreciar una diferenciación entre ambas clases. En cuanto a los verdaderos y falsos negativos, estos se encuentran distribuidos de forma más uniforme en el intervalo [0.1,0.5], reflejando la ausencia de evidencias significativas de que se tratan de anomalías. Comparando los tres métodos, $CV+_{AD}$ genera una mayor separación entre los verdaderos y los falsos positivos, lo que sugiere un mejor comportamiento, seguido de *split-conformal*, al igual que



(a) Histograma de los p-valores corregidos obtenidos mediante el método *split-conformal*.



(b) Histograma de los p-valores corregidos obtenidos mediante el método CV_{AD}.



(c) Histograma de los p-valores corregidos obtenidos mediante el método $CV+_{AD}$.

Figura 4.13: Distribución de los p-valores corregidos para cada técnica de predicción conformal, diferenciando observaciones según la clasificación obtenida en el modelo de referencia y su categoría real, para el método *Extended Isolation Forest*.

ocurría para el *Isolation Forest*.

De forma generalizada, el comportamiento de los p -valores obtenidos para el algoritmo *Extended Isolation Forest* es más satisfactorio que el de los producidos por el *Isolation Forest*. Por tanto, se puede concluir que la combinación de *Extended Isolation Forest* con medición de incertidumbre conformal resulta superior, tanto en rendimiento general como en estabilidad frente a distintas configuraciones de FDR. Esta mejora es posible que se deba a la capacidad del *Extended Isolation Forest* de captar relaciones más complejas entre variables y generar puntuaciones de anomalía más precisas y carentes de sesgos, lo cual facilita la calibración posterior de los métodos *split-conformal* y *cross-conformal*.

ExCeed

Tras haber aplicado los distintos métodos de predicción conformal, se procede a utilizar el método *ExCeeD* sobre el *Extended Isolation Forest* de referencia. Al igual que en el caso del *Isolation Forest*, se desea dotar al sistema de detección de anomalías de una medida de confianza adicional.

En este caso, se sigue el mismo procedimiento descrito anteriormente: se aplica el método *ExCeeD* exclusivamente sobre aquellas observaciones del conjunto de test que han sido clasificadas como anómalas por el *Extended Isolation Forest* de referencia. Se mantiene la tasa de contaminación fijada en $\delta = 0{,}05$ con el objetivo de mitigar el posible impacto de ruido o errores de etiquetado en el conjunto de entrenamiento, por lo que se emplea la Ecuación 3.1 para estimar las probabilidades. En la Figura 4.14a, se muestra la distribución de las probabilidades obtenidas. En esta queda reflejado como la aplicación del método ExCeeD no logra una adecuada separación entre verdaderos positivos y falsos positivos en términos de probabilidad de anomalía, ya que únicamente asigna probabilidades bajas a un subconjunto minoritario de los falsos positivos, otorgando al resto de observaciones detectadas como anómalas una confianza del 100 %. Este comportamiento pone de manifiesto una limitación del método para capturar de forma fiable la certeza del modelo en sus predicciones más relevantes.

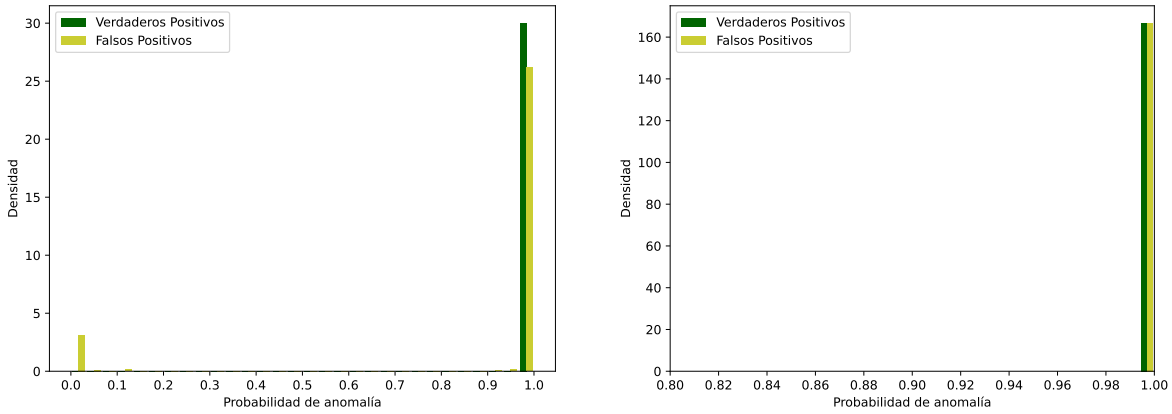
Comparando con los resultados obtenidos para el *Isolation Forest*, el método presenta un comportamiento notablemente distinto. En este caso, la totalidad de los verdaderos positivos recibe probabilidades de anomalía elevadas, lo cual no ocurre para el algoritmo más sencillo. No obstante, el problema de asignar probabilidades altas a falsos positivos sí que se mantiene. Esto indica que el método no es capaz de asignar mayor incertidumbre a las observaciones incorrectamente clasificadas, sino que simplemente otorga valores extremos sin aportar información adicional.

Finalmente, se procede a aplicar sobre este modelo el método *boostrap* desarrollado en este trabajo.

Bootstrap

Se sigue un procedimiento análogo al realizado para el *Isolation Forest*, generando 1000 remuestras *bootstrap* y se entrenan los modelos *Extended Isolation Forest* correspondientes. A continuación, se evalúan las observaciones del conjunto test clasificadas como anómalas de acuerdo al *Extended Isolation Forest* de referencia y se estima su probabilidad de haber sido clasificadas correctamente, su probabilidad de anomalía.

Las distribuciones de las probabilidades obtenidas se muestran en la Figura 4.14b. Ha sido necesario



(a) Probabilidad de anomalía calculada mediante *ExCeeD*.

(b) Probabilidad de anomalía calculada mediante *bootstrap*.

Figura 4.14: Probabilidad de anomalía calculada mediante métodos de medición de incertidumbre para cada observación detectada como anómala del conjunto test a partir del *Extended Isolation Forest* de referencia. Se distingue entre los falsos positivos (amarillo) y los verdaderos positivos (verde).

modificar la escala para mejorar la visualización, ya que, como resulta evidente, todas las observaciones clasificadas como anómalas del conjunto test reciben una probabilidad de anomalía de uno, tanto los verdaderos positivos como los falsos positivos. Esto deja en evidencia que el modelo no está proporcionando información útil que permita cuantificar la incertidumbre de sus predicciones.

En este contexto, esta metodología no discrimina entre observaciones clasificadas correctamente y aquellas que son errores del modelo, otorgando el mismo nivel de confianza, la máxima, a todas. Esto no solo invalida su utilidad para priorizar o filtrar alertas, sino que además transmite una falsa sensación de fiabilidad que puede comprometer la interpretación del sistema en un entorno real. No obstante, indica que existe una incertidumbre en el modelo muy reducida, ya que clasifica siempre en la misma categoría pese a perturbaciones en el conjunto de entrenamiento.

En resumen, estos resultados muestran que para este algoritmo y este conjunto de datos, ni *ExCeeD* ni *bootstrap* son herramientas adecuadas para llevar a cabo medición de incertidumbre, siendo más notable el caso del *bootstrap*.

Capítulo 5

Conclusiones

Para finalizar este Trabajo Fin de Máster, se presentan en este capítulo las distintas conclusiones que se han alcanzado tras realizar un análisis en profundidad de distintos métodos de medición de incertidumbre aplicados a algoritmos de detección de anomalías, en un contexto no supervisado. También se incluyen posibles líneas de trabajo futuro que podrían ampliar y complementar el trabajo realizado.

En un primer lugar, se han aplicado, con resultados satisfactorios, las técnicas de medición de incertidumbre basadas en predicción conformal. Se ha comprobado que estas, además de mantener el buen comportamiento de los modelos *Isolation Forest* y *Extended Isolation Forest*, proporcionan garantías estadísticas a través del control de la tasa de falsos descubrimientos mediante la resolución de contrastes de hipótesis. En función de las características específicas de cada contexto en el que se emplee un sistema de detección de anomalías, será necesario (y posible) controlar distintas tasas, en función del riesgo asociado que conlleve la falta de detección, o en su defecto, el exceso de falsos positivos. Por lo tanto, esta metodología proporciona una mejora sustancial frente al enfoque tradicional, en el cual se establecen tasas de contaminación de forma heurística para calcular a partir de estas los umbrales de decisión de los algoritmos.

Comparando entre los métodos de predicción conformal, en el caso del *Isolation Forest*, es el *split-conformal* el que presenta un mejor comportamiento, como se ve reflejado tanto en las métricas de evaluación como en las gráficas de los p -valores. En el caso del *Extended Isolation Forest*, además del método *split-conformal*, CV+$_{AD}$ también presenta un comportamiento muy satisfactorio, proporcionando una mayor diferenciación entre verdaderos y falsos positivos y valores para las métricas próximos a uno. La elección de cuál de estos dos métodos utilizar puede depender del tamaño del conjunto de datos disponible. Si este es muy amplio, es posible dedicar parte para el conjunto de calibración, mientras que si es más reducido, puede ser preferible emplear CV+$_{AD}$ y así aprovechar los datos de forma más eficiente, pese al incremento del coste computacional.

Por otra parte, también se han aplicado metodologías que capturan la incertidumbre mediante la estimación de la probabilidad de que cada alerta detectada esté correctamente clasificada, *ExCeeD* y la implementación *bootstrap* desarrollada en este trabajo. En el caso del conjunto de datos de tráfico de red, ninguno de los dos métodos reflejan incertidumbre en las anomalías incorrectas y certeza en las

correctas. Comparando entre ambos, *ExCeeD* proporciona medidas extremas (probabilidades de cero o uno para la mayoría de las observaciones) lo cual no aporta información ni sobre la incertidumbre del modelo ni de las predicciones. Para el conjunto de las dos lunas el método presenta un comportamiento similar, ya que otorga probabilidades o bien de cero, o bien de uno, para la extensa mayoría de los datos. Por su parte, el método *bootstrap*, pese a también presentar ciertas limitaciones, sí que es capaz de producir una distribución de probabilidades ligeramente más variada para el *Isolation Forest*. En el caso del *Extended Isolation Forest*, no produce ningún tipo de diferenciación entre las anomalías correctamente clasificadas (verdaderos positivos) y las clasificadas incorrectamente (falsos positivos), ya que asigna para todas las observaciones una probabilidad de uno. No obstante, esto sí que proporciona información sobre la robustez y la certeza del modelo, ya que todos los modelos *bootstrap* entrenados sobre las distintas remuestras *bootstrap* producen la misma clasificación. De esta forma, pese a que no se genera información útil a nivel de observaciones, existen indicaciones de que el modelo es estable frente a modificaciones en los datos de entrenamiento. Adicionalmente, en el caso del conjunto de las lunas sí que se ve reflejada la variabilidad en las probabilidades calculadas en las observaciones que se encuentran próximas a las dos lunas, reflejando la falta de certeza para clasificar estos datos.

Por último, desde el punto de vista de usabilidad, especialmente en entornos donde las salidas del sistema son empleadas por desarrolladores de modelos o perfiles sin formación específica en estadística, las probabilidades de anomalía obtenidas a través de métodos como *bootstrap* o *ExCeeD* resultan más intuitivas y accesibles que los p -valores. Además, proporcionan una métrica adicional sobre las anomalías detectadas que facilita la toma de decisiones o la priorización de alertas. Sin embargo, los p -valores proporcionan garantías estadísticas a través del control de la FDR, por lo que pueden ser preferibles en entornos críticos como es el caso de la detección de intrusiones en contextos de ciberseguridad.

5.1. Trabajo futuro

Tras exponer las distintas conclusiones a las que se ha llegado, resulta claro que existen distintas formas de continuar y extender este trabajo. La primera es analizar el comportamiento de los métodos empleados de forma más extensa, con otros conjuntos de datos reales y con anomalías auténticas (sin ser generadas en un laboratorio).

También es pertinente realizar un análisis de los resultados obtenidos, relacionándolos directamente con el conjunto de datos empleado. En el caso de los datos de tráfico de red, es de interés analizar de forma separada los resultados obtenidos en función del tipo de ataque de denegación de servicio. De esta forma, sería posible comprobar si un ataque se detecta más que otro o si sus observaciones presentan más incertidumbre. Adicionalmente, se podría estudiar en particular aquellas observaciones que fueron clasificadas incorrectamente como anómalas pero que obtuvieron, o bien probabilidades altas de anomalía o p-valores conformales próximos a cero. A través de este análisis es posible determinar si dichas observaciones presentan un comportamiento muy similar a los datos normales, explicándose así la confianza errónea de los métodos.

Otro aspecto relevante aún pendiente es emplear datos reales del proyecto GIC-TEL, que permitiría evaluar las técnicas propuestas en un entorno operativo real. También sería conveniente incorporar la

medición de incertidumbre en el sistema de detección de anomalías desplegado en la aplicación Councilbox, con el fin de evaluar su comportamiento en la práctica y establecer de qué forma proporciona el sistema la incertidumbre asociada a cada observación. En el caso de los métodos de medición de incertidumbre basados en predicción conformal, esto último se podría llevar a cabo mediante la clasificación obtenida tras resolver los contrastes de hipótesis correspondientes. En cambio, para los métodos que estiman probabilidades, las anomalías detectadas podrán ir acompañadas de estos valores, indicando así un orden de prioridad para analizar las alertas. Tras la implementación e integración de la medición de incertidumbre, también es conveniente medir el impacto que tiene esta sobre los investigadores que se dedican a investigar las alertas detectadas.

Sobre el método *bootstrap* desarrollado en este trabajo, se ha observado que en el caso del conjunto de datos sintéticos y sencillos, funciona correctamente. En cambio, para el conjunto de datos de tráfico de red, su comportamiento ha sido mejorable, proporcionando medidas de incertidumbre poco útiles a nivel de observación pero informando sobre la robustez del modelo de forma global. Por ello, es conveniente llevar a cabo un estudio de simulación, en el que se empleen distintos conjuntos de datos y algoritmos, para comprobar si esta metodología puede aportar medición de incertidumbre sobre las anomalías detectadas o si simplemente no es una buena aproximación para llevar esto a cabo.

Por último, otra posibilidad es estudiar y aplicar las distintas técnicas de medición de incertidumbre que existen a otros modelos de detección de anomalías en contextos no supervisados. Específicamente, esto es de especial relevancia para modelos tipo *autoencoders*. Estos modelos de aprendizaje profundo son ampliamente empleados en detección de anomalías, especialmente cuando se tienen datos de alta dimensionalidad. No obstante, al tratarse de algoritmos "caja negra", sus decisiones y resultados son difíciles de analizar, por lo que ampliar la confianza en estos modelos mediante la incorporación de mecanismos de medición de incertidumbre permitiría aumentar su transparencia y su fiabilidad en contextos críticos, como es el caso de la ciberseguridad.

Referencias

Aggarwal, C. C. (2017). *Outlier analysis* (2nd ed.). Springer. Descargado de https://link.springer.com/book/10.1007/978-3-319-47578-3 doi: 10.1007/978-3-319-47578-3

Angelopoulos, A. N., y Bates, S. (2022). *A gentle introduction to conformal prediction and distribution-free uncertainty quantification.* Descargado de https://arxiv.org/abs/2107.07511

Bates, S., Candès, E., Lei, L., Romano, Y., y Sesia, M. (2023). Testing for outliers with conformal p-values. *The Annals of Statistics*, *51*(1). Descargado de http://dx.doi.org/10.1214/22-AOS2244 doi: 10.1214/22-aos2244

Benjamini, Y., y Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300. Descargado 2025-04-22, de http://www.jstor.org/stable/2346101

Bennett, K., y Demiriz, A. (1998). Semi-supervised support vector machines. En M. Kearns, S. Solla, y D. Cohn (Eds.), *Advances in neural information processing systems* (Vol. 11). MIT Press. Descargado de https://proceedings.neurips.cc/paper_files/paper/1998/file/b710915795b9e9c02cf10d6d2bdb688c-Paper.pdf

Bhatia, S., Kush, N. S., Djamaludin, C., Akande, A. J., y Foo, E. (2014). Practical modbus flooding attack and detection. En *Proceedings of the twelfth australasian information security conference (aisc)* (Vol. 149, pp. 57–65). Auckland, New Zealand: Australian Computer Society, Inc.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., y Sander, J. (2000, junio). Lof: identifying density-based local outliers. *ACM SIGMOD Record*, *29*(2), 93–104. Descargado de https://dl.acm.org/doi/10.1145/335191.335388 doi: 10.1145/335191.335388

Cortes, C., y Vapnik, V. (1995, septiembre). Support-vector networks. *Mach. Learn.*, *20*(3), 273–297. Descargado de https://doi.org/10.1023/A:1022627411411 doi: 10.1023/A:1022627411411

Councilbox. (2024). *Gestión de identidad y ciberseguridad en procesos telemáticos.* Descargado de https://www.councilbox.com/gestion-de-identidad-y-ciberseguridad-en-procesos-telematicos/ (Accedido el 9 de julio de 2025)

Cover, T., y Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21-27. doi: 10.1109/TIT.1967.1053964

Dempster, A. P., Laird, N. M., y Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38. Descargado 2025-07-15, de http://www.jstor.org/stable/2984875

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*(1), 1–26. Descargado 2025-07-18, de http://www.jstor.org/stable/2958830

Ester, M., Kriegel, H.-P., Sander, J., y Xu, X. (1996, enero). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. En D. W. Pfitzner y J. K. Salmon (Eds.), *Second international conference on knowledge discovery and data mining (kdd'96). proceedings of a conference held august 2-4* (p. 226-331).

Frazão, I., Abreu, P., Cruz, T., Araújo, H., y Simões, P. (2019). *Cyber-security modbus ics dataset.* IEEE Dataport. Descargado de https://dx.doi.org/10.21227/pjff-1a03 doi: 10.21227/pjff-1a03

Gal, Y., y Ghahramani, Z. (2016, 20–22 Jun). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. En M. F. Balcan y K. Q. Weinberger (Eds.), *Proceedings of the 33rd international conference on machine learning* (Vol. 48, pp. 1050–1059). New York, New York, USA: PMLR. Descargado de https://proceedings.mlr.press/v48/gal16.html

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., ... Zhu, X. X. (2022). *A survey of uncertainty in deep neural networks.* Descargado de https://arxiv.org/abs/2107.03342

Graves, A. (2011). Practical variational inference for neural networks. En *Neural information processing systems.* Descargado de https://api.semanticscholar.org/CorpusID:14885866

Hariri, S., Kind, M. C., y Brunner, R. J. (2021, abril). Extended isolation forest. *IEEE Transactions on Knowledge & Data Engineering*, *33*(04), 1479-1489. Descargado de https://doi.ieeecomputersociety.org/10.1109/TKDE.2019.2947676 doi: 10.1109/TKDE.2019.2947676

Hennhöfer, O., y Preisach, C. (2024). *Uncertainty quantification in anomaly detection with cross-conformal p-values.* Descargado de https://arxiv.org/abs/2402.16388

Horak, T., Strelec, P., Huraj, L., Tanuska, P., Vaclavova, A., y Kebisek, M. (2021). The vulnerability of the production line using industrial iot systems under ddos attack. *Electronics*, *10*(4). Descargado de https://www.mdpi.com/2079-9292/10/4/381 doi: 10.3390/electronics10040381

Khaire, P., y Kumar, P. (2022). A semi-supervised deep learning based video anomaly detection framework using rgb-d for surveillance of real-world critical environments. *Forensic Science International: Digital Investigation*, *40*, 301346. Descargado de https://www.sciencedirect.com/science/article/pii/S2666281722000154 doi: https://doi.org/10.1016/j.fsidi.2022.301346

Kingma, D. P., y Welling, M. (2014). Auto-Encoding Variational Bayes. En *2nd international conference on learning representations, ICLR 2014, banff, ab, canada, april 14-16, 2014, conference track proceedings.*

Klaus, B., y Strimmer, K. (2024). fdrtool: Estimation of (local) false discovery rates and higher criticism [Manual de software informático]. Descargado de https://CRAN.R-project.org/package=fdrtool (R package version 1.2.18)

Kriegel, H.-P., Schubert, M., y Zimek, A. (2008). Angle-based outlier detection in high-dimensional data. En *Proceedings of the 14th acm sigkdd international conference on knowledge discovery and data mining* (p. 444–452). New York, NY, USA: Association for Computing Machinery. Descargado de https://doi.org/10.1145/1401890.1401946 doi: 10.1145/1401890.1401946

Lakshminarayanan, B., Pritzel, A., y Blundell, C. (2017). *Simple and scalable predictive uncertainty estimation using deep ensembles.* Descargado de https://arxiv.org/abs/1612.01474

Laxhammar, R. (2014). *Conformal anomaly detection: Detecting abnormal trajectories in surveillance applications* (Informatics). University of Skövde, Sweeden.

Liu, F. T., Ting, K. M., y Zhou, Z.-H. (2008). Isolation forest. En *2008 eighth IEEE international conference on data mining* (p. 413-422). doi: 10.1109/ICDM.2008.17

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms.* Copyright Cambridge University Press.

Modbus Organization. (2012). *Modbus tcp/ip protocol specification.* Application-layer messaging protocol developed for Industrial Control Systems. Descargado 2025-06-24, de https://www.modbus.org/docs/Modbus_Application_Protocol_V1_1b3.pdf (Se emplea un encabezado MBAP de 7 bytes sobre TCP/IP, con modelo cliente-servidor y sin CRC adicional, ya que TCP y Ethernet ofrecen mecanismos de verificación de errores)

National Institute of Standards and Technology. (s.f.). *Programmable logic controller (plc).* https://csrc.nist.gov/glossary/term/programmable-logic-controller. (Accessed: 2025-06-24)

Ortega-Fernandez, I. (2024). *Machine learning approaches and explainability for real-time cyberattack detection* (Tesis Doctoral). Descargado de http://hdl.handle.net/11093/7256

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, *12*(Oct), 2825–2830.

Perini, L., Vercruyssen, V., y Davis, J. (2021). Quantifying the confidence of anomaly detectors in their example-wise predictions. En F. Hutter, K. Kersting, J. Lijffijt, y I. Valera (Eds.), *Machine learning and knowledge discovery in databases* (pp. 227–243). Cham: Springer International Publishing.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, *43*(3-4), 353-360. Descargado de https://doi.org/10.1093/biomet/43.3-4.353 doi: 10.1093/biomet/43.3-4.353

R Core Team. (2021). R: A language and environment for statistical computing [Manual de software informático]. Vienna, Austria. Descargado de https://www.R-project.org/

Saetta, E., Tognaccini, R., y Iaccarino, G. (2024). Uncertainty quantification in autoencoders predictions: Applications in aerodynamics. *Journal of Computational Physics*, *506*, 112951. Descargado de https://www.sciencedirect.com/science/article/pii/S0021999124002006 doi: https://doi.org/10.1016/j.jcp.2024.112951

Sakurada, M., y Yairi, T. (2014, diciembre). Anomaly detection using autoencoders with nonlinear dimensionality reduction. En *Proceedings of the mlsda 2014 2nd workshop on machine learning for sensory data analysis* (p. 4–11). Gold Coast Australia QLD Australia: ACM. Descargado de https://dl.acm.org/doi/10.1145/2689746.2689747 doi: 10.1145/2689746.2689747

Salakhutdinov, R., y Mnih, A. (2008). Bayesian probabilistic matrix factorization using markov chain monte carlo. En *Proceedings of the 25th international conference on machine learning* (p. 880–887). New York, NY, USA: Association for Computing Machinery. Descargado de https://doi.org/10.1145/1390156.1390267 doi: 10.1145/1390156.1390267

Shyu, M.-L., Chen, S.-C., Sarinnapakorn, K., y Chang, L. (2003). A novel anomaly detection scheme based on principal component classifier..

Storey, J. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(3), 479-498. Descargado de https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00346 doi: https://doi.org/10.1111/1467-9868.00346

Storey, J. D., y Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, *100*(16), 9440-9445. Descargado de https://www.pnas.org/doi/abs/10.1073/pnas.1530509100 doi: 10.1073/pnas.1530509100

Stracuzzi, D., Chen, M., Darling, M., Peterson, M., y Vollmer, C. (2017). *Uncertainty quantification for machine learning.* Office of Scientific and Technical Information (OSTI). Descargado de http://dx.doi.org/10.2172/1733262

Tukey, J. W. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, *29*(2), 614.

Tyralis, H., y Papacharalampous, G. (2024). A review of predictive uncertainty estimation with machine learning. *Artificial Intelligence Review*, *57*, 94. doi: https://doi.org/10.1007/s10462-023-10698-8

van Engelen, J. E., y Hoos, H. H. (2020, febrero). A survey on semi-supervised learning. *Machine Learning*, *109*(2), 373–440. Descargado de https://doi.org/10.1007/s10994-019-05855-6 doi: 10.1007/s10994-019-05855-6

Van Rossum, G., y De Boer, J. (1991). Interactively testing remote servers using the python programming language. *CWI quarterly*, *4*(4), 283–303.

Vovk, V., Gammerman, A., y Shafer, G. (2005). *Algorithmic learning in a random world.* Springer New York, NY. doi: https://doi.org/10.1007/b106715

Wenchong, H., Zhe, J., Tingsong, X., Zelin, X., y Yukun, L. (2024). *A survey on uncertainty quantification methods for deep learning.* Descargado de https://arxiv.org/abs/2302.13425

Yong, B. X., y Brintrup, A. (2022). Bayesian autoencoders with uncertainty quantification: Towards trustworthy anomaly detection. *Expert Systems with Applications*, *209*, 118196. Descargado de https://www.sciencedirect.com/science/article/pii/S0957417422013562 doi: https://doi.org/10.1016/j.eswa.2022.118196

Yong, B. X., Fathy, Y., y Brintrup, A. (2020). Bayesian autoencoders for drift detection in industrial environments. En *2020 ieee international workshop on metrology for industry 4.0 & iot* (p. 627-631). doi: 10.1109/MetroInd4.0IoT48571.2020.9138306