CONSTRUCCIÓN DE UN MODELO QUE PERMITA IDENTIFICAR FALLAS EN GENERADORES DE CENTRALES HIDROELÉCTRICAS

Adrián Rodríguez Amaya
David Andrés Pérez Aponte

Nota de Aceptación

Certificamos que el presente Trabajo de Grado Satisface, en alcances y calidad, todos los requisitos que demanda un Trabajo de Grado de Maestría.

María Constanza Pabón
Director

Julián Gil
Jurado

Eugenio Tamura
Jurado

Aprobado en cumplimiento de los requisitos exigidos por la Pontificia Universidad Javeriana Cali, para optar el título de Magister en Ciencia de Datos.

HERNÁN CAMILO ROCHA NIÑO Ph. D.
Decano Facultad de Ingeniería y Ciencias

JUAN CARLOS MARTÍNEZ ARIAS
Director Posgrados de Ingeniería y Ciencias

Santiago de Cali, 28 junio de 2023.

Acta de Correcciones al Documento de Trabajo de Grado

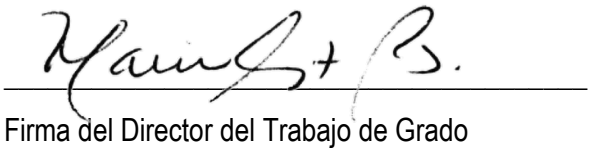
Santiago de Cali, 28 junio de 2023

Autor: Adrián Mauricio Rodríguez Amaya; David Andrés Pérez Aponte ID: 8972790; 8972743

Título del Trabajo de Grado: "CONSTRUCCIÓN DE UN MODELO QUE PERMITA IDENTIFICAR FALLAS EN GENERADORES DE CENTRALES HIDROELÉCTRICAS**"**

Director: María Constanza Pabón

Como indica el artículo 2.13 de las Directrices para Trabajo de Grado de Maestría, he verificado que el estudiante indicado arriba ha implementado todas las correcciones que los Jurados del Proyecto de Trabajo de Grado definieron que se efectuaran, como consta en el Acta de Evaluación correspondiente.



Firma del Director del Trabajo de Grado

Santiago de Cali, 29 de mayo del 2023

Doctora
Gloría Inés Alvarez V.
Directora Maestría en Ciencia de Datos
Facultad de Ingeniería y Ciencias
Pontificia Universidad Javeriana de Cali

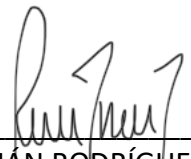
Asunto: Presentación para evaluación del proyecto aplicado

Cordial Saludo

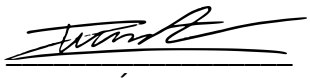
Con el fin de cumplir con los requisitos exigidos por la Universidad para optar por el título de Magíster en Ciencia de Datos, nos permitimos presentar a su consideración el proyecto denominado "CONSTRUCCIÓN DE UN MODELO QUE PERMITA IDENTIFICAR FALLAS EN GENERADORES DE CENTRALES HIDROELÉCTRICAS", el cual fue realizado por el (los) estudiante (s) ADRIÁN MAURICIO RODRÍGUEZ AMAYA, DAVID ANDRÉS PÉREZ APONTE con código (s) 8972790, 8972743 pertenecientes a la Maestría en Ciencia de Datos, bajo la dirección de María Constanza Pabón.

El suscrito director del Proyecto Aplicado autoriza para que se proceda a hacer la evaluación de este proyecto, toda vez que ha revisado cuidadosamente el documento y avala que ya se encuentra listo para ser presentado y sustentado oficialmente.

Atentamente,



ADRIÁN RODRÍGUEZ AMAYA
C.C. 1101813866 de Ovejas-Sucre



DAVID PÉREZ APONTE
C.C. 1038404254 de Marinilla

MARIA CONSTANZA PABÓN
C.C. 34.559.226 de Popayán

Documentación anexa:
Resumen del Proyecto Aplicado en formato digital (máximo 1 página).
Una copia digital (PDF) del documento del proyecto aplicado

ACUERDO DE CONFIDENCIALIDAD SUSCRITO ENTRE ENEL COLOMBIA S.A. ESP y PONTIFICIA UNIVERSIDAD JAVERIANA - CALI

Entre los suscritos a saber: **(i) HECTOR LIZCANO TARAZONA**, mayor de edad, vecino de Bogotá D.C, identificado con cédula de ciudadanía/extranjería No. 91.258.618 de Bucaramanga, actuando como **HEAD OF HYDRO POWER PLANT RÍO BOGOTÁ COL**, en nombre y representación de ENEL COLOMBIA S.A. ESP, sociedad constituida según escritura pública No.3480 otorgada en la Notaría 18 del Círculo de Bogotá D.C.; inscrita en la Cámara de Comercio de Bogotá bajo el No. 01151755 del 17 de agosto de 2007, y que para todos los efectos del presente acuerdo se denomina ENEL COLOMBIA, y **(ii) PONTIFICIA UNIVERSIDAD JAVERIANA - CALI**, representada por **LUIS FELIPE GÓMEZ RESTREPO, S.J.**, mayor de edad, identificado(a) con cédula de ciudadanía número 79.152.231 de Bogotá, y quién en adelante y para todos los efectos del presente documento se denomina **PUJ-CALI,** han acordado celebrar el presente Acuerdo de Confidencialidad (en adelante el "Acuerdo"), previas las siguientes cláusulas:

CONSIDERACIONES:

1. ENEL COLOMBIA y PUJ-CALI, en adelante las partes, en beneficio mutuo desean revelarse determinada información verbal o escrita, en general de carácter mercantil que puede incluir entre otros, planes de proyectos, inversión y desarrollo, información técnica y financiera, planes de productos y servicios, información de precios, análisis y proyecciones, especificaciones, diseños, dibujos, software, datos, prototipos know how y otra información de negocios o técnica, para desarrollo del proyecto de grado denominado "CONSTRUCCIÓN DE UN MODELO QUE PERMITA IDENTIFICAR FALLAS EN GENERADORES DE CENTRALES HIDROELÉCTRICAS" (en adelante el Proyecto), en virtud del programa académico: MAESTRÍA EN CIENCIA DE DATOS, al cual se encuentra matriculado el señor ADRIAN MAURICIO RODRIGUEZ AMAYA, quien a su vez actúa como colaborador de ENEL COLOMBIA.

2. El presente Acuerdo de Confidencialidad tiene como finalidad establecer el uso y la protección que darán Las Partes a la información que se han entregado y se entregarán en desarrollo del Proyecto.

3. De conformidad con lo anterior, Las Partes de este acuerdo se someten a las siguientes:

CLÁUSULAS

PRIMERA: DEFINICIONES

1. **INFORMACIÓN CONFIDENCIAL:** Información Confidencial significa cualquier información acerca de productos, nuevas tecnologías, modelos de negocios, información técnica, financiera, comercial, de mercado, estratégica y cualquiera otra relacionada con las operaciones de negocios presentes y futuros de las partes**,** de sus socios o accionistas, incluyendo, sus matrices, filiales, subsidiarias,

afiliadas o sucursales, que haya sido o sea entregada o comunicada en virtud del desarrollo del Proyecto.

Adicionalmente, cualquier información suministrada, previa a la celebración del presente Acuerdo, se considerará como confidencial y estará sujeta a los términos del mismo.

Dicha información podrá ser escrita, verbal, en medio magnético, en cualquier forma, tangible o no. Incluye entre otras, procesos, proyectos, esbozos, fotografías, plantas, diseños, bases de datos, directorios, tablas, conceptos de producto, especificaciones, muestras (incluso de equipos y herramientas), informativos, nombres de clientes, vendedores y/o distribuidores, información de precios, definiciones de mercado, invenciones e ideas o, en cualquier otra forma y en cualquier otro medio, accesos remotos y/o físicos a máquinas, servicios, contraseñas, usuarios y toda información que se suministren o divulguen las partes en ejecución del Proyecto.

De igual forma son constitutivos de información confidencial, todos los análisis, recopilaciones, datos, estudios, memorandos, informes y documentos, en cualquier forma y en cualquier medio, elaborados por las partes o sus directores, funcionarios, empleados o representantes que se deriven de, o se relacionen con la información que trata el párrafo anterior; o que contengan o se basen en todo o en parte, en dicha información; al igual que cualquier idea, concepto, know-how, conocimiento o técnica relacionada con las actividades propias del negocio de **LA PARTE REVELADORA;** contenidas en la información que trata el mencionado párrafo, y toda la información que permanezca en la memoria de los empleados de **LA PARTE RECEPTORA** que han tenido acceso a la Información Confidencial bajo el presente Acuerdo.

PARÁGRAFO PRIMERO. - Dado lo anteriormente expuesto, **LA PARTE RECEPTORA** se compromete a mantener la información suministrada por **LA PARTE REVELADORA** bajo la más estricta confidencialidad comprometiéndose a no disponerla, publicarla, o revelarla a terceras partes no autorizadas expresamente por **LA PARTE REVELADORA.** En especial se obliga a no utilizar en su propio provecho o en el de terceras personas naturales o jurídicas la información a que tenga acceso en desarrollo de las mesas de trabajo del Proyecto.

PARÁGRAFO SEGUNDO: Se exceptúa de lo anterior la Información que:

(i) Haya sido de dominio público, o sea publicada sin que medie ninguna acción y/o intervención de la PARTE RECEPTORA.
(ii) Antes de revelarla estuviera en posesión legítima de la PARTE RECEPTORA.
(iii) Posteriormente a la revelación de ésta, sea legalmente recibida de un tercero que tenga derechos para distribuir la información sin notificación de ninguna restricción de su derecho a revelarla posteriormente y que no esté bajo una prohibición de carácter contractual o legal de suministrar o divulgar dicha información.

(iv) Se revele con la aprobación previa y escrita de la PARTE REVELADORA.
(v) La revelación y/o divulgación de la información se realice en desarrollo o por mandato de una ley, decreto o sentencia u orden de autoridad competente en ejercicio de sus funciones legales. En este caso, la parte obligada a divulgar la información confidencial se obliga a avisar inmediatamente haya tenido conocimiento de esta obligación a la otra parte del presente Acuerdo, para que pueda tomar las medidas necesarias para proteger su información confidencial, y de igual manera se compromete a tomar las medidas necesarias para atenuar los efectos de tal divulgación.

2. **PARTE REVELADORA:** Se constituye en **PARTE REVELADORA**, cualquiera de LAS PARTES que suministre a la otra información confidencial.

3. **PARTE RECEPTORA:** Se constituye en **PARTE RECEPTORA**, cualquiera de Las Partes que reciba de la otra información confidencial.

4. **ENCARGADO:** es **PUJ-CALI**, quien, de forma independiente, acepta y acuerda realizar el Tratamiento a los Datos Personales de los Titulares por cuenta de EL RESPONSABLE, en concordancia con las instrucciones que EL RESPONSABLE le imparta y de acuerdo con los términos de este Acuerdo. Para los efectos del presente documento se entenderá Encargado la PARTE RECEPTORA.

5. **RESPONSABLE(S):** es **ENEL COLOMBIA**, persona**(s)** jurídica**(s)** que, de forma independiente, determinará**(n)** la finalidad y la forma como se hará el Tratamiento de los Datos Personales. Para los efectos del presente documento se entenderá**(n)** Responsable**(s)** la**(s)** PARTE REVELADORA**(S)**.

6. **LEY APLICABLE DE PROTECCIÓN DE DATOS PERSONALES:** se refiere a cualquier norma de protección de datos personales, privacidad o normas relacionadas, referentes a la recolección, tratamiento, transferencia o transmisión de los Datos Personales de los Titulares de acuerdo con este Acuerdo, incluyendo pero sin limitarse a la Ley 1581 de 2012, Decreto Único Reglamentario 1074 de 2015, y la Circular Externa 005 de 2017 de la Superintendencia de Industria y Comercio y aquellas que las modifiquen, adicionen o deroguen, así como las demás normas concordantes vigentes.

7. **POLÍTICA DE TRATAMIENTO:** es la política de tratamiento de los Datos Personales del RESPONSABLE, y que se adjunta como Anexo al presente Acuerdo.

8. **TITULAR:** es una persona natural determinada o identificable, cuyos Datos Personales están sujetos al Tratamiento de los mismos por parte de EL RESPONSABLE o EL ENCARGADO.

9. **TRANSMISIÓN**: es el Tratamiento de Datos Personales que implica la comunicación de los mismos dentro o fuera del territorio de la República de Colombia, con el objeto de permitir a EL RESPONSABLE y a EL ENCARGADO realizar el Tratamiento de los Datos Personales.

10. **TRATAMIENTO:** será cualquier operación o conjunto de operaciones realizadas sobre los Datos Personales, consistentes en la recolección, grabación, organización, almacenamiento, consulta, uso, divulgación por transmisión, actualización, transferencia, diseminación o cualquier otra forma y/o el bloqueo, eliminación o destrucción de los mismos.

SEGUNDA.- OBJETO. Las Partes hemos convenido celebrar el presente Acuerdo de Confidencialidad que tiene como finalidad establecer los términos que rigen el uso y la protección que le darán a la Información Confidencial que se revelen en desarrollo del Proyecto.

En este sentido, Las Partes se obligan a no revelar, divulgar, exhibir, mostrar, comunicar, utilizar y/o emplear la Información Confidencial con persona natural o jurídica, en su favor o en el de terceros, y en consecuencia se obligan a mantenerla de manera confidencial y privada y a proteger dicha información para evitar su divulgación no autorizada, ejerciendo sobre esta el mismo grado de diligencia que utilizan para proteger información confidencial de su propiedad.

PARÁGRAFO.- Las Partes admiten y consienten que toda la Información Confidencial que aquí se trata, es propiedad exclusiva de quien revela la información y que ésta se da a conocer únicamente con el propósito de facilitar el desarrollo y ejecución del Proyecto.

TERCERA.- USO DE LA INFORMACIÓN Respecto al uso que se le debe dar a la Información Confidencial, Las Partes acuerdan las siguientes reglas:

1) Toda la Información Confidencial de **LA PARTE REVELADORA** será en todo momento propiedad de la misma.
2) **LA PARTE RECEPTORA** se obliga a mantener en estricta reserva la Información Confidencial de **LA PARTE REVELADORA** y no podrá divulgarla, usarla, exhibirla, mostrarla, utilizarla, emplearla, explotarla y/o transmitirla, en beneficio propio o de un tercero o en perjuicio de **LA PARTE REVELADORA** y, en consecuencia, se obliga a mantenerla de manera confidencial y privada y a proteger dicha información para evitar su divulgación o Tratamiento no autorizado, ejerciendo sobre esta el mismo grado de diligencia que utiliza para proteger información confidencial de su propiedad, y en todo caso, por lo menos el grado de diligencia que utilizaría un buen empresario en la protección de su información confidencial.
3) La Información Confidencial podrá usarse, explotarse y/o transmitirse única y exclusivamente para el desarrollo del Proyecto y del presente Acuerdo o cuando medie el consentimiento previo y por escrito de **LA PARTE REVELADORA** y siempre y cuando el tercero receptor de dicha información se obligue expresamente a mantener la confidencialidad y reserva sobre ella.
4) **LA PARTE RECEPTORA** utilizará los medios necesarios para que sus directivos, funcionarios, representantes, empleados y asesores (en adelante "Representantes") guarden la debida reserva sobre la Información Confidencial.
5) **LA PARTE RECEPTORA** no podrá mencionar a terceros que conoce o mantiene en su poder la Información Confidencial, salvo que exista la autorización previa y

por escrito de **LA PARTE REVELADORA** y siempre y cuando el tercero receptor de dicha información se obligue expresamente a mantener la confidencialidad y reserva sobre ella.

6) Las Partes convienen que la Información Confidencial que sea proporcionada, incluyendo aquella que haya sido suministrada antes de la firma de este Acuerdo, será conservada en estricta reserva.
7) Las Partes reconocen que la Información Confidencial obtenida o que se obtenga con ocasión del desarrollo del Proyecto y del presente Acuerdo, tiene un valor comercial, que no es del dominio público y que su divulgación podría afectar gravemente los intereses de **LA PARTE REVELADORA** y del compromiso que se pretende suscribir formalizando el presente Acuerdo.
8) **LA PARTE RECEPTORA** acuerda que no infringirá ninguno de los derechos de propiedad intelectual ni otros derechos de **LA PARTE REVELADORA** que recaigan sobre la Información Confidencial.
9) Si **LA PARTE RECEPTORA** o cualquiera de sus Representantes deben, por ley o por orden de autoridad competente, revelar alguna Información Confidencial de **LA PARTE REVELADORA**, **LA PARTE RECEPTORA** informará previamente por escrito a **LA PARTE REVELADORA** del requerimiento que le sea formulado para la revelación de la Información Confidencial y de los términos y circunstancias del mismo, de tal forma que **LA PARTE REVELADORA** pueda buscar una protección adecuada u otro recurso para mantener el amparo de la Información Confidencial.
10) La autorización de uso de la información confidencial no concede, ni expresa ni implícitamente, autorización, permiso o licencia de uso de marcas comerciales, patentes, derechos de autor o de cualquier otro derecho de propiedad industrial o intelectual.
11) **LA PARTE RECEPTORA** velará por el cumplimiento del presente acuerdo de confidencialidad, además de su pronta y oportuna divulgación hacia todas las personas involucradas en el desarrollo y ejecución del Proyecto
12) Una vez firmado el presente acuerdo **LA PARTE REVELADORA y LA PARTE RECEPTORA** acordarán los medios y condiciones para compartir la información, transportarla o transferirla.

PARÁGRAFO PRIMERO: OBLIGACIONES RESPECTO EL TRATAMIENTO DE DATOS PERSONALES

OBLIGACIONES DE LA PARTE RECEPTORA EN CALIDAD DE ENCARGADO:

1) Que el tratamiento de los datos se efectuará de conformidad con la legislación vigente, así como con los criterios, requisitos y especificaciones establecidos en el Contrato en caso de existir, las estipulaciones del presente Acuerdo y con las instrucciones que emanen de **LA PARTE REVELADORA**.
2) Que cuando obtenga por cualquier medio datos de carácter personal, se obliga a obtener la debida autorización de su titular y a informar adecuadamente sobre el uso que le dará a la información. Deberá tener soporte o prueba de esta autorización.
3) Que los datos personales a los que tenga acceso no serán aplicados ni utilizados para un fin distinto al que figura en el presente Acuerdo o a un tratamiento diferente al autorizado por su titular.

4) Que dará trámite a las consultas y reclamos que interpongan los titulares de los datos personales en los términos señalados en la normatividad vigente.
5) Que realizará oportunamente la actualización, rectificación o supresión de los datos de los titulares en los términos señalados en la normatividad vigente.
6) Que actualizará la información de los titulares reportada por **LA PARTE REVELADORA** dentro de los cinco (5) días hábiles siguientes contados a partir de su recibo.
7) Que adoptará un manual interno de políticas y procedimientos para garantizar la adecuada atención a las consultas y reclamos que interpongan los titulares.
8) Que se abstendrá de circular información que esté siendo controvertida por el titular y cuyo bloqueo haya sido ordenado por la autoridad competente en la materia.
9) Que devolverá a **LA PARTE REVELADORA** los datos de carácter personal que hayan sido objeto de tratamiento, en un plazo no mayor a quince (15) días contados desde la fecha de terminación del Contrato en caso de existir o del presente Acuerdo, lo cual será certificado por el representante legal de **LA PARTE RECEPTORA**.
10) Que destruirá cualquier documento, soporte o copia de los datos de carácter personal que hayan sido objeto de tratamiento y que no hayan podido ser objeto de devolución. No obstante, no procederá a la destrucción de los datos cuando exista una previsión legal que exija su conservación, en cuyo caso las Partes conservarán, debidamente protegidos los mencionados datos, lo cual será certificado por el representante legal de **LA PARTE RECEPTORA**.
11) Que no comunicará, ni cederá a otras personas físicas o jurídicas los datos personales que le sean suministrados con motivo de la relación jurídica y guardará la debida confidencialidad respecto del tratamiento que se le autorice.
12) Que adoptará, en el tratamiento de los datos suministrados, las medidas de índole técnica y organizativa necesarias exigidas por la normativa legal que al respecto resulte de aplicación, de forma que se garantice la seguridad de los datos de carácter personal y se evite su alteración, pérdida, tratamiento o acceso no autorizado, habida cuenta del estado de la tecnología, la naturaleza de los datos almacenados y los riesgos a que están expuestos, ya provengan de la acción humana, del medio físico o natural. Las medidas abarcarán, a título enunciativo, hardware, software, procedimientos de recuperación, copias de seguridad y datos extraídos de datos personales en forma de exhibición en pantalla o impresa.
13) Que se compromete a tener autorización del manejo, tratamiento y circulación de los datos personales de cada uno de sus empleados y trabajadores con el fin de verificar el cumplimiento de las obligaciones jurídico laborales, de seguridad social, de prevención de riesgos laborales y demás.
14) Que en el caso que para la prestación del servicio fuera necesaria la realización de alguna transferencia internacional de datos, el Proveedor se obliga a informar a **LA PARTE REVELADORA** con carácter previo y con la suficiente antelación para que ésta pueda solicitar las correspondientes autorizaciones, sin las cuales, **LA PARTE RECEPTORA** no podrá realizar dichas transferencias.
15) Dar tratamiento, a nombre de **LA PARTE REVELADORA**, a los datos personales conforme a los principios que los tutelan.

16) Salvaguardar la seguridad de las bases de datos en los que se contengan datos personales.
17) Guardar confidencialidad respecto del tratamiento de los datos personales.
18) Llevar a cabo el Tratamiento de los Datos Personales de acuerdo con los términos de (i) el consentimiento otorgado por el Titular para el Tratamiento de los Datos Personales, (ii) la Política de Tratamiento de EL RESPONSABLE y, (iii) los principios y normas contenidas en la Ley Aplicable de Protección de Datos.
19) **LA PARTE RECEPTORA** garantizará mecanismos de para la atención de las solicitudes que determinen las autoridades colombianas en materia de protección de datos personales.
20) **LA PARTE RECEPTORA** implementará mecanismos para la atención de consultas, reclamos y la garantía del derecho de los titulares de la información personal.
21) **LA PARTE RECEPTORA** se compromete a notificar a la **PARTE REVELADORA** el o los países donde se llevará a cabo el tratamiento de los datos personales recibidos.
22) **LA PARTE RECEPTORA** se compromete a no **transferir, transmitir** y tratar los datos en países no autorizados por escrito por la **PARTE REVELADORA.**

OBLIGACIONES DE LA PARTE REVELADORA EN CALIDAD DE RESPONSABLE:

1) **LA PARTE REVELADORA** declara y garantiza que, durante el término de este Acuerdo, deberá tratar los Datos Personales aplicando los más altos estándares de confidencialidad y en cumplimiento de su Política de Tratamiento y este Contrato.
2) **LA PARTE REVELADORA** deberá notificar oportunamente a **LA PARTE RECEPTORA** acerca de cualquier cambio en la legislación que **LA PARTE REVELADORA** considere que puede afectar el Tratamiento de los Datos Personales o este Contrato.
3) **LA PARTE REVELADORA** deberá resolver oportuna y apropiadamente todas las solicitudes realizadas por **LA PARTE RECEPTORA** en relación con el Tratamiento de los Datos Personales sujetos a la Transmisión y acatar las recomendaciones, instrucciones y órdenes de la Autoridad de Protección de Datos en cumplimiento con la Ley Aplicable de Protección de Datos.
4) Solicitar y conservar prueba de la Autorización de los Titulares e informar la finalidad del Tratamiento a los Titulares.
5) Garantizar que la información que se le suministre a **LA PARTE RECEPTORA** de los Datos Personales sea conforme con la información que suministren los Titulares.
6) **LA PARTE REVELADORA** implementará mecanismos para la atención de consultas, reclamos y la garantía del derecho de los titulares de la información personal.

CUARTA. PROPIEDAD DE LA INFORMACIÓN

a). La **PARTE RECEPTORA** por este acto acusa recibo y acuerda que toda Información Confidencial de la **PARTE REVELADORA** es propiedad exclusiva de ésta, y que se revela únicamente con el propósito de facilitar la ejecución del Proyecto.

b). La Información Confidencial de la **PARTE REVELADORA** deberá ser tratada como tal y resguardada bajo este aspecto por la **PARTE RECEPTORA**, durante el término que se fija en el presente acuerdo a partir de la fecha en que se ha hecho entrega de la misma.

c) La PARTE REVELADORA se reserva el derecho de auditar el cumplimiento de los requisitos del presente acuerdo.

d). Ninguna parte adquirirá derechos de propiedad o disposición respecto de la Información suministrada por la otra parte.

QUINTA. - SEGURIDAD. Las partes garantizan que aplican las mismas medidas de seguridad razonables para evitar divulgación, fuga o uso no autorizado de información confidencial o patentada y aceptan que protegerán la Información Confidencial de cada una, de la misma manera y en el mismo grado en que protegen su propia Información Confidencial.

Se conviene que toda la Información Confidencial sea guardada por la **PARTE RECEPTORA** en un lugar con acceso limitado únicamente a los representantes o a quienes en forma razonable requieran conocer la Información Confidencial en relación con el Proyecto.

Si así lo considera, La **PARTE REVELADORA** solicitará por escrito las medidas de seguridad requeridas para el acceso y protección de la información confidencial, así como para su transporte, almacenamiento o procesamiento.

A menos que por escrito se describa otras, las medidas de seguridad mínimas para la información que tendrá que implementar la **PARTE RECEPTORA** serán las determinadas en las políticas de seguridad de la información de Enel: PL 33 Protección y Clasificación de la Información y la PL 487 Seguridad de la información.

SEXTA. CONFIDENCIALIDAD EN LA CADENA DE ABASTECIMIENTO. La **PARTE RECEPTORA** debe solicitar a La PARTE REVELADORA autorización para subcontratar actividades del proyecto. Si La **PARTE REVELADORA** autoriza la subcontratación, es obligación de La **PARTE RECEPTORA** establecer medidas para proteger la información confidencial de que trata este acuerdo. La **PARTE RECEPTORA** será responsable solidariamente con sus subcontratistas por los perjuicios que con el incumplimiento del presente acuerdo le sean causados a la **PARTE REVELADORA**.

SÉPTIMA. INCIDENTES, CONSULTAS Y RESOLUCIÓN DE CONFLICTOS SOBRE LA CONFIDENCIALIDAD DE LA INFORMACIÓN. Es una obligación de la **PARTE RECEPTORA** realizar el reporte de cualquier hecho que amenace las condiciones del presente acuerdo, entre otras el compromiso de la, seguridad o la violación de leyes aplicables a la información de propiedad de la PARTE REVELADORA dentro de los dos (2) días calendario siguientes a la ocurrencia de cualquier incidente.

La **PARTE REVELADORA** designará una persona, así como los medios necesarios para recibir los reportes, consultas o conflictos sobre el manejo de la información.

Si surgiere alguna diferencia, disputa o controversia entre LAS PARTES por razón o con ocasión del presente Acuerdo, LAS PARTES buscarán de buena fe un arreglo directo antes de acudir a la justicia ordinaria. En consecuencia, si surgiere alguna diferencia, cualquiera de LAS PARTES notificará a la otra la existencia de dicha diferencia y una etapa de arreglo directo surgirá desde el día siguiente a la respectiva notificación. Esta etapa de arreglo directo culminará a los treinta (30) días siguientes a la fecha de su comienzo.

En caso de no lograrse acuerdo en el citado término, la controversia será resuelta por los jueces de la República de Colombia.

OCTAVA. ACCESO A LA INFORMACION CONFIDENCIAL. Cuando así lo requiera La **PARTE REVELADORA**, le podrá solicitar a la **PARTE RECEPTORA** una lista de las personas autorizadas para acceder a la información confidencial, y ésta deberá enviarla dentro de los tres (3) días hábiles siguientes al requerimiento.

NOVENA. NO OTORGAMIENTO DE DERECHOS. La entrega de información, sea confidencial o no, no concede, ni expresa ni implícitamente, autorización, permiso o licencia de uso de marcas comerciales, patentes, derechos de autor o de cualquier otro derecho de propiedad industrial o intelectual. Ni este Acuerdo, ni la entrega o recepción de información, sea confidencial o no, constituirá o implicará promesa de efectuar contrato alguno por cualquiera de las partes.

DÉCIMA. DEVOLUCIÓN DE INFORMACIÓN.- En caso de terminación del Proyecto **LA PARTE REVELADORA** hará la devolución a **LA PARTE RECEPTORA** de toda la Información Confidencial que le haya sido entregada, así como de todo el material escrito que contenga o refleje cualquier Información Confidencial. **LA PARTE RECEPTORA** no mantendrá copias, extractos o reproducciones parciales o totales de la Información Confidencial. Todos los documentos, comunicados, análisis, notas, estudios, y demás escritos preparados por **LA PARTE RECEPTORA** o sus representantes, basados en la Información Confidencial serán destruidos. **LA PARTE RECEPTORA** deberá notificar por escrito a **LA PARTE REVELADORA** de tal destrucción, dentro del mismo término para la devolución de la información aquí señalada.

PARÁGRAFO.- Cualquier Información Confidencial y cualquier copia de la misma, que con base en lo establecido en esta cláusula no sea devuelta a **LA PARTE REVELADORA**, continuará sometida íntegramente a lo consagrado en este acuerdo. En caso de ser necesario que **LA PARTE RECEPTORA** conserve La Información Confidencial o una parte de la misma a efectos de la atención de disposiciones legales, esta continuará sometida íntegramente a lo consagrado en este acuerdo.

DÉCIMA PRIMERA.- ÉTICA Y ANTICORRUPCIÓN. ENEL COLOMBIA, como compañía que pertenece al Grupo Enel, declara que, y **PUJ-CALI** por la presente reconoce, que el Grupo Enel en la gestión de sus actividades comerciales y sus

relaciones, se adhiere a los principios contenidos en su Código de Ética, el Plan de Tolerancia Cero a la Corrupción, el Enel Global Compliance Program adoptado de conformidad con el Decreto Legislativo italiano 231/2001 y la Política de Derechos Humanos de la Compañía. Dichos documentos están disponibles en la dirección web: https://www.enel.com/investors/governance/internal-controls.

Además, el Grupo Enel, en el desarrollo de sus actividades comerciales y en la gestión de su relación con terceros, espera que sus contrapartes se acojan a principios equivalentes en la gestión de sus actividades comerciales y en sus demás relaciones, conforme a los mismos principios adoptados por Enel según el inciso anterior.

El Grupo Enel se adhiere al Pacto Mundial de las Naciones Unidas ("GC por sus siglas en inglés") y en cumplimiento del 10º principio del GC, se propone continuar con su compromiso de lucha contra cualquier tipo de corrupción. Por lo tanto, el Grupo Enel prohíbe recurrir a cualquier tipo de promesa, oferta o demanda de pago ilícito, en dinero u otro beneficio, con el fin de obtener una ventaja en relación con sus grupos de interés, esta prohibición se extiende a todos sus empleados. **PUJ-CALI** declara tener en cuenta los compromisos del Grupo Enel y se compromete a no prometer, ofrecer o exigir un pago ilícito en la ejecución de este Acuerdo en interés del Grupo Enel y/o en beneficio de sus empleados.

Las Partes también declaran cumplir con obligaciones legales en materia de corrupción y blanqueo de capitales; trabajo infantil y protección de la mujer; igualdad de trato, prohibición de la discriminación, abusos y acoso; libertad sindical, de asociación y de representación; trabajo forzado; salud y seguridad; protección del medio ambiente; así como el cumplimiento de las leyes aplicables, condiciones salariales, contributivas, de seguros y fiscales para todos los trabajadores contratados por cualquier motivo durante la ejecución del contrato. Se entiende que se aplican los convenios de la OIT, o la legislación vigente del país donde se deben realizar las actividades, en caso de estas ser más restrictivas.

Enel se reserva la facultad de realizar cualquier actividad de control y seguimiento encaminada a verificar el cumplimiento de las obligaciones antes mencionadas.

En caso de incumplimiento de cualquiera de dichas obligaciones durante la ejecución del presente Acuerdo, y de los principios relacionados en la presente clausula, Enel se reserva el derecho a terminar el Acuerdo y a reclamar daños y perjuicios.

Cada Parte cumplirá íntegramente con todos los requisitos legales relacionados con las Sanciones Internacionales en relación con el cumplimiento de este Acuerdo.

Cada Parte mantendrá en vigor y hará cumplir las políticas y procedimientos diseñados para asegurar el cumplimiento de las Sanciones y comunicará inmediatamente por escrito a la otra Parte cualquier cambio en las obligaciones y representaciones antes mencionadas que puedan ocurrir durante la vigencia de este Acuerdo, así como la ocurrencia de cualquier circunstancia que resulte o pueda

resultar en un incumplimiento de cualquiera de las obligaciones y representaciones mencionadas anteriormente durante la vigencia de este Acuerdo.

Ninguna de las Partes (i) contribuirá o pondrá a disposición la totalidad o parte de los ingresos del Acuerdo, directa o indirectamente, para financiar las actividades, negocios o inversiones de cualquier persona sancionada y / o no autorizada o para el beneficio de esta (ii) participar en cualquier transacción, actividad o conducta que pudiera causar que una de las Partes del acuerdo incumpla alguna Sanción.

Cada Parte declara por la presente que no es una Persona Sancionada y se compromete a no involucrar, directa o indirectamente, a ninguna Persona No Autorizada en el cumplimiento de este Acuerdo.

Cada Parte indemnizará y mantendrá indemne a la otra Parte por cualquier daño, pérdida, costo o gasto que surja o esté relacionado con la violación de las obligaciones y representaciones anteriores y cada Parte también podrá rescindir este Acuerdo en caso de que, a partir de su ejecución, la otra Parte viole los términos de las obligaciones y representaciones establecidas anteriormente.

En tal caso, cada Parte podrá notificar la rescisión a la otra Parte indicando el motivo de la misma y las Partes podrán negociar de buena fe para mitigar en la medida de lo posible cualquier pérdida o daño en relación con las Sanciones o que surjan de ellas. A falta de dicho acuerdo, dentro de los 30 días siguientes a la notificación de terminación, este Acuerdo se dará por terminado automáticamente y cada Parte renunciará a cualquier reclamo, acción o petición en relación con las Sanciones o que surjan de ellas, sujeto a cualquier otro recurso que pudiera tener en virtud de la ley o en virtud del contrato, que surja de cualquier otra obligación incumplida en virtud del Acuerdo.

DÉCIMA SEGUNDA.- INTEGRIDAD DEL ACUERDO.- Ninguna modificación o enmienda de este Acuerdo será efectiva salvo que la misma sea pactada por escrito por las partes.

DÉCIMA TERCERA.- CESIÓN.- Ninguna de las partes podrá ceder ni transmitir total o parcialmente los derechos y obligaciones derivados de éste Acuerdo, sin el previo consentimiento por escrito de la otra parte.

DÉCIMA CUARTA.- RENUNCIA.- Las obligaciones de las partes y los derechos que este convenio confiere a cada una de ellas, no serán considerados como renunciables, en virtud de prácticas o costumbres contrarias. La tolerancia de una de las partes en soportar el incumplimiento de cualquier obligación a cargo de la otra, no podrá ser considerada como aceptación del hecho tolerado, ni como precedente para su repetición; tampoco impedirá o limitará el derecho de la parte cumplida de hacer valer todas y cada una de las disposiciones de conformidad con los términos de este Acuerdo.

DÉCIMA QUINTA.-NOTIFICACIONES.- Todas las notificaciones y comunicaciones relacionadas con el presente Acuerdo se harán por escrito y se enviarán a las siguientes direcciones:

ENEL COLOMBIA
Atención: HECTOR LIZCANO T
Dirección: Calle 93 # 13 - 45
Bogotá
Teléfono: 601 514 7000

POTIFICIA UNIVERSIDAD JAVERIANA CALI
Atención: GLORIA ALVAREZ VARGAS
Dirección: Calle 18 No. 118-250
Cali
Teléfono: 602 321 8200
E- Mail: galvarez@javerianacali.edu.co

DÉCIMA SEXTA.- La violación del presente Acuerdo por cualquiera de las partes dará derecho a la otra a reclamar judicialmente el resarcimiento económico de todos los daños y perjuicios que tal violación pudiera representarle, sin perjuicio de que así mismo pueda adelantar las acciones penales pertinentes, si a ellas hubiere lugar.

DÉCIMA SÉPTIMA.- VIGENCIA.- El presente Acuerdo subsistirá el tiempo que perdure el Proyecto y, en todo caso, la obligación de confidencialidad continuará vigente hasta por tres (3) años más, aún después de la terminación del mismo, por cualquier causa.

DÉCIMA OCTAVA.- DISPOSICIONES INVÁLIDAS: Si alguna de las disposiciones de este Acuerdo llegare a ser declarada ilegal, inválida o sin vigor bajo las leyes presentes o futuras, dicha disposición deberá excluirse, y este Acuerdo deberá, al alcance posible y sin destruir su propósito, ser realizado y ejecutado como si dicha disposición ilegal, inválida o sin vigor, no hubiera hecho parte del mismo y las restantes disposiciones aquí contenidas deberán conservar el mismo valor y efecto y no deben ser afectadas por la disposición declarada ilegal, inválida o sin vigor.

DÉCIMA NOVENA .- LEGISLACIÓN APLICABLE.- El presente Acuerdo y los compromisos que suscriban las partes formalizando la intención expresada a lo largo de este documento, se regirán por las leyes de la República de Colombia.

VIGÉSIMA. - VÁLIDEZ DE LA FIRMA ELECTRÓNICA/DIGITAL. Las Partes reconocen y aceptan que las firmas electrónicas o digitales plasmadas en el presente documento son confiables y vinculantes para obligarlas legal y contractualmente en relación con su contenido y tienen la misma validez y los mismos efectos jurídicos de la firma manuscrita.

Para constancia se firman dos (2) originales del mismo tenor, a los QUINCE (15) días del mes de SEPTIEMBRE de dos mil VEINTIDOS (2022).

Por ENEL COLOMBIA,

Firmado por HECTOR LIZCANO TARAZONA
el 23/09/2022 a las 11:47:13 COT

HECTOR LIZCANO TARAZONA
HEAD OF HYDRO POWER PLANT RÍO BOGOTÁ COL

Por PONTIFICIA UNIVERSIDAD JAVERIANA - CALI;

Firmado digitalmente por LUIS FELIPE GOMEZ RESTREPO
Fecha: 2022.10.05 15:20:26 -05'00'

LUIS FELIPE GÓMEZ RESTREPO, S.J.
RECTOR
REPRESENTANTE LEGAL

Iniciales de HUGO ADOLFO TAVAS VARELA
el 16/09/2022 a las 03:56:38 CEST

CONSTRUCCIÓN DE UN MODELO QUE PERMITA IDENTIFICAR FALLAS EN GENERADORES DE CENTRALES HIDROELÉCTRICAS

Adrián Rodríguez Amaya
Código 8972790

David Andrés Pérez Aponte
Código 8972743

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director(a)
María Constanza Pabón

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, MAYO 29 DE 2023

FICHA RESUMEN

TÍTULO: CONSTRUCCIÓN DE UN MODELO QUE PERMITA IDENTIFICAR FALLAS EN GENERADORES DE CENTRALES HIDROELÉCTRICAS

1. ÁREA DE TRABAJO: Sector de generación de energía eléctrica
2. TIPO DE PROYECTO: Aplicado
3. ESTUDIANTE(S): Adrián Rodríguez Amaya, David Andrés Pérez Aponte
4. CORREO ELECTRÓNICO:

   adrianrodriguez@javerianacali.edu.co, davand043@javerianacali.edu.co
5. DIRECCIÓN Y TELEFONO:
6. DIRECTOR: María Constanza Pabón
7. VINCULACIÓN DEL DIRECTOR: Facultad Ingeniería y Ciencias
8. CORREO ELECTRÓNICO DEL DIRECTOR: mcpabon@javerianacali.edu.co
9. CO-DIRECTOR: N.A.
10. GRUPO O EMPRESA QUE LO AVALA: N.A.
11. OTROS GRUPOS O EMPRESAS: N.A.
12. PALABRAS CLAVE: Aprendizaje automático, Centrales hidroeléctricas, Generadores eléctricos, Generación de energía, Mantenimiento predictivo
13. FECHA DE INICIO: 4-Jul-2022
14. DURACIÓN ESTIMADA: 12 meses
15. RESUMEN:

Las indisponibilidades no planeadas en la generación eléctrica representan multas para las empresas generadoras de energía, por parte del administrador del mercado mayorista; el modelo propuesto permitirá identificar, predecir fallas en generadoras de centrales hidroeléctricas, y ayudar a los ingenieros de operación a programar mantenimientos proactivos.

En el presente trabajo se analizaron las variables involucradas en un conjunto de datos descargados del SCADA de la operación de las unidades de generación, seleccionando los atributos más relevantes para la construcción de un modelo que identificó posibles fallas en los generadores eléctricos de una central hidroeléctrica, este conocimiento se aplicó en el contexto local para beneficio de la industria con el fin de reducir el impacto económico causado por las fallas, mediante el uso de la ciencia de datos.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE ANEXOS

INTRODUCCIÓN

Las empresas generadoras de energía eléctrica en Colombia deben cumplir con una programación de generación de energía eléctrica que es administrada por el operador XM, el no cumplimiento de esta programación de generación de energía genera penalizaciones de acuerdo con la resolución 24 de 1995 de la CREG [1], este incumplimiento se presenta generalmente por indisponibilidades no planeadas.

Predecir la ocurrencia de dichas indisponibilidades no planeadas con al menos dos horas de anticipación ayudaría a los ingenieros de operación a programar mantenimientos proactivos y no reactivos, que son los que usualmente se programan, una vez la indisponibilidad haya ocurrido.

En el presente trabajo se desarrolló un modelo predictivo para identificar fallas en generadores de centrales hidroeléctricas. La central hidroeléctrica "A" involucrada en el presente estudio se encuentra ubicada en el departamento de Cundinamarca, a una altura aproximada de 900 m.s.n.m., con una temperatura promedio aproximada al 20 °C y humedad relativa promedio de 70%.

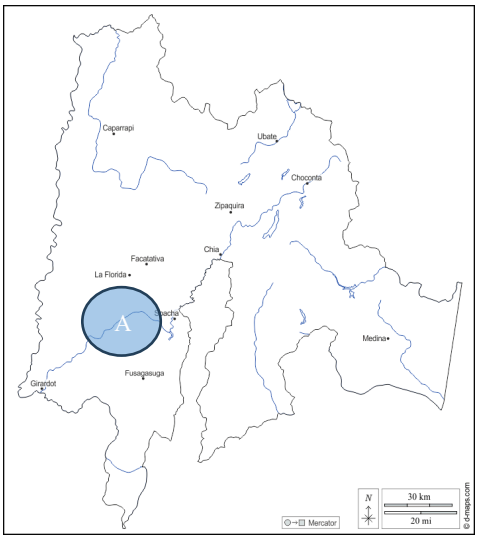


Fig. 1. Ubicación Central Hidroeléctrica A en el departamento de Cundinamarca.[1]

Esta central hidroeléctrica cuenta con tres unidades de generación ubicadas en la misma locación, de las cuales se extrajo toda la información relevante para conformar diferentes conjuntos de

[1] Tomado de: https://d-maps.com/carte.php?num_car=77997&lang=es

datos para su análisis. Las unidades de generación en mención tienen la capacidad de generar una potencia bruta de 50MW, tienen dos turbinas Pelton de eje horizontal y poseen un caudal de diseño de 6,9 metros por segundo y su velocidad nominal es de 514 revoluciones por minuto, la excitación es de tipo estática, estado sólido con tiristores controlados.

Para la construcción del modelo se tomaron los datos de las tres unidades de generación, para luego analizar las variables involucradas en las fallas ocurridas en un periodo de tiempo desde febrero de 2021 hasta septiembre de 2022, y posteriormente se realizó una correlación de variables con el fin de identificar las variables de mayor relevancia en la ocurrencia de determinada falla, también se exploraron diferentes modelo predictivos, para construir el mejor modelo que se ajuste a los datos procesados, posteriormente se realizó una evaluación del modelo de acuerdo a los indicadores de desempeño Recall y F2-Score .

Como resultado, se identificaron los tipos de fallas más recurrentes en la unidad de generación seleccionada para el proyecto, también se encontraron las relaciones entre los conjuntos de variables o señales y como afectan estas en las fallas más recurrentes de la unidad de generación, por último, se construyeron treinta y seis modelos predictivos, tomando periodos de tiempo de 2, 3, 4, 5, 9, 10, 11 y 12 horas antes de la ocurrencia de una falla; esto, utilizando tres algoritmos de aprendizaje supervisado, Random Forest, Support Vector Machines y XGBOOST.

1. DEFINICIÓN DEL PROBLEMA

1.1. PLANTEAMIENTO DEL PROBLEMA

El mercado mayorista de energía eléctrica en Colombia es administrado por XM, quien es el encargado de realizar la coordinación de la operación con los distintos operadores de red. Uno de los principales elementos de la operación es el cumplimiento del despacho económico de energía; que, de acuerdo con [2] "es la programación de la generación para cubrir la demanda esperada, de tal forma que para cada hora se utilicen los recursos de menor precio, cumpliendo con las condiciones límites que tiene el sistema como son los requisitos de reserva rodante, las inflexibilidades y las restricciones".

Para efectuar una buena coordinación de la operación con XM y asegurar la generación de energía prevista es indispensable evitar indisponibilidades. Las indisponibilidades pueden ser de tipo planeadas o de tipo no planeadas; estas últimas son generalmente causadas por operación de las protecciones eléctricas o mecánicas de los equipos frente a fallas en sus sistemas. Las indisponibilidades no planeadas están divididas en varios grupos, dentro de las cuáles las de mayor impacto son por mantenimientos correctivos no planificados, ineficiencia técnica y fallo general.

La operación de las protecciones puede ser visualizada en los SCADAs, de acuerdo con [3] se define un sistema SCADA como: "cualquier software que permita el acceso a datos remotos de un proceso y permita, utilizando las herramientas de comunicación necesarias en cada caso, el control del mismo". El software SCADA opera a nivel local en la central y desde el centro de control de la operación, con la limitación que no es posible realizar predicciones de ningún tipo. Dado lo anterior, en la mayoría de los casos las acciones por parte de los ingenieros son de tipo reactivas, es decir, se programan los mantenimientos una vez que ya ocurrió algún evento que causó indisponibilidad de las unidades de generación.

Estas indisponibilidades no planeadas causan una desviación con respecto al programa de despacho económico y redespachos, que a su vez implican penalizaciones por parte del administrador de red. De acuerdo con la resolución 24 de 1995 de la CREG (Comisión de Regulación de Energía y Gas) "Si la generación real está por fuera de la banda del 5 % aplicada al despacho programado de cada unidad o planta ofertada, el generador deberá retribuir a la cuenta por penalizaciones el valor absoluto de la diferencia entre la generación real y el despacho programado, multiplicado por el valor absoluto de la diferencia entre el precio de oferta y el precio de la bolsa."[1].

Toda la coordinación de la operación teniendo en cuenta los anteriores factores, resulta en una tarea complicada para los ingenieros de operación en el centro de control. Este escenario incrementa su complejidad en la medida que aumenta el número de unidades de generación.

Los ingenieros de operación del centro de control no cuentan con herramientas predictivas para identificación de posibles fallas; si bien el SCADA brinda facilidades para seguimiento de tendencias, no es suficiente para operar de manera óptima varias unidades de generación al tiempo.

En la actualidad, los ingenieros de operación realizan la identificación de posibles fallas de manera manual; a través de gráficas de tendencias de valores anómalos en los sistemas de las unidades de generación sobre las cuales se toman decisiones que buscarían evitar indisponibilidades por falla.

Este análisis manual no es óptimo, ya que el parque de generación hidráulico cuenta con alrededor de 32 unidades de generación y a que un solo recurso humano es insuficiente para realizar una revisión conjunta de muchas variables, que pudieran tener o no algún tipo de relación.

1.2. FORMULACIÓN DEL PROBLEMA

El área de operación tiene como reto realizar la coordinación de la operación y evitar indisponibilidades no planeadas; bajo este contexto ¿Cómo predecir ciertos tipos de fallas con la finalidad de prevenir indisponibilidades no planeadas?

SISTEMATIZACIÓN:

- ¿Existe relación entre las variables del conjunto de datos?
- ¿Cuáles son los atributos de mayor relevancia?
- ¿Qué modelos se pueden utilizar para realizar análisis predictivos de fallas?
- ¿Cómo se evaluará el desempeño del modelo predictivo diseñado?

2. OBJETIVOS DEL PROYECTO

2.1. OBJETIVO GENERAL

Construir un modelo que permita predecir fallas en generadores de centrales hidroeléctricas con la finalidad de prevenir indisponibilidades no planeadas de los activos frente al administrador del mercado eléctrico.

2.2. OBJETIVOS ESPECÍFICOS

- Analizar las variables involucradas en el conjunto de datos y la correlación entre ellas.
- Seleccionar los atributos que se utilizarán para elaborar el modelo.
- Evaluar los modelos disponibles para realizar análisis predictivo de fallas en generadores eléctricos en centrales hidroeléctricas.
- Construir un modelo que permita identificar posibles fallas en generadores eléctricos en centrales hidroeléctricas.
- Evaluar el desempeño del modelo predictivo seleccionado.

3. MARCO TEÓRICO Y ANTECEDENTES

3.1. MARCO TEÓRICO

La sociedad actual que conocemos transita por un proceso de digitalización masivo, donde se utiliza una infinidad de dispositivos eléctricos y electrónicos para uso cotidiano, como ejemplos más comunes podemos resaltar el uso del celular y el computador, tanto en ambientes laborales y educativos como en ambientes de ocio y esparcimiento.

Estos dispositivos eléctricos y electrónicos dependen directamente de la electricidad para poder funcionar, es por eso por lo que la energía eléctrica se ha convertido en parte fundamental y funcional de la vida moderna; sería difícil imaginar la vida hoy en día sin la energía eléctrica.

Hay muchas formas de generar energía eléctrica, desde fuentes alternativas como el viento y la energía solar, así como también plantas de energía nuclear y una de las más comunes es la energía producida por medio de centrales hidroeléctricas; gracias a sus grandes afluentes hídricos, este el medio más utilizado en Colombia para la generación de energía eléctrica. De acuerdo con [2] durante el mes de noviembre de 2021 “el 86.46% de la generación fue producto de recursos renovables y el 13.54% restante de recursos no renovables”.

El funcionamiento de una central hidroeléctrica se basa en la transformación de distintos tipos de energía; la energía potencial del agua almacenada en un embalse se transforma en energía cinética al ser conducida por túneles o tuberías hasta una turbina acoplada a un generador eléctrico. Esta energía cinética es aprovechada por el generador eléctrico para ser transformada en energía eléctrica para luego ser transportada y distribuida a los centros poblados para el consumo de los usuarios industriales y residenciales.

Una central hidroeléctrica (Fig. 2 Central hidroeléctrica. [4]) se conforma por un conjunto de componentes generales y específicos que permiten su óptimo funcionamiento, según [4] estos componentes generales se encuentran definidos por: una presa que contiene y almacena el recurso hídrico, también cuenta con conducciones que transportan el agua hasta la central hidroeléctrica, una sala de máquinas que aloja los activos de la unidad generadora, turbinas que son quienes transforman la energía potencial en cinética, alternador o generador que se encarga de transformar la energía cinética en energía eléctrica.

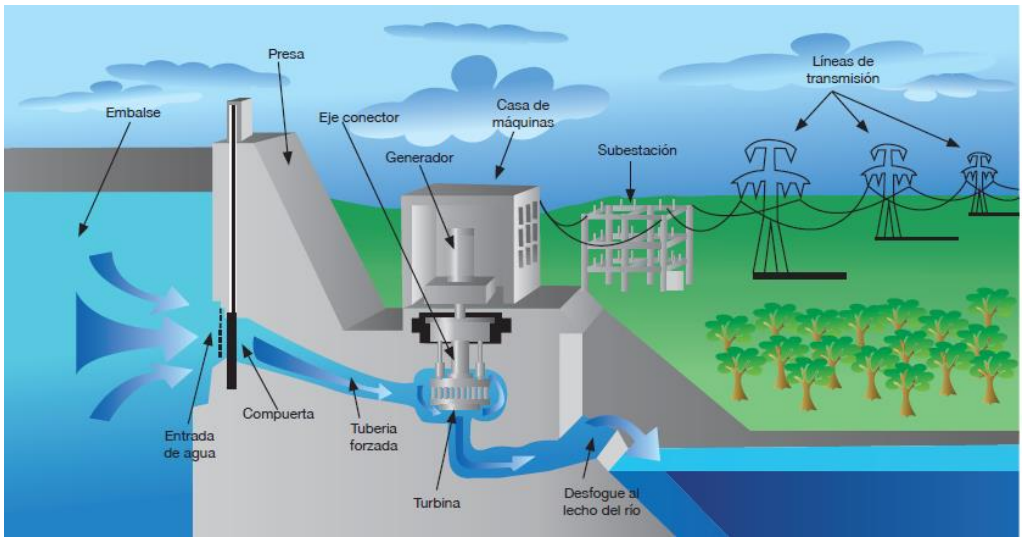


Fig. 2. Central hidroeléctrica [4]

Existen también componentes muchos más específicos dentro de la casa de máquinas que realizan el control de los diferentes parámetros de operación del generador eléctrico. El principal elemento de corte hidráulico es la válvula de admisión, que permite o bloquea el paso de agua desde la tubería de conducción hacia la turbina. Para el control de las revoluciones en relación con la potencia generada el elemento encargado es el regulador de velocidad, mientras que para el control del voltaje de referencia y potencia reactiva se encarga el regulador de tensión. El calor generado por la dinámica del proceso de generación se regula por el sistema de refrigeración que tiene un rol muy importante en el día a día de la operación.

La operación de estas centrales hidroeléctricas puede llegar a ser muy compleja, debido a la gran cantidad de elementos electrónicos y de control que la componen, con la finalidad de lograr centralizar el control y la supervisión de todos los sistemas se utilizan controles de unidad empleando controladores lógicos programables (PLCs). Estos controladores se han vuelto cada vez más sofisticados y a su vez procesan una gran cantidad de información relevante para el proceso.

Toda la información recopilada en los PLCs se centraliza en un nivel superior, el sistema encargado de centralizar toda la información es el SCADA. Aunque esta herramienta permite la supervisión y control de todos los equipos en niveles inferiores, este no es capaz de realizar, por ejemplo, correlación entre variables de proceso y normalmente no cuentan con modelos para análisis de comportamientos del proceso.

Para realizar un análisis objetivo de comportamientos se pueden emplear distintos modelos, en la Fig. 3 se presentan las principales técnicas en minería de datos [2], estos pueden ser de tipo supervisado (predictivos) o no supervisado (descriptivos). Para objeto del presente documento nos enfocaremos en los modelos de tipo supervisado.

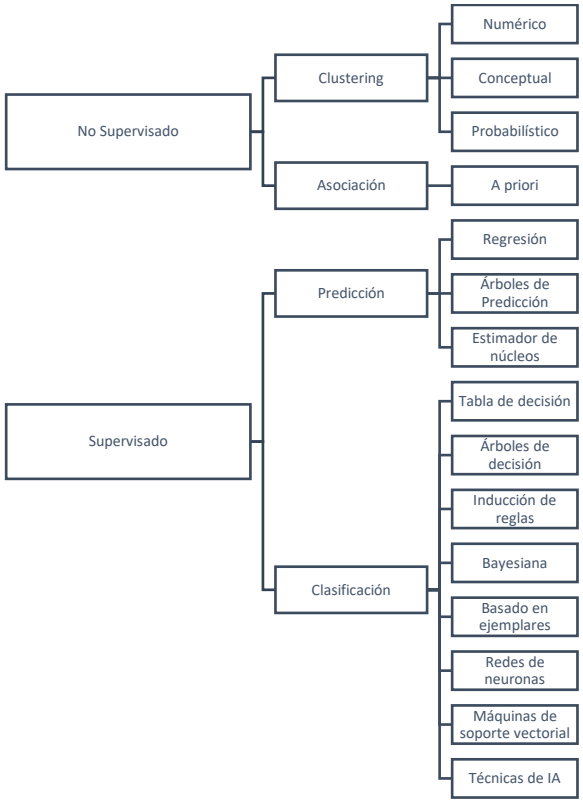


Fig. 3. Técnicas de minería de datos [5]

Según[5], los modelos supervisados "Tienen como principal objetivo aproximar posibles valores del futuro o desconocidos a través de los datos de los que ya se dispone. Los datos van acompañados de una salida (clase, categoría o valor numérico). La regresión y la clasificación son técnicas comúnmente usadas en este tipo de modelos".

Estos modelos supervisados acarrean una serie de tareas que de acuerdo con el autor [5] se pueden desarrollar para dar solución al problema. Dentro de dichas actividades se encuentra la clasificación, cuya finalidad es la de predecir la clase de nuevas instancias, esta a su vez se deriva en clasificación suave cuando involucra un porcentaje de certeza. También existen otras tareas como categorización, priorización y regresión; donde esta última busca "aprender una función real para asignar un valor real a una instancia".

[2] Esta categorización representa la finalidad para lo que normalmente son más utilizados.

En [6]se define aprendizaje supervisado como “cuando entrenamos un algoritmo de Machine Learning dándole las preguntas (características) y las respuestas (etiquetas). Así en un futuro el algoritmo pueda hacer una predicción conociendo las características”. Además, indica que el Machine Learning se soporta bajo los siguientes modelos: Modelos Lineales, Modelos de Árbol y Redes neuronales. Así mismo, precisa que Machine Learning “Es una rama de la Inteligencia Artificial que se encarga de generar algoritmos que tienen la capacidad de aprender y no tener que programarlos de manera explícita. El desarrollador no tendrá que sentarse a programar por horas tomando en cuenta todos los escenarios posibles ni todas las excepciones posibles. Lo único que hay que hacer es alimentar el algoritmo con un volumen gigantesco de datos para que el algoritmo aprenda y sepa qué hacer en cada uno de estos casos”.

Con la evolución de tecnologías, como es el caso de Machine Learning, se han empezado a abordar principales retos de la operación en la generación de energía como son[7]:

- ¿Esta operación es normal o no?
- ¿Cuál será la oferta, la demanda o la condición operativa en el futuro?
- ¿A qué categoría pertenece este patrón?
- ¿Se pueden sustituir las medidas complejas, costosas, frágiles o de laboratorio por un cálculo?
- ¿Cómo se ajustan los puntos de control en tiempo real para mantener estable el proceso?
- ¿Cómo y cuándo se cambian los puntos establecidos para mejorar el proceso de acuerdo con alguna medida de éxito?

Esto ha convertido a la tecnología de Machine Learning en la mejor opción costo eficiente para responder a estas preguntas.

Especialmente, el mantenimiento predictivo se ha convertido en uno de los principales tópicos del Machine Learning, que en combinación con la predicción en el tiempo y el diagnostico, puede ayudar a los operadores de la planta a programar mantenimientos proactivos, previniendo una falla y sus daños colaterales, que suele ser el 90% del costo financiero total de una falla [7].

3.2. ANTECEDENTES

Uno de los grandes retos para utilizar un modelo de aprendizaje automático dentro de la detección de anomalías es la escogencia de las variables más importantes para el análisis, Mulongo et al [8], utilizaron variables como: horas de trabajo del generador, la tasa de consumo, combustible consumido y la cantidad de combustible añadido en el generador, para le detección de anomalías en plantas de generación Diesel usando machine Learning.

Xayyasith et al [9], utilizaron los datos de temperatura de entrada y salida de un intercambiador de calor, para entrenar el algoritmo de machine Learning para construir un modelo de mantenimiento predictivo de un sistema de enfriamiento en la planta hidroeléctrica de Nam Ngum-1.

Vallim Filho et al [10], proponen un marco de referencia para encontrar las variables más importantes del proceso asociadas al Ciclo de carga de una turbina, la cual es una medida que se utiliza para definir la programación de mantenimiento de la turbina, encontrando 23 variables representativas para el proceso y el modelo de aprendizaje.

Por su parte, Betty et al [11] enfocaron su publicación en dos plantas hidroeléctricas distintas, en la primera de ellas emplearon un conjunto de datos de 630 señales análogas, mientras que para la segunda solo utilizaron 60 señales de tipo análogas. Estas señales se tomaron de componentes de la captación, compuertas, turbina, generador y transformadores de potencia.

A diferencia de los anteriores autores, en [12] utilizaron solo las medidas de seis sensores registradas cada hora durante los años 2011 al 2016, en donde se obtuvieron alrededor de 8600 muestras por año. Dentro de las señales seleccionadas se encuentran la medida de potencia activa, apertura de álabes y medidas de temperatura en cojinetes. A partir de las medidas seleccionadas se definieron tres modos de trabajo mediante un algoritmo K-means de tres clústeres, utilizando la distancia euclidiana y 3000 iteraciones, esto con la finalidad de dividir las observaciones recolectadas en tres modos de trabajo para ser modeladas a través de un algoritmo SOM.

Cao et al [13], emplearon los valores de las oscilaciones del cojinete guía superior y valores de potencia activa, donde se registraron 172396 valores de oscilación entre el 01/04/2015 a las 7:44:41 hasta el 07/04/2015 a las 23:59:17. Con estos valores y partiendo del teorema del límite central, los autores realizaron distintas pruebas de distribución normal, además propusieron una estrategia de alarma basada en características de distribución con los datos de oscilación del rodamiento guía superior de una unidad hidroeléctrica en China.

Luego de seleccionar las variables e identificar sus correlación e incidencias en el proceso, se define cuáles son los algoritmos de machine Learning más adecuados para utilizar, por ejemplo, Mulongo et al [8], utilizaron los algoritmos: SVM (Support Vector Machines), MLP (MultiLayer Perceptron), KNN (K-Nearest Neighbors) and LR (Logistic Regression) classifiers.

Anucha et al [9], utilizaron los algoritmos SVM (Support Vector Machines), regresión logística, KNN (k-Nearest Neighbors) y clasificación de conjuntos para su modelo.

Algoritmos de mapas autoorganizados basados en redes neuronales fueron utilizados por [11], esto, con la finalidad de construir un modelo del comportamiento nominal del sistema, teniendo como base para el entrenamiento los conjuntos de datos para las dos plantas. Este tipo de algoritmos pertenecen a la categoría de aprendizaje no supervisado.

La metodología utilizada por [12] comprende dos etapas, la primera de ellas es una etapa preparatoria y la segunda corresponde a una etapa de evaluación de comportamiento. Al igual que [11], se emplean mapas autoorganizados que reducen el conjunto de datos en clústeres de observaciones similares sin la necesidad de tener un conocimiento previo. Luego de esto se comparan los comportamientos de dos patrones o se compara un patrón de referencia con observaciones discretas. El siguiente paso cuantifica numéricamente las discrepancias utilizando tres indicadores, indicador de similitud, indicador de desviación e indicador compuesto. Finalmente se realiza el análisis de los resultados basados en transición y desviación entre patrones de comportamiento.

En contraste con otros autores, Cao et al [13] enfocaron su publicación en el uso el muestreo aleatorio sobre los datos de monitoreo de una unidad hidroeléctrica de forma repetida conforme al teorema del límite central. Basados en la probabilidad de distribución se explora la factibilidad de utilizarla para establecer ajustes de alarmas para los datos bajo supervisión. Todo el análisis y pruebas se realizaron recopilando información con la unidad de generación bajo condiciones estables.

A diferencia de las publicaciones mencionadas, donde se adoptó un enfoque en sistemas específicos, el objeto de este documento se orientó en el análisis de la ocurrencia de eventos o fallas durante los años 2021 y 2022 que pueden o no pertenecer a un mismo sistema. También, se emplearon distintos conjuntos de data sets en diferentes periodos de tiempo con impactos diferentes sobre la unidad de generación.

4. DESARROLLO DEL PROYECTO

4.1. METODOLOGIA

La metodología utilizada para la realización del modelo fue la propuesta por [14], la cual se encuentra apoyada en la Gestión de Procesos de Negocio o BPM; esta metodología utiliza la información proveniente de cada proceso alineándola con los objetivos empresariales de una empresa. Con el conocimiento de cada parte de los procesos se logra obtener transparencia de estos, aumentando su velocidad y flexibilidad mediante automatización continua, disminuyendo los errores y ahorrando tiempo[3]. El modelo en cuestión consta de cinco fases que se ejecutan de forma repetitiva, a saber:

1. *Mapeo de Proceso*:
   En esta fase se realiza el entendimiento general del proceso de generación de energía en una central hidroeléctrica, la cual se realiza a través de documentación técnica y entrevistas con los operadores de la central.

2. *Definición y Construcción de Variables Derivadas*:
   En esta fase se identifican las variables principales o críticas del proceso, así como variables derivadas a partir de métricas construidas sobre puntos de operación como límites de temperaturas entre otras.

3. *Recopilación, preparación, transformación y almacenamiento de datos*:
   En esta fase se realiza todo lo referente a la preparación, transformación y procesamiento de los datos, como la eliminación de ruido, la validación de outliers y la "suciedad" del dataset.

4. *Modelado Predictivo*:
   Esta es la fase central del framework en la cual se utiliza una variedad de técnicas para definir cuál es el mejor modelo para lograr el objetivo del principal.

5. *Aplicación en un caso de estudio*:
   Esta fase permite realizar una aplicación de las fases anteriores en una central hidroeléctrica.

[3] GBTEC Software AG. "¿Qué es BPM? Definición y aplicaciones" gbtec.com. [En línea]. Disponible en: https://www.gbtec.com/es/recursos/bpm/#:~:text=El%20término%20Business%20Process%20Management,para%20orientarlos%20a%20objetivos%20concretos. (Consultado May. 23, 2023).

Para el desarrollo del proyecto la fase 5 esta implícita en las primeras cuatro fases, ya que el proceso se desarrolló sobre un caso de estudio de una central hidroeléctrica.

4.1.1. Mapeo del proceso

El proceso de generación de energía en centrales hidroeléctricas se encuentra definido a grandes rasgos por algunas etapas que se muestran la Fig. 4:

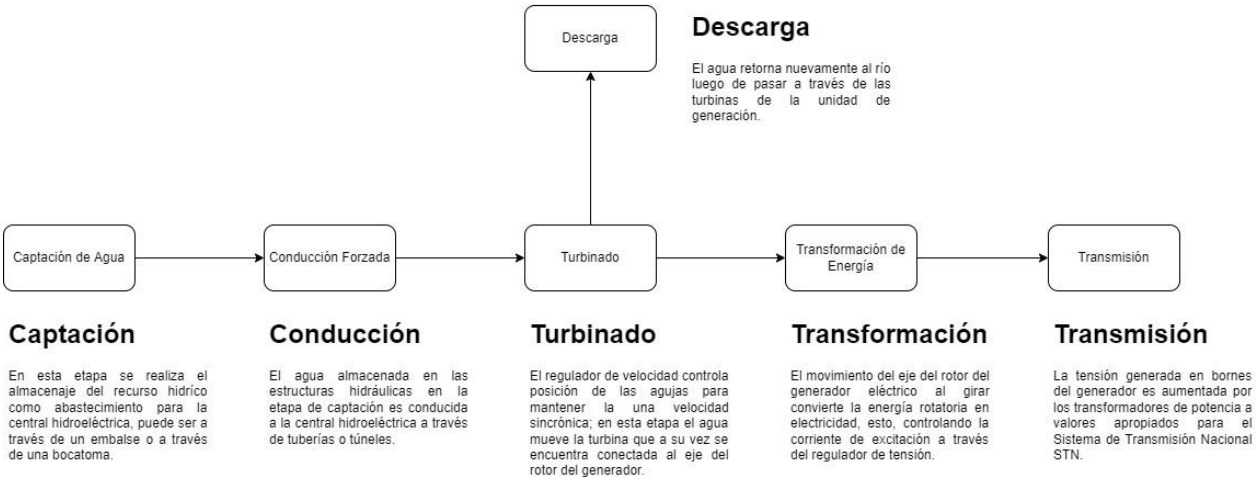


Fig. 4. Proceso de generación de energía en centrales hidroeléctricas. Fuente: Elaboración propia.

Las centrales hidroeléctricas pueden estar conformadas por una o más unidades de generación, cada una de ellas con subsistemas propios y otros comunes a todas las unidades. Algunos de los subsistemas propios comprenden válvula esférica, sistema de refrigeración, regulador de velocidad, regulador de tensión, entre otros.

Toda la información de los subsistemas propios y comunes se recopila en los servidores del sistema de Supervisión, Control y Adquisición de Datos (SCADA), donde, a través de este software se consulta la información almacenada en las bases de datos y se ejerce control sobre los equipos que así lo tengan dispuesto.

Dentro de la central de generación existen diferentes tipos de SCADA de acuerdo con la pirámide de automatización. Algunos de ellos ofrecen mayor o menor detalle en lo que corresponde a información de los distintos sistemas asociados a la generación de energía eléctrica.

En la pirámide de automatización encontramos en el nivel inferior toda la instrumentación, sensores y actuadores; en el nivel 1, se ubican los controladores lógicos programables PLCs. En el nivel 2, encontramos los sistemas de supervisión, control y adquisición de datos SCADA, en el nivel 3 encontramos los sistemas de gestión de producción y en el nivel 4 se ubican los sistemas de

planificación de recursos empresariales, en soluciones más modernas encontramos servicios en la nube para analítica de datos, computación en la nube, entre otros.

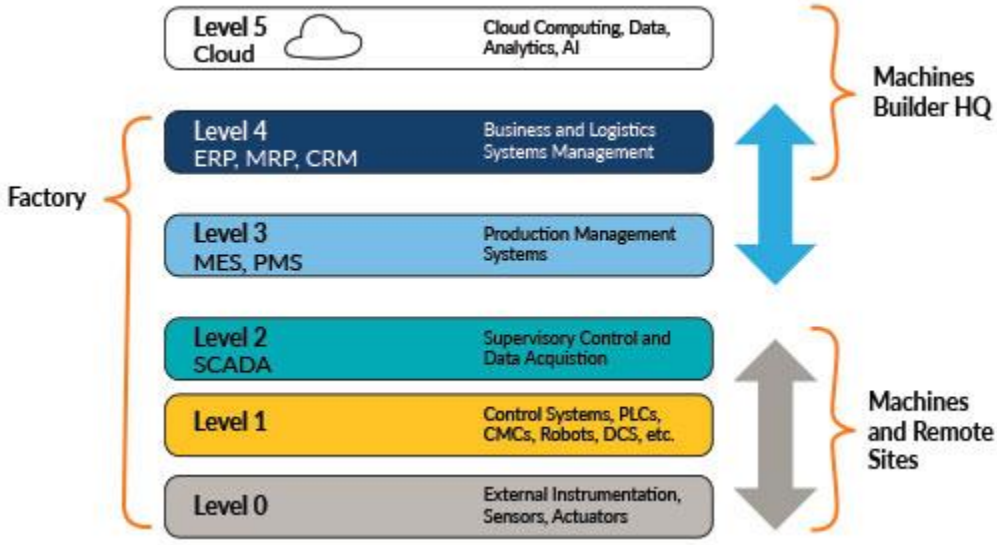


Fig. 5. Pirámide de Automatización [15]

La base de datos del SCADA aloja una cantidad considerable de señales asociadas a una unidad de generación, se consideró necesario establecer un listado de las señales que pudieran tener mayor relación con la ocurrencia de fallas en las unidades de generación.

Esta base de datos recopila la información de alrededor de 45.000 puntos de información (atributos) provenientes de todas las unidades de generación de ocho centrales hidroeléctricas. Estos puntos pueden ser consultados con una frecuencia de muestreo de máximo un (1) segundo. Durante la exploración inicial, se identificó que una (1) unidad de generación cuenta con alrededor de 200 atributos que se pueden consultar desde la base de datos del sistema SCADA.

Dentro del listado de estos 200 atributos podemos identificar señales provenientes de los sensores, que de forma general indican temperaturas, presiones, flujos, niveles, tensiones, corrientes, entre otras.

4.1.2. Selección de atributos más relevantes

De los 200 atributos previamente identificados no se tuvieron en cuenta todos ellos, ya que algunos o muchos de ellos pueden no aportar información relevante para la identificación de fallas. En ese sentido, se establecieron algunos criterios que permitieron realizar una selección oportuna de los atributos más relevantes relacionados con las fallas.

El primer criterio para realizar una buena selección de atributos fue explorar el registro de indisponibilidades de las unidades. Este registro recopila información referente al activo,

relevancia, instalación, inicio de indisponibilidad, fin de indisponibilidad, descripción y causa. Esta última, relaciona el o los atributos que preliminarmente condujeron a una falla, es decir, si la causa de una falla fue un bajo aislamiento de la resistencia de aislamiento del rotor, este atributo obtiene relevancia para ser seleccionado para la construcción del modelo.

Cabe resaltar que una indisponibilidad refleja el estado de la unidad de generación cuando no es posible entrar en operación comercial, a partir del registro de indisponibilidades se identificaron cuáles fueron las fallas consignadas (el tipo de falla que se produjo y posteriormente se registró) y su posible causa.

Un segundo criterio que se tuvo en cuenta fue la comparación del registro de indisponibilidades con el registro de eventos de estado de la unidad, este último refleja cuando una unidad se encuentra en estado en servicio (*unidad en operación comercial*), indisponible (*unidad detenida y no lista para entrar en servicio*) y reserva (*unidad detenida y lista para entrar en servicio*). Es necesario aclarar que un evento de unidad no siempre está asociado a una indisponibilidad, con algunas excepciones como la ocurrencia de una falla. El evento de unidad permite identificar la fecha y hora de la ocurrencia de una falla, que al estar asociada a una indisponibilidad permite trazar toda la información referente al tiempo de ocurrencia de una falla y su posible causa.

Luego de realizado el filtrado de la información de las fallas, indisponibilidades y estado de unidad, se realizó la selección de los atributos considerados más relevantes a partir del criterio y experiencia profesional de los operadores e ingenieros del área de mantenimiento y operación de central hidroeléctrica; se identificaron 58 atributos considerados como los más relevantes que aportan información frente a la ocurrencia de las fallas registradas en los periodos de tiempo descritos.

Con la finalidad de evaluar la pertinencia de la selección de estos atributos, se realizó una exploración de los datos durante algunos periodos de falla en una ventana de tiempo de cinco horas, en la Fig. 6, Fig. 7 y Fig. 8, la línea vertical roja indica el momento de la ocurrencia de la falla. Se obtuvieron las siguientes gráficas:

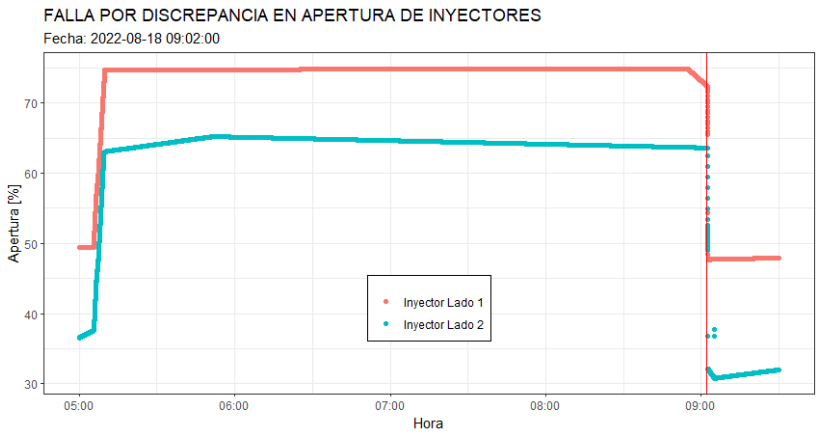


Fig. 6. Discrepancia en apertura de inyectores. Fuente: Elaboración propia.

La posición de los inyectores debe ser proporcional para mantener un buen balance en el eje del rotor, esto puede traducirse en vibraciones más estables. Se puede observar cómo aumenta la diferencia de posición entre el inyector lado 1 y el inyector lado 2 a partir de las 06:00 horas, lo que finalmente se refleja en una falla por discrepancia en apertura de inyectores a las 09:00 horas en la línea roja vertical.

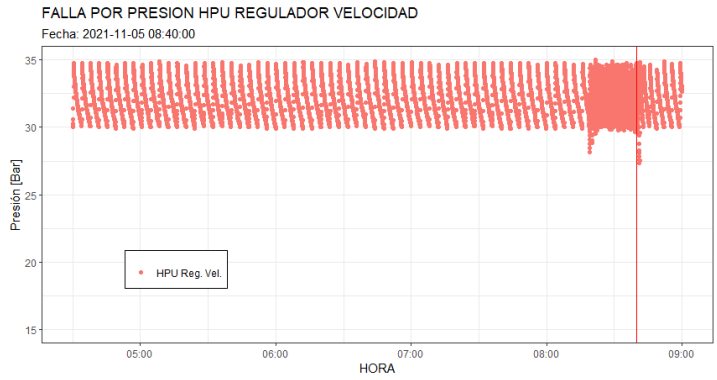
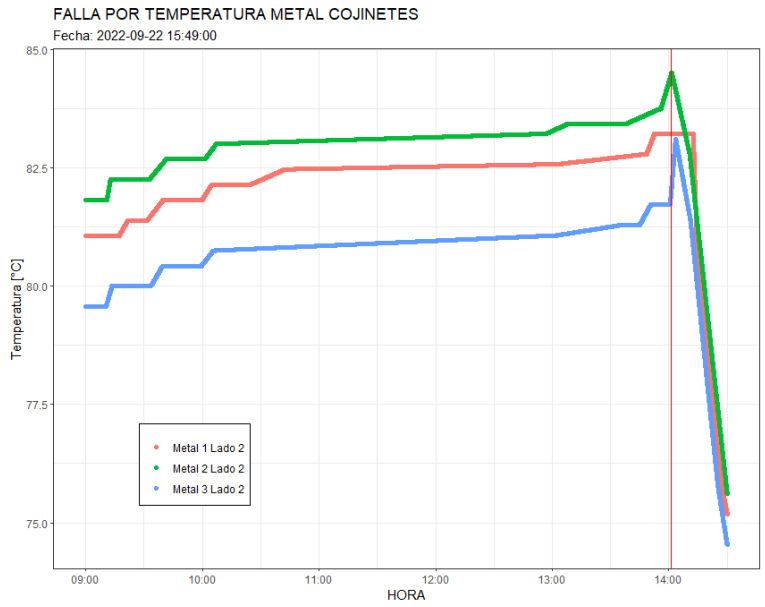


Fig. 7. Baja presión HPU regulador de velocidad. Fuente: Elaboración propia.

La unidad hidráulica debe mantener una presión constante a través de las bombas de la HPU (Unidad de potencia hidráulica, por sus siglas en inglés) del regulador de velocidad. Se puede observar cómo en la línea roja vertical esta presión cae por debajo de 30 Bar y se presenta el evento de falla.



La temperatura del metal del cojinete turbina se debe mantener bajo ciertos niveles operativos, esto indicaría una buena refrigeración del aceite. Se puede apreciar cómo la temperatura incrementa exponencialmente hasta alcanzar el punto de disparo de la unidad.

Fig. 8. Alta temperatura metal cojinete turbina. Fuente: Elaboración propia.

En la exploración de los datos se observó que algunos de ellos contenían muchos registros con valores corruptos, con etiquetas tales como "*No Sample*" donde no se almacenó ningún dato para este atributo, "*Bad*" donde el valor del atributo fue descartado por mala calidad, "*Tag not found*" donde no se encontró el tag del atributo, "No Data" donde no se registró ninguna información.

En la Fig. 9, Fig. 10, Fig. 11, Fig. 12 y Fig. 13, se pueden observar ejemplos de la cantidad de registros con valores corruptos.

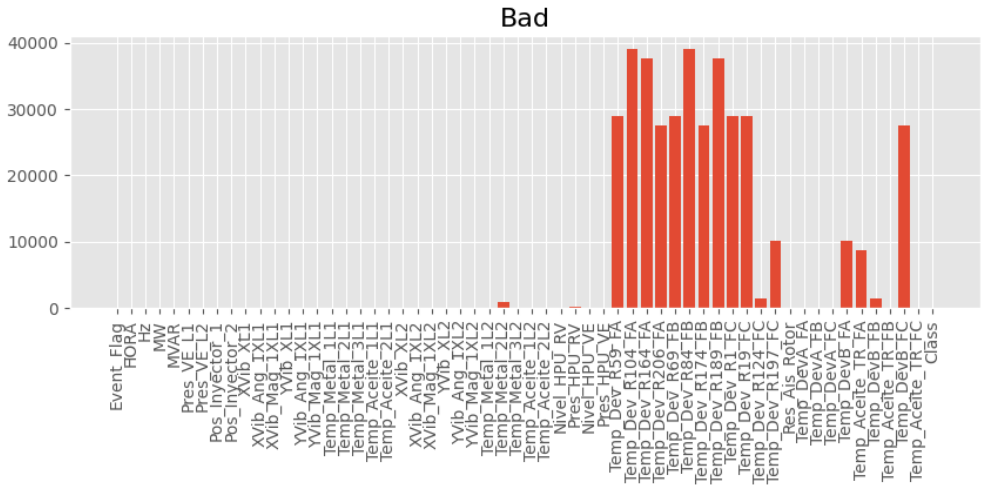


Fig. 9. Cantidad de registros con valor "Bad". Fuente: Elaboración propia.

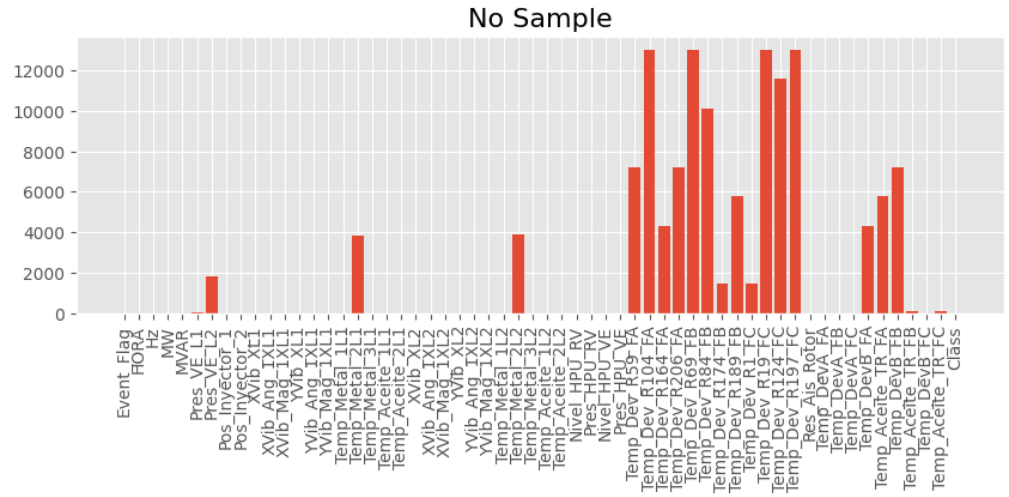


Fig. 10. Cantidad de registros con valor "No Sample". Fuente: Elaboración propia.

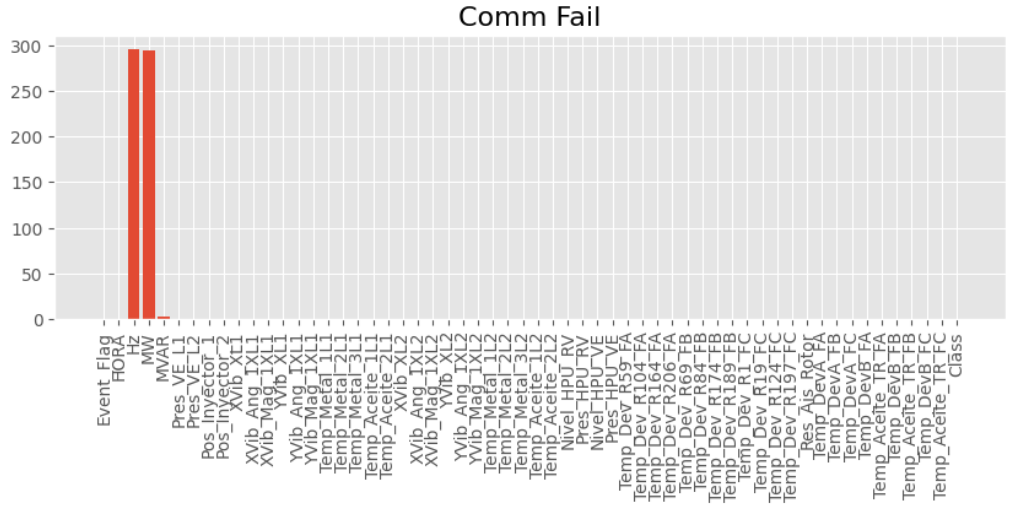


Fig. 11. Cantidad de registros con valor "Comm Fail". Fuente: Elaboración propia.

Fig. 12. Cantidad de registros con valor "No Data". Fuente: Elaboración propia.

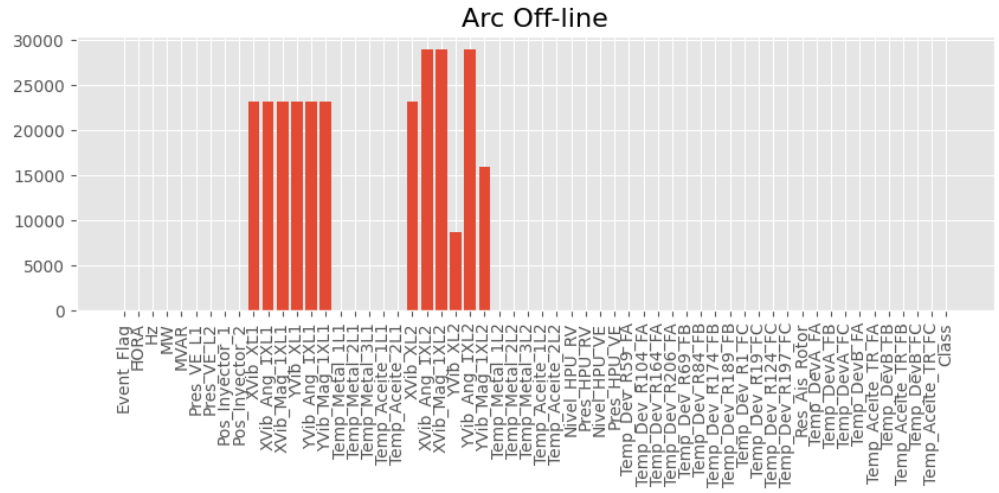


Fig. 13. Cantidad de registros con valor "Arc Off-line". Fuente: Elaboración propia.

Se optó por eliminar los atributos que tuvieran más del 50% de valores corruptos, ya que no aportaban información para la predicción de fallas.

Luego de los anteriores análisis, quedaron los siguientes 23 atributos, donde 21 de ellos corresponden a los extraídos de la base de datos del SCADA, mientras que los 2 atributos adicionales, Event_Flag y Class, se construyeron para la manipulación de los datos:

TABLA I. DATASET FINAL CON 23 ATRIBUTOS

Index	Sistema	Descripción	Nombre	Tipo de Dato
1	NA	Clase derivada	Event_Flag	Object
2	SCADA	TimeStamp del SCADA	Hora	datetime64
3	Transmisión	Frecuencia	Hz	float64
4	Transmisión	Potencia Activa	MW	float64
5	Transmisión	Potencia Reactiva	MVAR	float64
6	Conducción	Presión Aguas Arriba Lado 1	Pres_VE_L1	float64
7	Turbina	Apertura Aguja 1	Pos_Inyetor_1	float64
8	Turbina	Apertura Aguja 2	Pos_Inyetor_2	float64
9	Generador	Temperatura Metal 1 Cojinete Lado 1	Temp_Metal_1L1	float64
10	Generador	Temperatura Metal 3 Cojinete Lado 1	Temp_Metal_3L1	float64
11	Generador	Temperatura Aceite 1 Cojinete Lado 1	Temp_Aceite_1L1	float64
12	Generador	Temperatura Aceite 2 Cojinete Lado 1	Temp_Aceite_2L1	float64
13	Generador	Temperatura Metal 1 Cojinete Lado 2	Temp_Metal_1L2	float64
14	Generador	Temperatura Metal 3 Cojinete Lado 2	Temp_Metal_3L2	float64
15	Generador	Temperatura Aceite 1 Cojinete Lado 2	Temp_Aceite_1L2	float64
16	Generador	Temperatura Aceite 2 Cojinete Lado 2	Temp_Aceite_2L2	float64
17	Turbina	Nivel Aceite HPU[4] Regulador Velocidad	Nivel_HPU_RV	float64
18	Turbina	Presión Aceite HPU Regulador Velocidad	Pres_HPU_RV	float64
19	Turbina	Nivel Aceite HPU Válvula Esférica	Nivel_HPU_VE	float64
20	Turbina	Presión Aceite HPU Válvula Esférica	Pres_HPU_VE	float64
21	Generador	Resistencia Aislamiento Rotor	Res_Ais_Rotor	float64
22	Transmisión	Temperatura Aceite Trafo Fase C	Temp_Aceite_TR_FC	float64
23	NA	Clase derivada, 1 para falla, 0 para no falla	Class	int

[4] Unidad de potencia hidráulica por sus siglas en inglés Horse Power Unit.

4.1.3. Recopilación, preparación, transformación y almacenamiento de datos

Para la construcción de cada conjunto de datos desde dos horas antes de un evento (H2) hasta cinco horas de un evento (H5) con los atributos de la TABLA I y siguiendo la metodología propuesta en [14] se convirtió un problema de serie de tiempo en un problema de clasificación supervisada, para lograr esto se realizaron los siguientes pasos:

Primero, se definieron unos rangos de tiempo de interés de los operadores para predecir una falla en la unidad de generación, como el tiempo mínimo para detener una unidad de generación de forma controlada es de 1 hora y media aproximadamente, se definió el tiempo mínimo de predicción de 2 horas, también se definieron unos rangos mayores para dar mayor tiempo a los operadores para tomar acciones de mantenimiento correctivo, estos tiempos son 3 horas, 4 horas y 5 horas.

Segundo, con base en el registro de indisponibilidades, se identificaron las fechas y horas de ocurrencia de las fallas y su posible causa, a través del listado de eventos.

Tercero, a partir del momento exacto de la falla, se devolvió hacia atrás en dichos conjuntos de datos en los rangos definidos, 2h, 3h, 4h, 5h etiquetando los registros para 4 minutos de cada falla, se agregaron las etiquetas H2, H3, H4 y H5 en el atributo *Event_Flag,* estas etiquetas permitieron crear cuatro conjuntos de datos desde dos horas antes de un evento (H2) hasta cinco horas de un evento (H5) para cada rango de tiempo.

Por último, se descargaron los distintos conjuntos de datos del SCADA para los atributos definidos en la TABLA I. En la figura Fig. 14 se muestran los pasos explicados anteriormente.

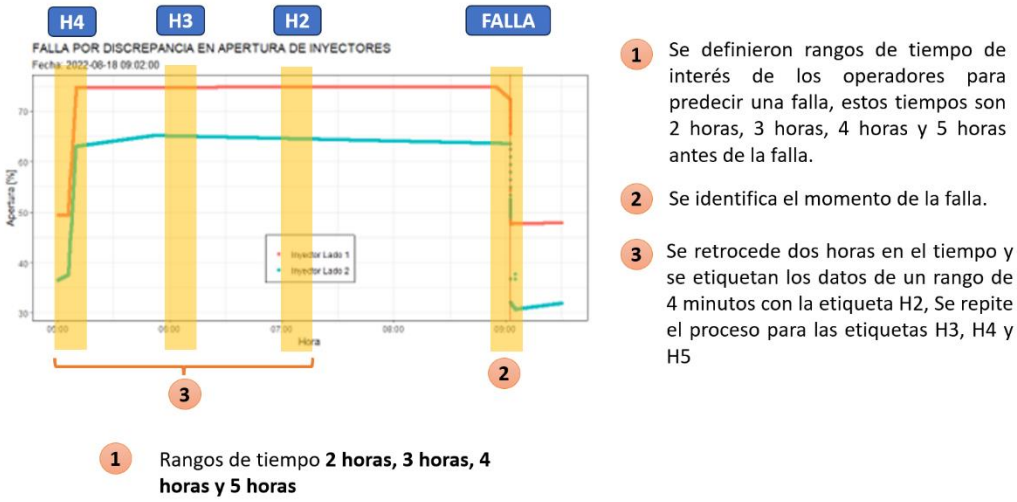


Fig. 14. Proceso de construcción de conjuntos de datos. Fuente: Elaboración propia

En la figura Fig. 14 se observa que la falla ocurrió a las 9:00, siguiendo la metodología presentada se etiquetaron los registros de las 7:00 a las 7:04 en el atributo Event_Flag con “H2”, los registros de las 6:00 a las 6:04 en el atributo Event_Flag con “H3”, los registros de las 5:00 a las 5:04 en el atributo Event_Flag con “H4” y los registros de las 4:00 a las 4:04 en el atributo Event_Flag con “H4”.

Se definió una ventana de tiempo de 4 minutos con la finalidad de poder observar el comportamiento de todos los atributos 2, 3, 4 y 5 horas antes de la ocurrencia de la falla y lo que esta desencadena en un tiempo posterior, esta definición se realizó bajo la hipótesis de que es un tiempo adecuado para apreciar las variaciones frente a algunos eventos que ocurren de forma muy rápida.

Para el tratamiento de los datos de eventos durante periodos de tiempo donde no se registraron fallas, se realizaron los mismos pasos, es decir, se tomaron ventanas de tiempo de 4 minutos 2, 3, 4 y 5 horas antes del tiempo definido.

Estos periodos de tiempo de no falla se tomaron aleatoriamente; teniendo en cuenta que la unidad de generación debería estar en operación comercial para que los datos recopilados tuvieran validez.

A continuación, se detallan algunas características de cada conjunto de datos desde dos horas antes de un evento (H2) hasta cinco horas de un evento (H5); estos tienen un total de 9841 observaciones y 23 atributos; los cuales se construyeron con 18 conjuntos de datos de eventos de falla y 24 conjuntos de datos de periodos de tiempo durante los cuales no se registraron fallas, para un total de 42 conjuntos de datos con información de dos horas antes de un evento (H2) hasta cinco horas de un evento (H5), cada uno de ellos contiene los datos de todos los atributos en una ventana de tiempo de cuatro (4) minutos, tomados en intervalos de tiempo de dos (2), tres (3), cuatro (4) y cinco (5) horas.

Los conjuntos de datos finales cuentan con los atributos de frecuencia, potencia activa, potencia reactiva, presión, nivel y temperatura de la HPU de la válvula esférica, vibraciones, temperaturas de cojinetes metal y aceite, presión, nivel y temperatura de la HPU del regulador de velocidad, temperatura devanados estator, temperatura devanados y aceite trasformador de potencia, resistencia de aislamiento rotor y el atributo clase, este último corresponde a sí las observaciones se encuentran asociadas a eventos de falla (valor 1) o no falla (valor 0).

A partir de las anteriores apreciaciones y teniendo en cuenta las características de estos conjuntos de datos, se destaca que el problema es de clasificación; para lo cual se emplean algoritmos de aprendizaje supervisado.

4.1.4. Selección de los modelos a aplicar

De acuerdo con distintos autores, se emplearon diferentes algoritmos de machine learning; se destaca que Mulongo et al [8], utilizaron los algoritmos basados en Support Vector Machines (SVM), MultiLayer Perceptron (MLP), K-Nearest Neighbors (KNN) and Logistic Regression classifiers (LR). Anucha et al [9], también utilizaron los algoritmos de SVM, KNN y LR; mientras que Betti et al [11] emplearon algoritmos basados en redes neuronales, Adhya et al [15] utilizaron XGBoot para la detección de fallas en sistemas fotovoltaicos, Vallim Filho et al [14] utilizaron algoritmos como árboles de decisión, redes neuronales artificiales y regresión logística para predecir las fallas en una unidad de generación basados en el ciclo de carga de los equipos.

Los algoritmos de aprendizaje supervisado utilizan los datos etiquetados para entrenar un algoritmo bajo tareas específicas. En nuestro caso, se entrenaron distintos modelos para clasificar si bajo ciertos parámetros se puede identificar la ocurrencia de una falla en la central de generación horas antes de que ésta ocurra.

Vallim Filho et al [14] en un estudio de trabajos relacionados en el uso de machine learning para mantenimiento predictivo en la industria, para un total de 562 artículos académicos publicados entre 2015 y 2020, encontraron que las técnicas más aplicadas fueron: redes neuronales artificiales, aprendizaje profundo y un tercer grupo con técnicas como: K vecinos más cercano (KNN), Arboles de decisión (DT) , Bosques aleatorios (RF), Naïve Bayes (NB), Máquinas de Soporte Vectorial (SVM) y XGBoost.

Con la finalidad de realizar la comparación de algunos modelos de aprendizaje supervisado, se seleccionaron tres (3) de ellos, a saber: Bosques Aleatorios, Máquinas de Soporte Vectorial y XGBoost, que de acuerdo con la literatura consultada son los que mejores resultados presentaron para este tipo de problemas, donde se tienen datos estructurados en formato de tabla como es el caso de los conjunto de datos descargados del SCADA, por esta misma razón se descartaron algoritmos de redes neuronales y aprendizaje profundo.

Random Forest:

Ramdon forest o bosques aleatorios, es un método popular de aprendizaje automático supervisado, el cual es una colección de árboles de clasificación y regresión, los cuales son modelos simples que, utilizan divisiones binarias en variables predictoras para determinar los resultados de la predicción, los árboles de decisión son un método intuitivo para predecir un resultado dividiendo los valores altos y bajos de un predictor relacionado con ese resultado [16].

Support Vector Machines (SVM):

Las máquinas de soporte vectorial son un método de aprendizaje automático supervisado usado para separar dos o más clases encontrado un hiperplano que maximiza el margen de separación entre ellos [8].

XGBOOST:

XGBoot es un método de aprendizaje automático supervisado para clasificación y regresión, es una abreviatura de “Extreme Gradient Boosting” (refuerzo de gradiente extremo).

Este método se basa en árboles de decisión, un factor clave de este algoritmo es su escalabilidad ya que el algoritmo puede manejar datos paralelos y escasos, XGBoost tiene la característica de manejar datos faltantes, la paralelización de la construcción de árboles y la capacidad de entrenamiento continúo garantizando su alta eficiencia y velocidad de ejecución con bajo consumo de memoria [15].

4.1.5. Construcción del modelo

Utilizando el registro de indisponibilidades y seleccionando los atributos relevantes, se conformaron cuatro conjuntos de datos que abarcan desde dos horas antes del evento (H2) hasta cinco horas antes del evento (H5). Estos conjuntos contienen observaciones de fallas y no fallas. Cada conjunto de datos representa una ventana de tiempo de cuatro minutos, capturada entre dos y cinco horas previas a una falla, así como períodos sin fallas.

El modelo de predicción se construye a partir de cada uno de estos conjuntos de datos, que abarcan desde dos horas antes del evento (H2) hasta cinco horas antes del evento (H5). Estos conjuntos se basan en la información extraída del sistema SCADA y cubren el período comprendido entre febrero de 2021 y septiembre de 2022.

```
import pandas as pd
fail_event_list = [None]*18

for i in range(0, 18):
    event = 'EI_{}'.format(i+1)
    fail_event_list[i] = pd.read_excel("./Data_UG.xlsx", header=0, sheet_name=event)
  return go(f, seed, [])
}
```

Fig. 15. Importación archivos de eventos. Fuente: Elaboración propia.

Una vez importados los datos se filtran las observaciones con valores anormales para evaluar opciones de imputación o eliminación, estos valores corresponden a etiquetas asignadas por el sistema SCADA cuando existen problemas con la comunicación. Las etiquetas mencionadas son: '*Arc Off-line*', '*No Data*', '*No Sample*', '*Bad*', '*Tag not found*', '*Comm Fail*'.

```
index_ = list({'Arc Off-line', 'No Data', 'No Sample', 'Bad', 'Tag not found', 'Comm Fail'})
columns_ = list(full_df.columns)
table_data = []

for i in index_:
    row = []
    for j in columns_:
        count = full_df.loc[full_df[j] == i].index
        count_ = len(count)
        row.append(count_)
    table_data.append(row)

missing_df = pd.DataFrame(table_data, index = index_, columns = columns_)
missing_df.head()
```

Fig. 16. Filtrado de observaciones con error en la comunicación. Fuente: Elaboración propia.

A partir del filtrado de las observaciones con error en la comunicación o calidad del dato, se creó una tabla donde se tabularon los valores para cada uno de los atributos. Con los resultados obtenidos, se realizó un análisis gráfico para cada una de las etiquetas descritas contra los atributos en los conjuntos de datos.

```
import matplotlib.pyplot as plt

for i in range(0, len(index_)):
    y = list(missing_df.iloc[i, :])
    x = list(full_df.columns)

    plt.style.use('ggplot')
    plt.figure(figsize=(10, 3))
    plt.title(index_[i], fontsize=16)
    plt.bar(x, y)
    plt.xticks(rotation=90)
    plt.show()
```

Fig. 17. Visualización etiquetas por error en comunicación. Fuente: Elaboración propia.

Se desarrollo una función para eliminar los atributos que tuvieran observaciones etiquetadas con error en la comunicación por encima del 50%, para los atributos que contenían pocas observaciones marcadas con estas etiquetas, solo se eliminaron las observaciones correspondientes. Los atributos que finalmente quedaron luego de la depuración de datos fueron cambiados a tipo numérico.

```
def adj_df(df):

  #Eliminar observaciones con etiquetas de error:
  df = df.loc[df['Hz'] != 'Comm Fail']
  df = df.loc[df['Pres_VE_L1'] != 'No Sample']
  df = df.loc[df['Pres_VE_L1'] != 'Bad']
  ...

  #Convertir atributos a numéricos:
  df['Hz'] = pd.to_numeric(df['Hz'])
  df['MW'] = pd.to_numeric(df['MW'])
  df['MVAR'] = pd.to_numeric(df['MVAR'])
  ...

  #Eliminar atributos con muchas observaciones etiquetadas:
  df = df.drop(['XVib_XL1'], axis=1)
  df = df.drop(['XVib_Ang_1XL1'], axis=1)
  df = df.drop(['XVib_Mag_1XL1'], axis=1)
  ...

  return(df)
```

Fig. 18. Eliminación de atributos con muchas etiquetas de datos erróneos. Fuente: Elaboración propia.

Una vez eliminadas las observaciones etiquetadas por el SCADA con error en la comunicación, para cada conjunto de datos desde dos horas antes de un evento (H2) hasta cinco horas de un evento (H5) de falla y no falla, se separaron algunos de estos conjuntos de datos para entrenamiento y otros para pruebas en una proporción de 70% y 30% respectivamente. Es decir, de 42 conjuntos de datos con información de dos horas antes de un evento (H2) hasta cinco horas antes de un evento (H5), 30 de ellos fueron tomados para entrenamiento y 10 para pruebas.

```
train_df_h2 = pd.DataFrame()
train_df_h3 = pd.DataFrame()
train_df_h4 = pd.DataFrame()
train_df_h5 = pd.DataFrame()

for elm in train_list:
    train_df_h2 = pd.concat([train_df_h2, elm.loc[elm.Event_Flag=='H2']], axis=0)
    train_df_h3 = pd.concat([train_df_h3, elm.loc[elm.Event_Flag=='H3']], axis=0)
    train_df_h4 = pd.concat([train_df_h4, elm.loc[elm.Event_Flag=='H4']], axis=0)
    train_df_h5 = pd.concat([train_df_h5, elm.loc[elm.Event_Flag=='H5']], axis=0)

train_data_2 = train_df_h2.drop(['Event_Flag'], axis = 1)
train_data_3 = train_df_h3.drop(['Event_Flag'], axis = 1)
train_data_4 = train_df_h4.drop(['Event_Flag'], axis = 1)
train_data_5 = train_df_h5.drop(['Event_Flag'], axis = 1)

train_data_2_ = train_data_2.reset_index().drop(['index'], axis = 1)
train_data_3_ = train_data_3.reset_index().drop(['index'], axis = 1)
train_data_4_ = train_data_4.reset_index().drop(['index'], axis = 1)
train_data_5_ = train_data_5.reset_index().drop(['index'], axis = 1)
```

Fig. 19. Separación de datos para entrenamiento y pruebas. Fuente: Elaboración propia.

También se realizaron pruebas para cada uno de los algoritmos de aprendizaje automático explorando diferentes combinaciones de los hiperparámetros, esto con la finalidad de obtener unos rangos preliminares de variación para configuración de los modelos.

```
X_train = train_data_2.drop(['Class'], axis=1)
y_train = pd.to_numeric(train_data_2['Class'])
X_test = test_data_2
estimators = list(range(10, 1010, 10))
deep = [None]*len(estimators)
samples_split = [2]*len(estimators)
samples_leaf = [1]*len(estimators)

macro_max, micro_max, weigh_max = f2_score(X_train=X_train,
                                           y_train=y_train,
                                           test=X_test,
                                           x_axis=estimators,
                                           xlab='n_estimators',
                                           estimators=estimators,
                                           deep=deep,
                                           samples_split=samples_split,
                                           samples_leaf=samples_leaf)
```

Fig. 20. Exploración rangos óptimos de hiperparámetros, Random Forest. Fuente: Elaboración propia.

Con la finalidad de encontrar los mejores estimadores para cada uno de los modelos a construir, se configuraron las funciones de Grid Search y Random Search de la siguiente manera para ser incorporadas posteriormente dentro de la construcción de los diferentes modelos:

```
gs = GridSearchCV(estimator=clf, param_grid=RF_hyper_param, scoring='accuracy',
                  cv=5, n_jobs=-1, verbose=0)

rs = RandomizedSearchCV(estimator=clf, param_distributions=RF_hyper_param, scoring=f2_scorer,
                        cv=5, n_jobs=-1, verbose=0, n_iter=10)
```

Fig. 21. Evaluación de mejores estimadores con Grid Search y Random Search. Fuente: Elaboración propia.

Luego de la exploración de distintas combinaciones de los hiperparámetros se encontraron los siguientes mejores rangos que aumentaron la presión de los modelos. Estos, fueron utilizados para aplicar las técnicas de Grid Search y Random Search para obtener los mejores estimadores.

```
RF_hyper_param = {
    'n_estimators': [250, 270, 290, 310, 330, 350, 370, 390, 410],
    'max_features' : ['sqrt'],
    'bootstrap' : [True, False]
}
```

Fig. 22. Definición de hiperparámetros Random Forest. Fuente: Elaboración propia.

```
SVM_hyper_param = {
    'kernel': ['linear', 'poly', 'rbf', 'sigmoid'],
    'degree': list(range(1, 5, 1)),
    'gamma': ['scale', 'auto'],
    'C': list(np.arange(2, 3, 0.1))
    'bootstrap' : [True, False]
}
```

Fig. 23. Definición de hiperparámetros SVM. Fuente: Elaboración propia.

```
XG_hyper_param = {
    'max_depth': [1],
    'learning_rate': [0.6, 0.7],
    'n_estimators': [15, 16, 17, 18, 19, 20],
    'gamma': [30, 35, 40, 45],
    'subsample': [0.7, 0.8, 0.9, 1],
    'colsample_bytree': [0.1],
    'alpha': [1, 3, 5, 8, 11],
    'reg_lambda': [10, 20, 30, 40, 50]
}
```

Fig. 24. Definición de hiperparámetros XGBOOST. Fuente: Elaboración propia.

Con la finalidad de optimizar el modelo, se creó una función que permitió calcular los mejores parámetros para su construcción, calcular las métricas de accuracy, precision, recall y F2 score del modelo y obtener una representación gráfica de los resultados.

```
def RF_clf(train, test):
    X_train = train.drop(['Class'], axis=1)
    y_train = pd.to_numeric(train['Class'])
    clf = RandomForestClassifier(random_state=42)

    gs = GridSearchCV(estimator=clf, param_grid=RF_hyper_param, scoring='accuracy',
                      cv=5, n_jobs=-1, verbose=0)

    gs.fit(X_train, y_train)
    X_test = test.drop(['Class'], axis=1)
    y_pred = gs.predict(X_test)
    y_true= test['Class']

    clf_report = precision_recall_fscore_support(y_true, y_pred, zero_division=0,
                                                 beta=2, average='micro')
    best_params = (gs.best_params_)
    cm = confusion_matrix(y_true, y_pred)

    return(best_params, clf_report, cm, gs)
```

Fig. 25. Cálculo de mejores parámetros y reporte de clasificación. Fuente: Elaboración propia.

Para mayor detalle en el código del algoritmo, se puede consultar el ***Anexo 1*** desarrollado para el cálculo de los distintos modelos de predicción.

4.1.6. Evaluación del desempeño del modelo predictivo

Las métricas utilizadas para medir el desempeño de los modelos son los siguientes:

Accuracy: El *accuracy* de la clasificación, se calcula con el número de clasificaciones correctas TP (True Positive[5]) + TN (True Negative[6]), por el número total de ejemplos FP (False Positive[7]) + FN (False Negative[8]) + TP + TN [17].

$$\text{Acc} = \frac{TP + TN}{FP + FN + TP + TN}$$

[5] Verdaderos Positivos
[6] Verdaderos Negativos
[7] Falsos Positivos
[8] Falsos Negativos

Precision: Es el porcentaje de verdaderos positivos TP, entre todos los ejemplos que el clasificador ha etiquetado como positivos TP + FP [17].

$$\mathrm{Pr} = \frac{TP}{TP + FP}$$

Recall: Es la probabilidad de que un ejemplo positivo sea correctamente reconocido como tal (por el clasificador), el valor se obtiene dividiendo el número de verdaderos positivos, por el número de positivos en el conjunto de datos [17].

$$\mathrm{Re} = \frac{TP}{TP + FN}$$

Fβ-Score: Combina las métricas de *precisión* y el *recall* en una sola métrica.

$$\mathrm{F}\beta = \frac{(\beta^2 + 1).Pr.Re}{\beta^2.Pr + Re}$$

El parámetro β ϵ [0, ∞], permitiendo definir el peso de la importancia relativa de los dos criterios, si β > 1, el mayor peso se le da al *recall,* si β < 1, se le da más peso a *Precision* [17]

Las métricas mencionadas tienen sus ventajas y desventajas, por ejemplo, la métrica de *Accuracy* no se recomienda para datos no balanceados, mientras si se quiere minimizar los falsos negativos sería más importante la métrica de *Recall*, y si se quiere minimizar los falsos positivos la métrica más importante es *Precision*.

Para nuestro problema de negocio, es más importante la métrica *Recall*, ya que queremos minimizar los falsos negativos, debido a que no apagar la unidad de generación y que se produzca un daño puede generar un alto impacto para el negocio.

Por eso las métricas seleccionadas para comparar los modelos son:

- Recall

$$\mathrm{Re} = \frac{TP}{TP + FN}$$

- F2-Score

$$\mathrm{F2} = 5(\frac{Pr.Re}{4.Pr + Re})$$

Para el cálculo de métrica F2-Score se utilizó la función precision_recall_fscore_support de scikit learn, la cual requiere otro hiperparámetro que es el de average, el cual determina el promedio realizado sobre los datos [18].

Las opciones para este parámetro son 'binary', 'micro', 'macro', 'samples', 'weighted', de los cuales se escogió el parámetro 'macro' ya que este parámetro no tiene en cuenta el desbalance de las clases[18] y los conjuntos de datos utilizados presentan desbalance en las clases.

Para cada uno de los algoritmos de aprendizaje automático seleccionados, Random Forest, Support Vector Machines y XGBOOST se extrajo la información de los mejores estimadores calculados a través de la búsqueda aleatoria de parámetros, el reporte de clasificación y la matriz de confusión. El índice asociado en cada una de las tablas corresponde a: dos horas antes del evento (H2), tres horas antes del evento (H3), cuatro horas antes del evento (H4) y cinco horas antes del evento (H5)

Random Forest:

TABLA II. MEJORES ESTIMADORES RANDOM FOREST

	n_estimators	max_features	bootstrap
H2	330	sqrt	True
H3	350	sqrt	False
H4	270	sqrt	True
H5	410	sqrt	True

De acuerdo con TABLA II, se identificó que, para cada modelo, dos, tres, cuatro y cinco horas antes, los estimadores son diferentes; con excepción del parámetro *max_features*.

TABLA III. CLASSIFICATION REPORT RANDOM FOREST

	precision	recall	f2_score	support
H2	0.832638	0.832638	0.832638	None
H3	0.832638	0.832638	0.832638	None
H4	0.928487	0.914660	0.913164	None
H5	0.832638	0.832638	0.832638	None

También fue posible observar cómo la precisión para los cuatro modelos se encuentra en un valor

entre 0.83 y 0.92, evidenciando una buena clasificación de las observaciones de prueba. Esto último se corroboró gráficamente a través de la matriz de confusión en la Fig. 26.

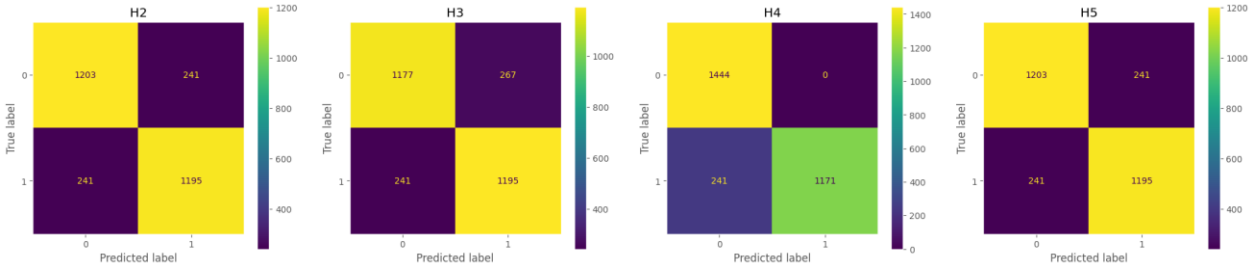


Fig. 26. Matriz de confusión Random Forest. Fuente: Elaboración propia.

Support Vector Machines (SVM):

TABLA IV. MEJORES ESTIMADORES SVM

	kernel	gamma	degree	C
H2	poly	scale	1	2.4
H3	poly	scale	1	2.3
H4	poly	scale	1	2.6
H5	linear	auto	4	2.2

Para el algoritmo de SVM se observa que todos los hiperparámetros van variando según el modelo, desde el *kernel* hasta *C*.

TABLA V. CLASSIFICATION REPORT SVM

	precision	recall	f2_score	support
H2	0.832638	0.832638	0.832638	None
H3	0.832638	0.832638	0.832638	None
H4	0.818429	0.818400	0.818318	None
H5	0.928487	0.916086	0.914387	None

Con respecto al reporte de clasificación de la TABLA V, se observa una respuesta similar a los resultados expuestos para Random Forest; donde se observa una buena clasificación de las observaciones de prueba. La matriz de confusión de la Fig. 27 representa los resultados de clasificación obtenidos.

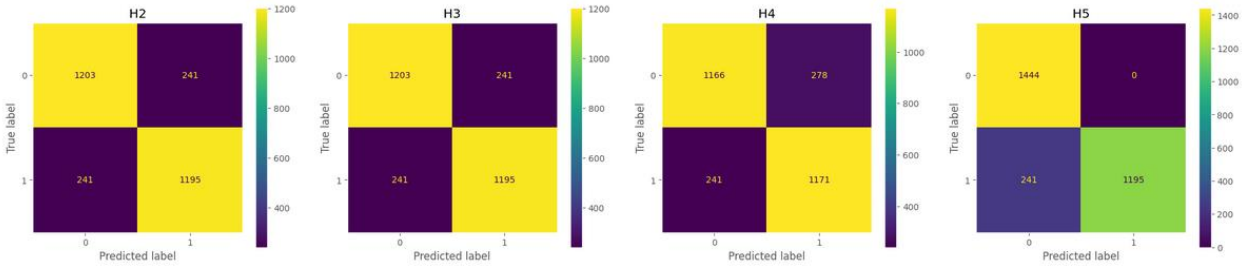


Fig. 27. Matriz de confusión SVM. Fuente: Elaboración propia.

XGBOOST:

TABLA VI. MEJORES ESTIMADORES XGBOOST

	subsample	reg_lambda	n_estimators	max_depth	learning_rate	gamma	colsample_bytree	alpha
H2	1.0	20	15	1	0.7	35	0.1	8
H3	0.8	50	16	1	0.7	40	0.1	3
H4	0.8	40	17	1	0.6	40	0.1	8
H5	0.9	20	19	1	0.6	35	0.1	11

Para XGBOOST se observan como varían cada uno de los hiperparámetros de acuerdo con cada modelo. Con respecto a la precisión del modelo se encontró diferencia en comparación con los otros dos algoritmos, situándose la precisión con XGBOOST en valores entre 0.77 a 0.87.

TABLA VII. CLASSIFICATION REPORT XGBOOST

	precision	recall	f2_score	support
H2	0.777189	0.773097	0.772158	None
H3	0.875260	0.832869	0.825548	None
H4	0.875260	0.830028	0.822935	None
H5	0.832638	0.832638	0.832638	None

En la siguiente matriz de confusión se validó gráficamente los resultados arrojados por el modelo.

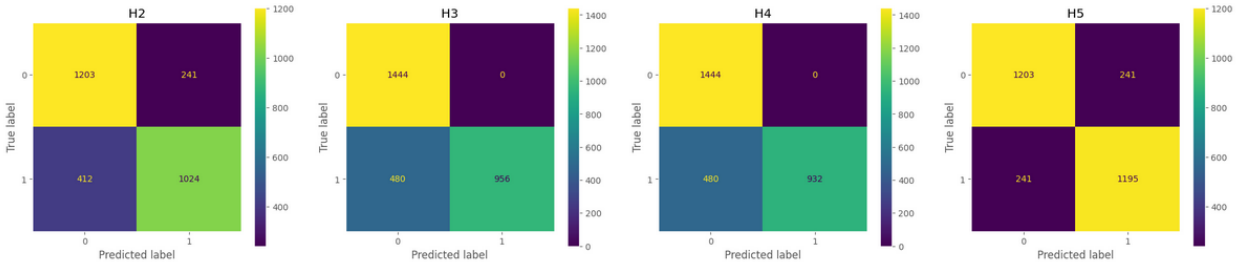


Fig. 28. Matriz de confusión XGBOOST. Fuente: Elaboración propia.

4.1.7. Evaluación del desempeño del modelo predictivo: tiempo ampliado

Con la finalidad de complementar los resultados obtenidos en la primera simulación, con periodos de tiempo de dos a cinco horas antes de la ocurrencia de una falla, se realizaron experimentos para periodos de tiempo entre nueve y doce horas antes de una falla. Para estos periodos de tiempo y para cada uno de los algoritmos de aprendizaje automático se extrajo la información referente a los reportes de clasificación y las diferentes matrices de confusión.

Random Forest:

TABLA VIII. CLASSIFICATION REPORT RANDOM FOREST, TIEMPO AMPLIADO

	precision	recall	f2_score	support
H09	1.0	1.0	1.0	None
H10	1.0	1.0	1.0	None
H11	1.0	1.0	1.0	None
H12	1.0	1.0	1.0	None

Para este algoritmo en particular se puede observar cómo todas las observaciones son clasificadas correctamente, en la siguiente sección se analizará este comportamiento. Gráficamente podemos evidenciarlo en la siguiente matriz de confusión:

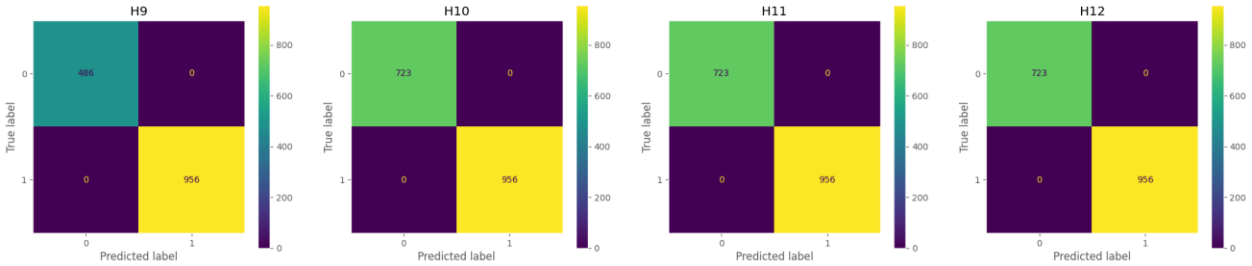


Fig. 29. Matriz de confusión Random Forest, tiempo ampliado. Fuente: Elaboración propia.

Support Vector Machines (SVM):

TABLA IX. CLASSIFICATION REPORT SVM, TIEMPO AMPLIADO

	precision	recall	f2_score	support
H09	0.980237	0.989540	0.987515	None
H10	0.887460	0.890167	0.880645	None
H11	0.875780	0.875000	0.863730	None
H12	1.000000	1.000000	1.000000	None

Al igual que con Random Forest, empleando el algoritmo de Support Vector Machines se obtiene una clasificación con valores de F2 altos, en la siguiente sección exploraremos este comportamiento. Gráficamente se pueden observar los resultados en la siguiente matriz de confusión:

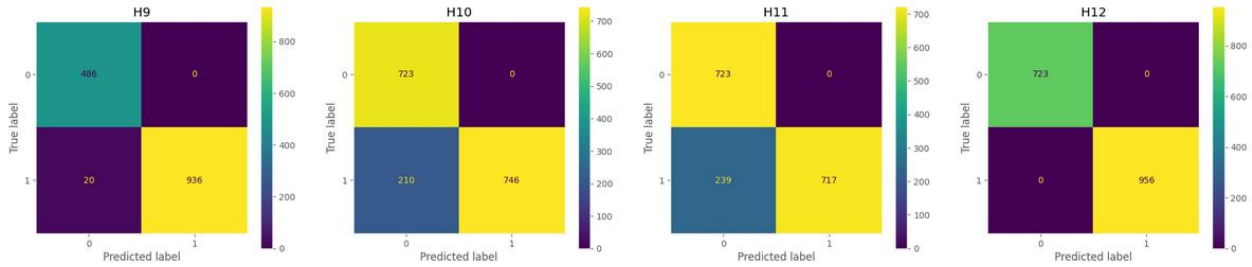


Fig. 30. Matriz de confusión SVM, tiempo ampliado. Fuente: Elaboración propia.

XGBOOST:

TABLA X. CLASSIFICATION REPORT XGBOOST, TIEMPO AMPLIADO

	precision	recall	f2_score	support
H09	1.000000	1.000000	1.000000	None
H10	1.000000	1.000000	1.000000	None
H11	1.000000	1.000000	1.000000	None
H12	0.891234	0.894874	0.885871	None

Con XGBOOST se tiene una precisión menor que 1.0 para el periodo de tiempo de H12, 12 horas antes de un evento de falla. La matriz de confusión Fig. 31 refleja estos resultados.

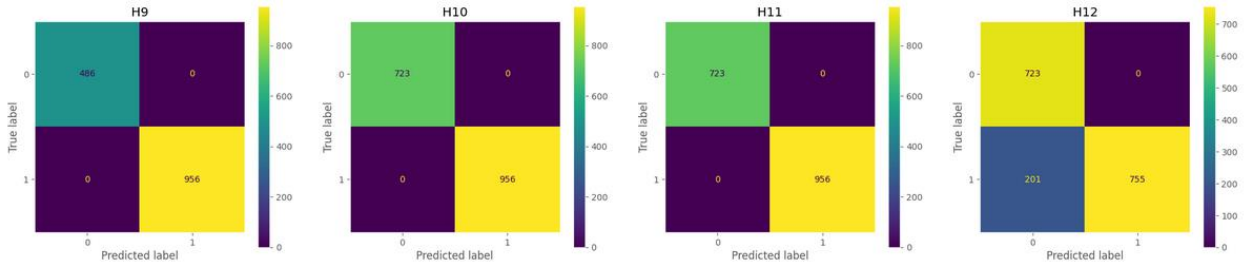


Fig. 31. Matriz de confusión XGBOOST, tiempo ampliado. Fuente: Elaboración propia.

4.1.8. Evaluación del desempeño del modelo predictivo: una observación por evento

El presente experimento tuvo como finalidad establecer si debido a la dinámica de los datos de las diferentes ventanas de tiempo era posible obtener los mismos resultados anteriormente observados, pero solo tomando una observación por hora, en contraste con la ventana de cuatro minutos definida. Para esto, se tomaron los periodos de tiempo entre dos y cinco horas para realizar este análisis. Los resultados obtenidos son los siguientes:

Random Forest:

TABLA XI. CLASSIFICATION REPORT RANDOM FOREST, UNA OBS. POR EVENTO

	precision	recall	f2_score	support
H02	0.833333	0.833333	0.833333	None
H03	0.833333	0.833333	0.833333	None
H04	0.833333	0.833333	0.833333	None
H05	0.833333	0.833333	0.833333	None

Los resultados de las métricas obtenidas son similares a las que se alcanzaron tomando una ventana de tiempo de cuatro minutos. Esto se puede evidenciar gráficamente en la siguiente figura:

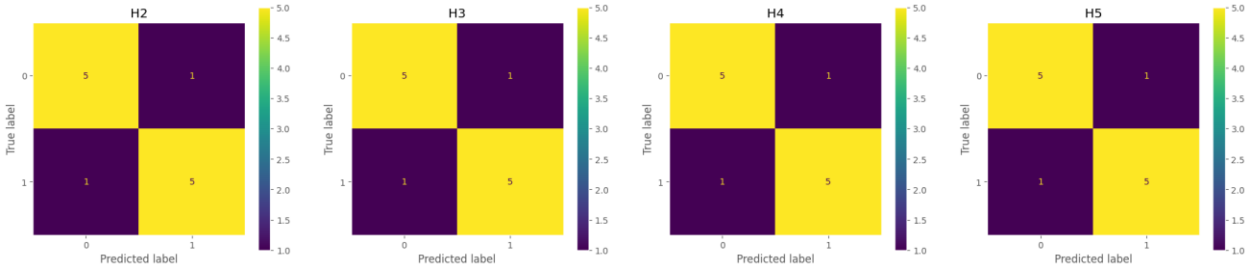


Fig. 32. Matriz de confusión Random Forest, una observación por evento. Fuente: Elaboración propia.

Support Vector Machines (SVM):

TABLA XII. CLASSIFICATION REPORT SVM, UNA OBS. POR EVENTO

	precision	recall	f2_score	support
H02	0.928571	0.916667	0.914905	None
H03	0.833333	0.833333	0.833333	None
H04	0.928571	0.916667	0.914905	None
H05	0.928571	0.916667	0.914905	None

De igual forma que para el algoritmo anterior, los resultados son similares, teniendo en cuenta que el número de observaciones se redujo a uno. La siguiente matriz de confusión refleja estos resultados de forma gráfica.

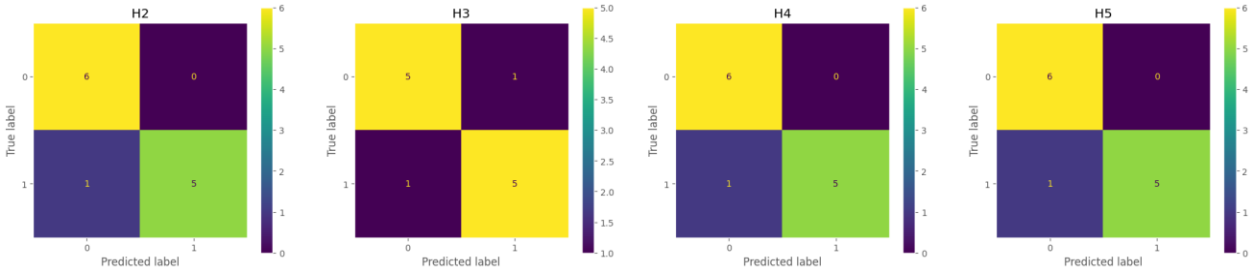


Fig. 33. Matriz de confusión SVM, una observación por evento. Fuente: Elaboración propia.

XGBOOST:

TABLA XIII. CLASSIFICATION REPORT XGBOOST, UNA OBS. POR EVENTO

	precision	recall	f2_score	support
H02	0.25	0.5	0.416667	None
H03	0.25	0.5	0.416667	None
H04	0.25	0.5	0.416667	None
H05	0.25	0.5	0.416667	None

A diferencia de los anteriores algoritmos, con XGBOOST se obtuvo un rendimiento menor con una reducción del alrededor del 50% del F2 score. Gráficamente los resultados están representados por la siguiente matriz de confusión:

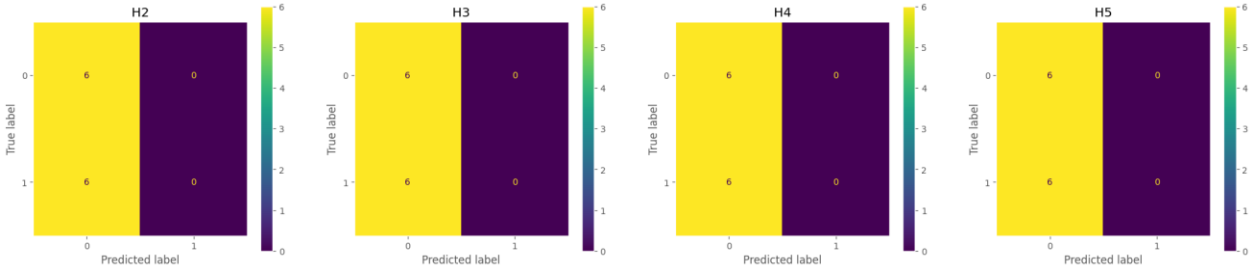


Fig. 34. Matriz de confusión XGBOOST, una observación por evento. Fuente: Elaboración propia.

5. RESULTADOS

A continuación, se presentan los resultados de importancia de atributos calculados para cada uno de los modelos, a partir del atributo **feature_names_in_** para los algoritmos de aprendizaje automático que ofrecieron mejor precisión, Random Forest y XGBOOST.

Random Forest:

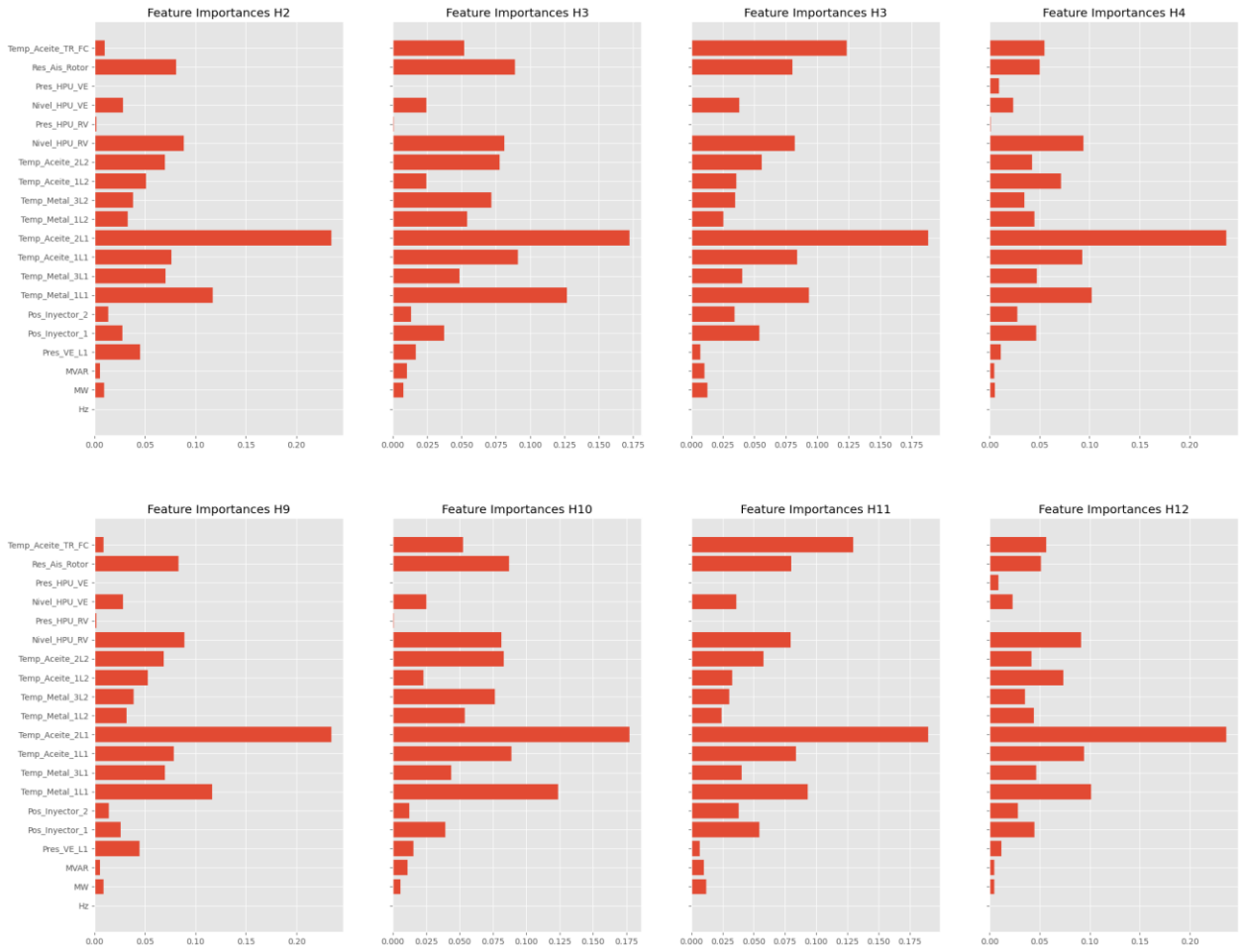


Fig. 35. Importancia de atributos Random Forest. Fuente: Elaboración propia.

De la Fig. 35 se pudo identificar que los atributos más representativos son Temperatura de Aceite 1 Lado 1 y Temperatura de Aceite 2 Lado 1.

XGBOOST:

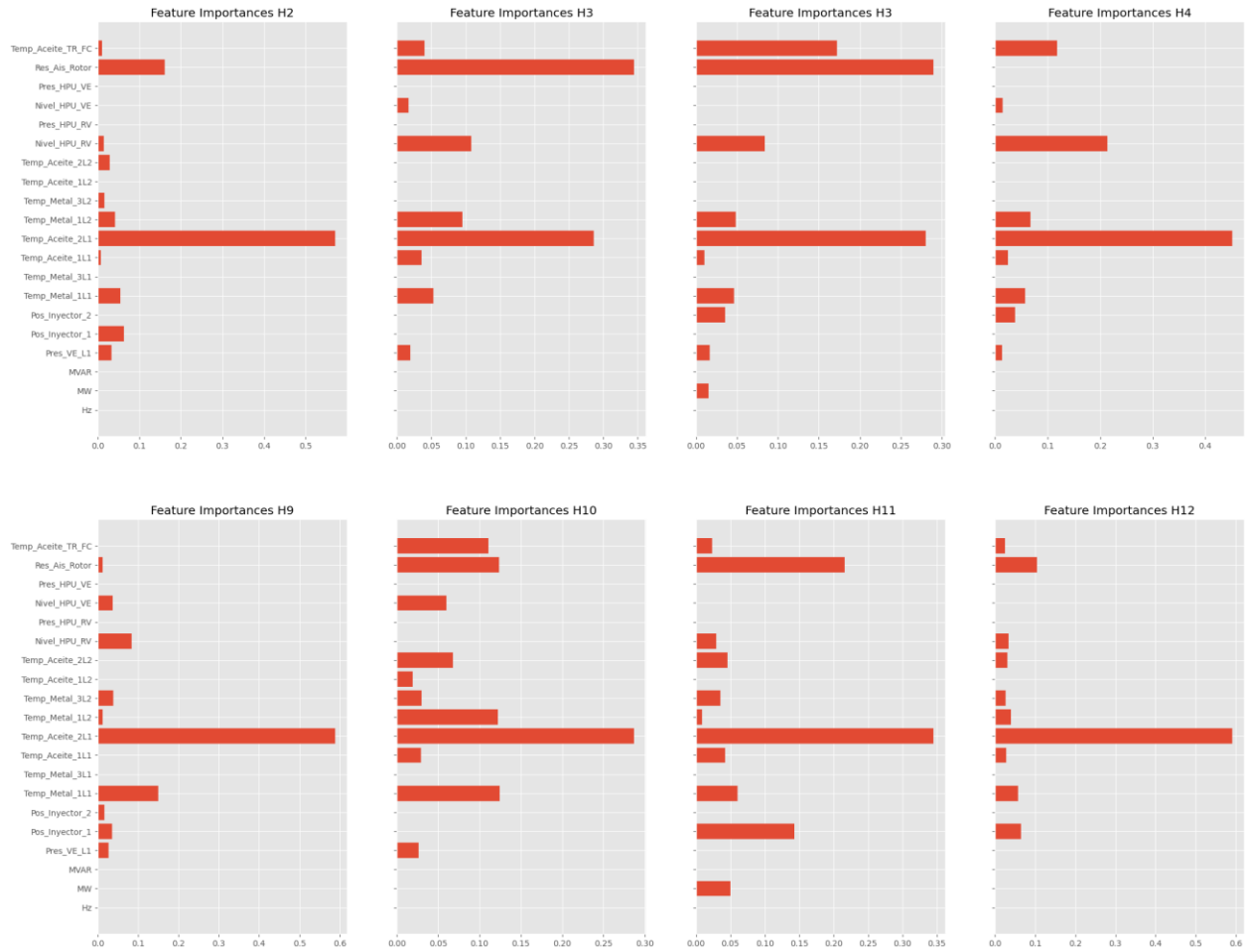


Fig. 36. Importancia de atributos XGBOOST. Fuente: Elaboración propia.

A diferencia de los resultados de importancia de variables obtenidos a través de Random Forest, con XGBOOST se observa una distribución, dándole el mayor peso a la Temperatura de Aceite 2 Lado 1.

Con la intención de analizar el comportamiento de este atributo e identificar la dispersión durante eventos de falla y no falla, se realiza una gráfica que ilustra todos los valores que ha tomado para todos los periodos de tiempo definidos.

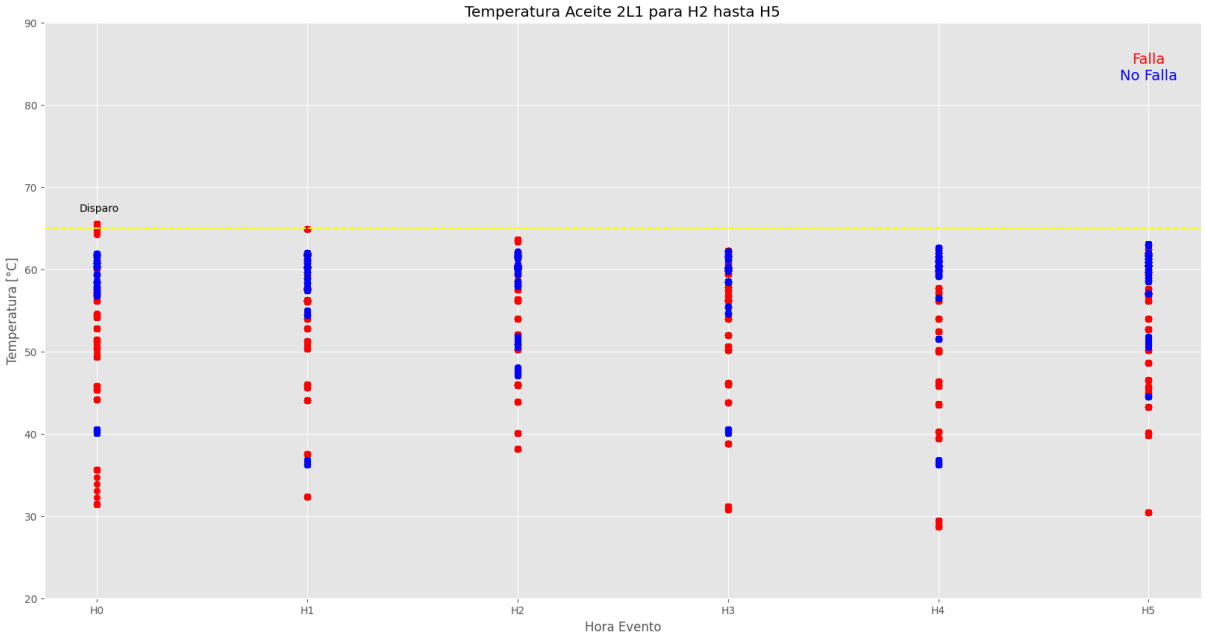


Fig. 37. Temperatura de Aceite 2L1 para H2 hasta H5. Fuente: Elaboración propia.

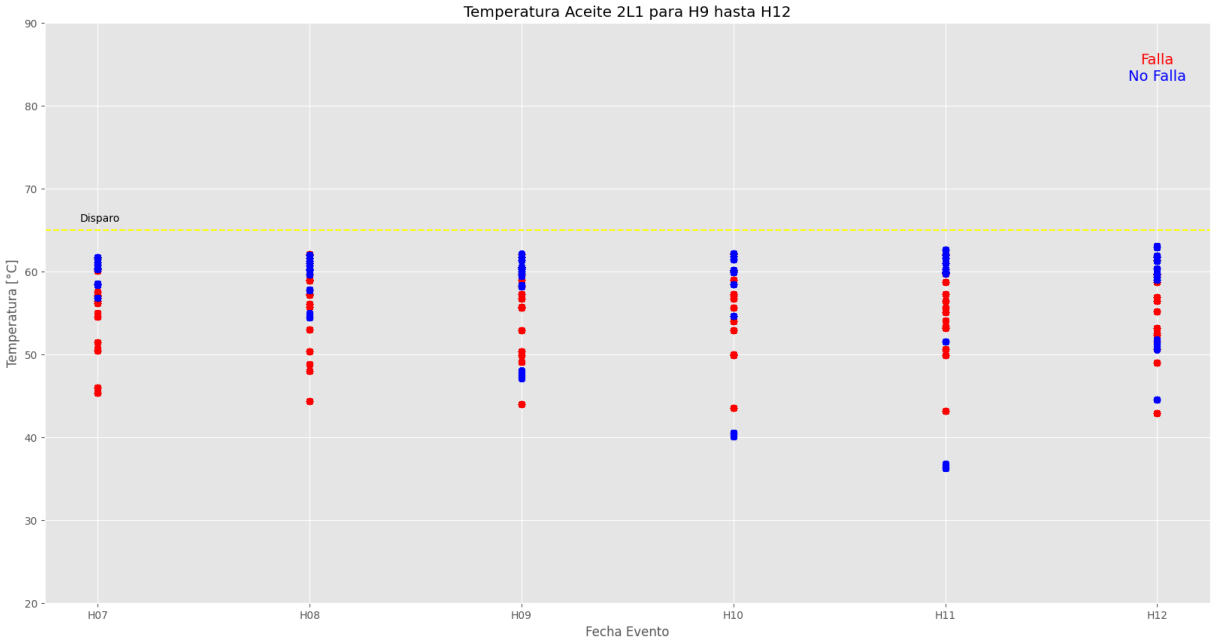


Fig. 38. Temperatura de Aceite 2L1 para H9 hasta H12. Fuente: Elaboración propia.

Conforme a estas figuras se puede identificar que los valores que tomó la Temperatura de Aceite 2L1 durante eventos de falla y de no falla desde periodos de tiempo H2 hasta H12, se encuentran dentro de los valores operativos normales; donde el valor máximo para el disparo de la unidad es de 65 °C.

Se realizó un análisis comparativo para cada uno de los atributos en los diferentes periodos de tiempo, desde dos horas antes de un evento (H2) hasta doce horas antes de un evento (H12), obteniendo los siguientes resultados:

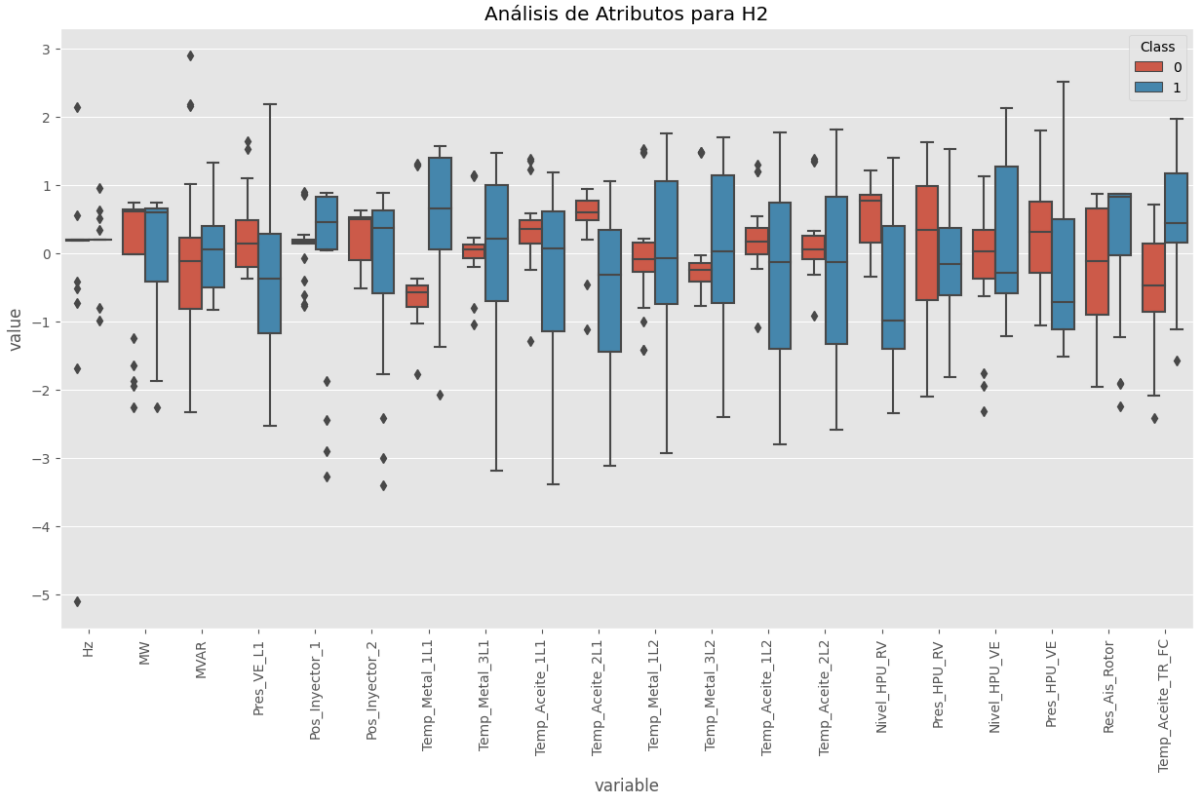


Fig. 39. Análisis de atributos para H2. Fuente: Elaboración propia.

Dos horas antes de un evento, Fig. 39, algunas de las variables tienen comportamientos parecidos, es decir, su valor se mueve dentro los mismos márgenes. Mientras que para otras variables se pudo observar que los márgenes donde se mueven estas variables son distintos. Se observó que la diferencia del valor de la Temperatura de Aceite 2 Lado 1 es muy evidente. También, la Temperatura de Aceite del Transformador de la Fase C tiene el mismo comportamiento. Es de anotar que las variables mencionadas coinciden con el análisis de importancia de atributos registrados en la Fig. 35 y la Fig. 36.

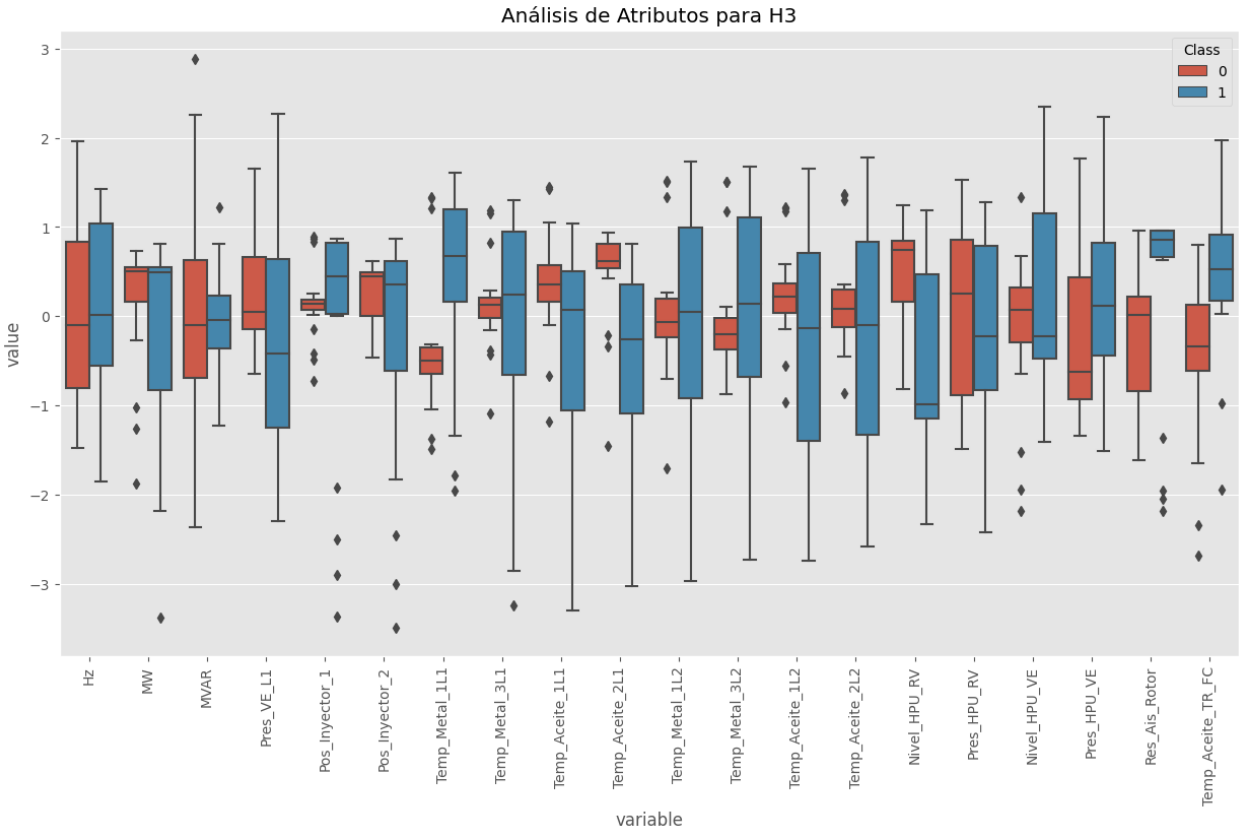


Fig. 40. Análisis de atributos para H3. Fuente: Elaboración propia.

Tres horas antes de un evento Fig. 40, se observaron algunas diferencias más marcadas en eventos de falla y de operación normal; es el caso de la Temperatura de Metal 1 Lado 1, Temperatura de Aceite 2 Lado 1 y Temperatura Aceite Transformador Fase C.

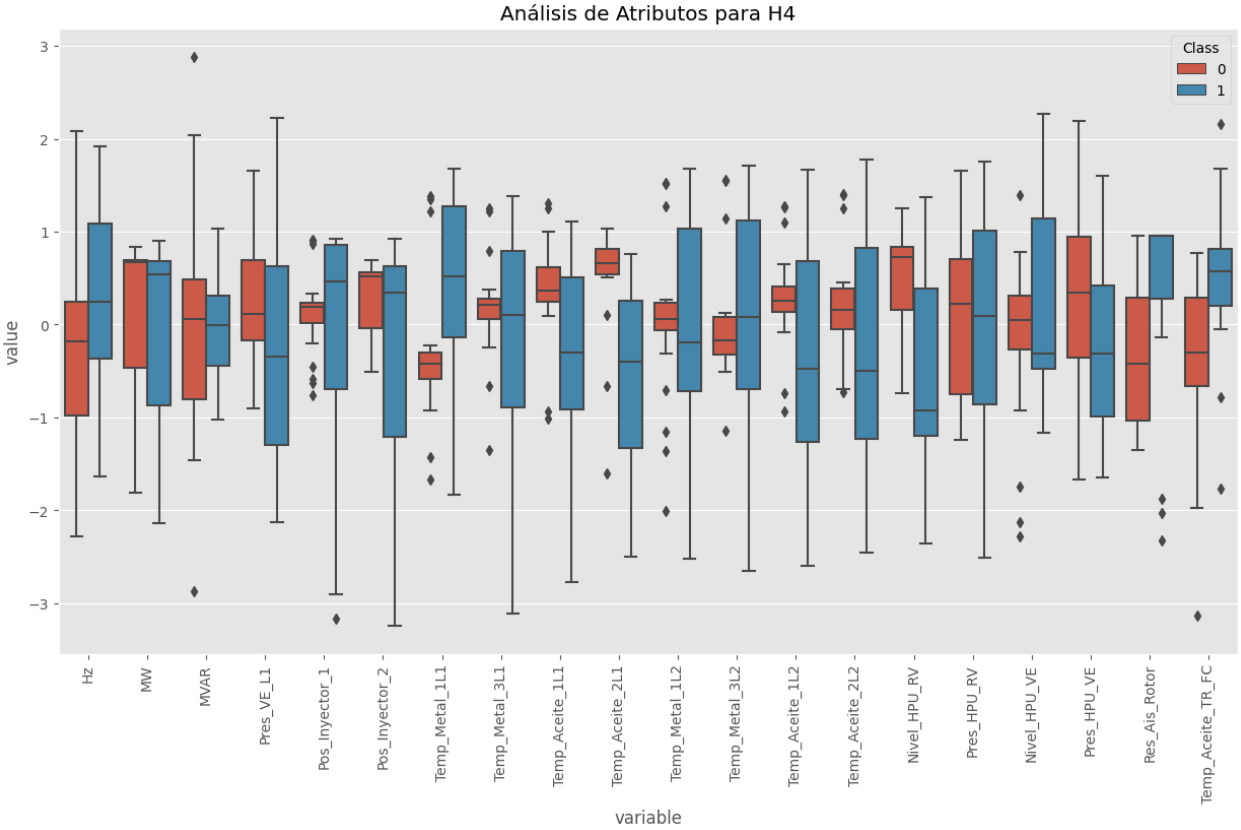


Fig. 41. Análisis de atributos para H4. Fuente: Elaboración propia.

Para el análisis de atributos cuatro horas antes de un evento Fig. 41, se apreció el mismo comportamiento que para tres horas antes; se observaron algunas variaciones muy marcadas de algunas variables.

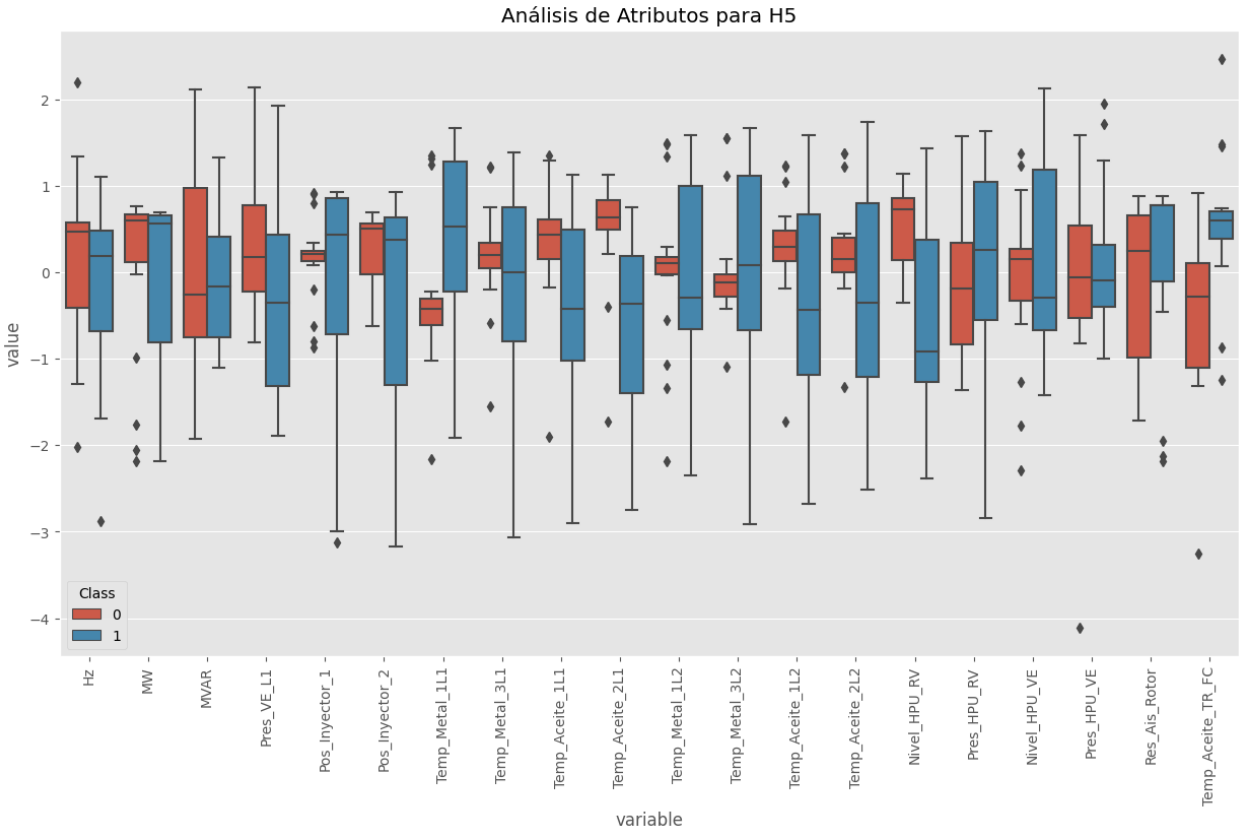


Fig. 42. Análisis de atributos para H5. Fuente: Elaboración propia.

Cinco horas antes de un evento Fig. 42, se observaron diferencias entre los valores de la Temperatura de Aceite del Transformador Fase C durante eventos de falla y durante periodos de tiempo de operación normal. Adicionalmente, se observaron diferentes rangos de variación para las variables Temperatura Metal 1 Lado 1 y Temperatura Aceite 2 Lado 1.

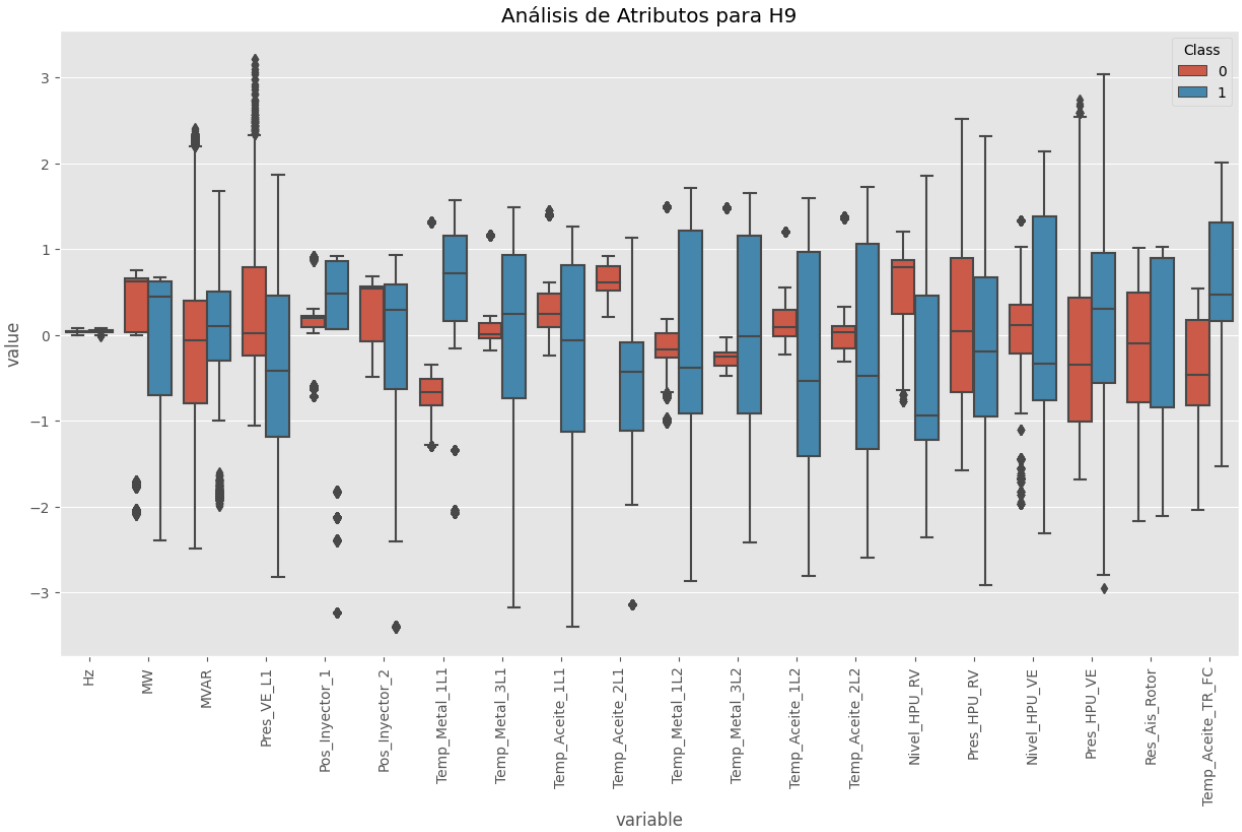


Fig. 43. Análisis de atributos para H9. Fuente: Elaboración propia.

De la Fig. 43, se puede identificar una mayor separación de la Temperatura de Aceite 2 Lado 1 frente a eventos de falla y períodos de tiempo donde no se registraron fallas, mientras que el resto de los atributos en su mayoría se comportan casi dentro de los mismos rangos.

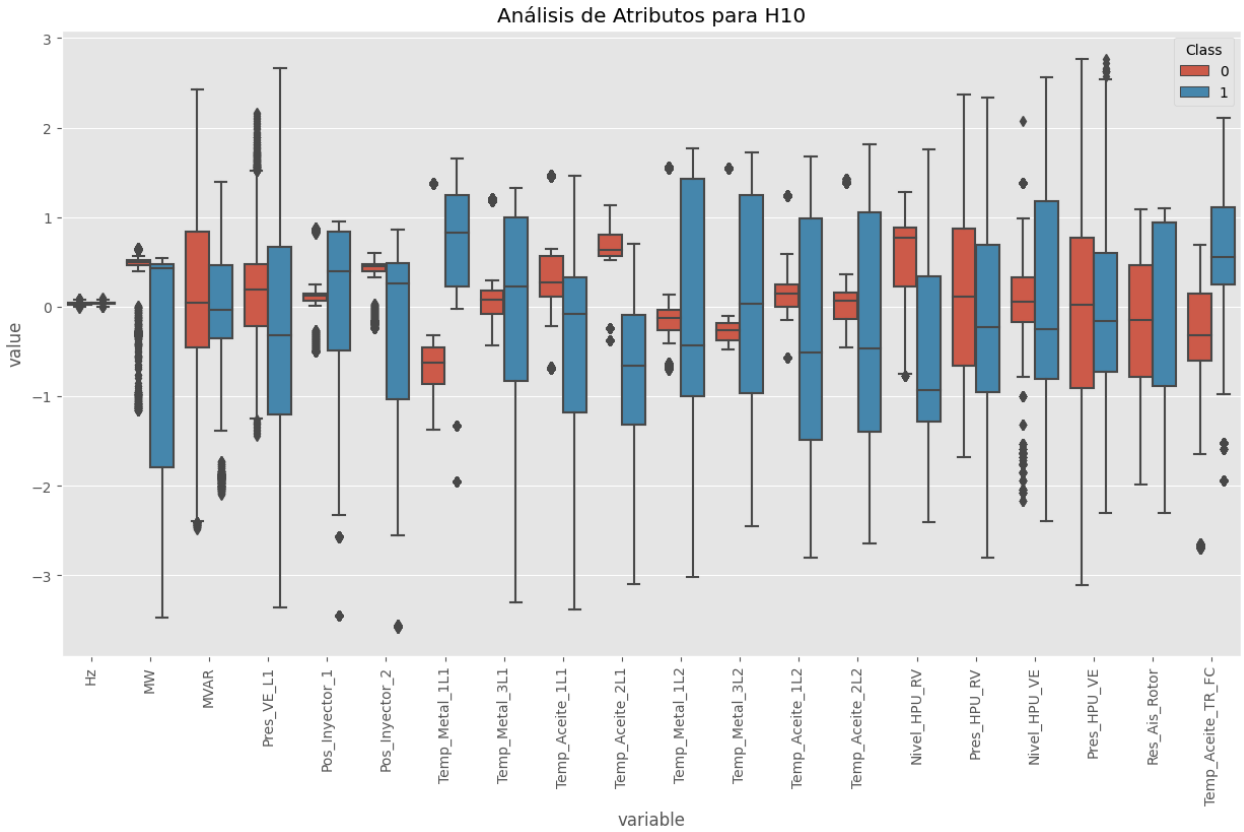


Fig. 44. Análisis de atributos para H10. Fuente: Elaboración propia.

El mismo efecto que para H9 se observa para H10, con la leve diferencia de la variación de los valores para la Temperatura de Aceite Transformador Fase C.

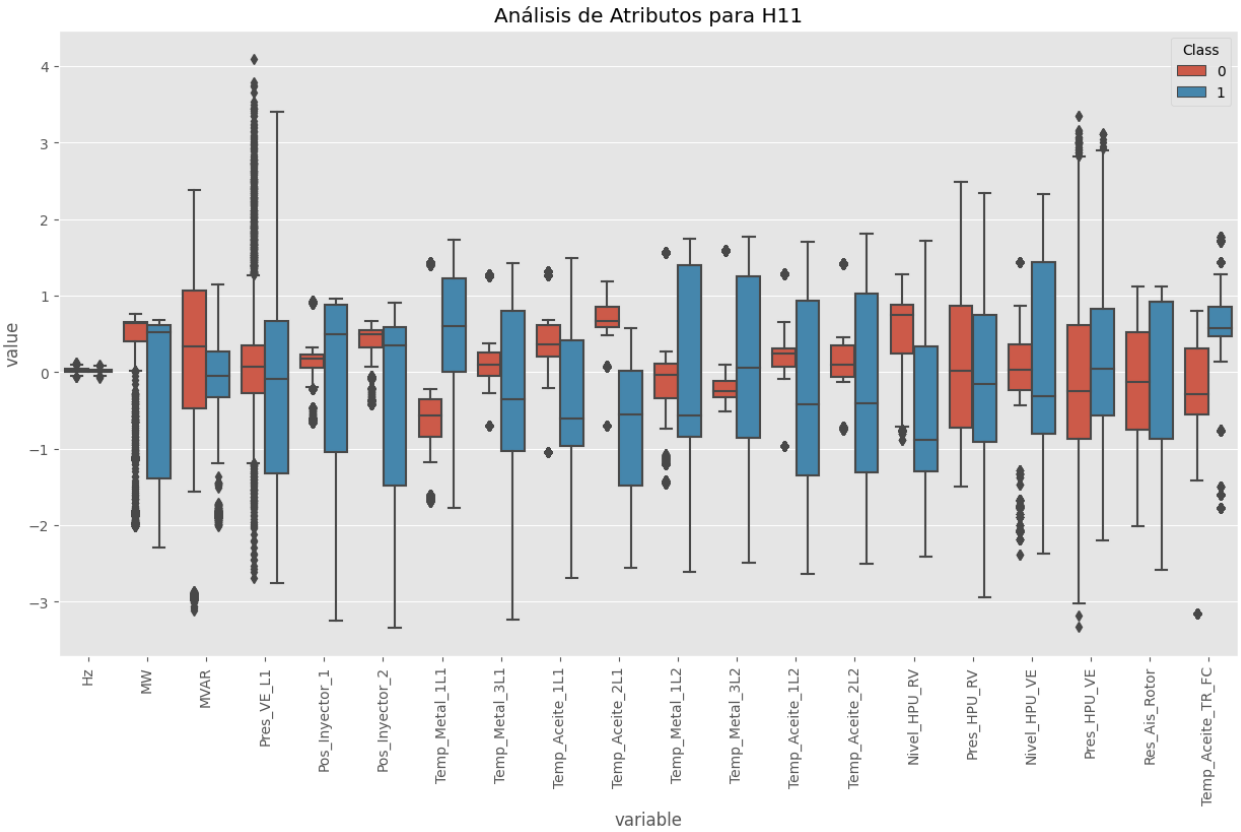


Fig. 45. Análisis de atributos para H11. Fuente: Elaboración propia.

Para el caso de H11, el comportamiento es similar a H10, resaltando las mismas variables con variación por fuera de los valores normales.

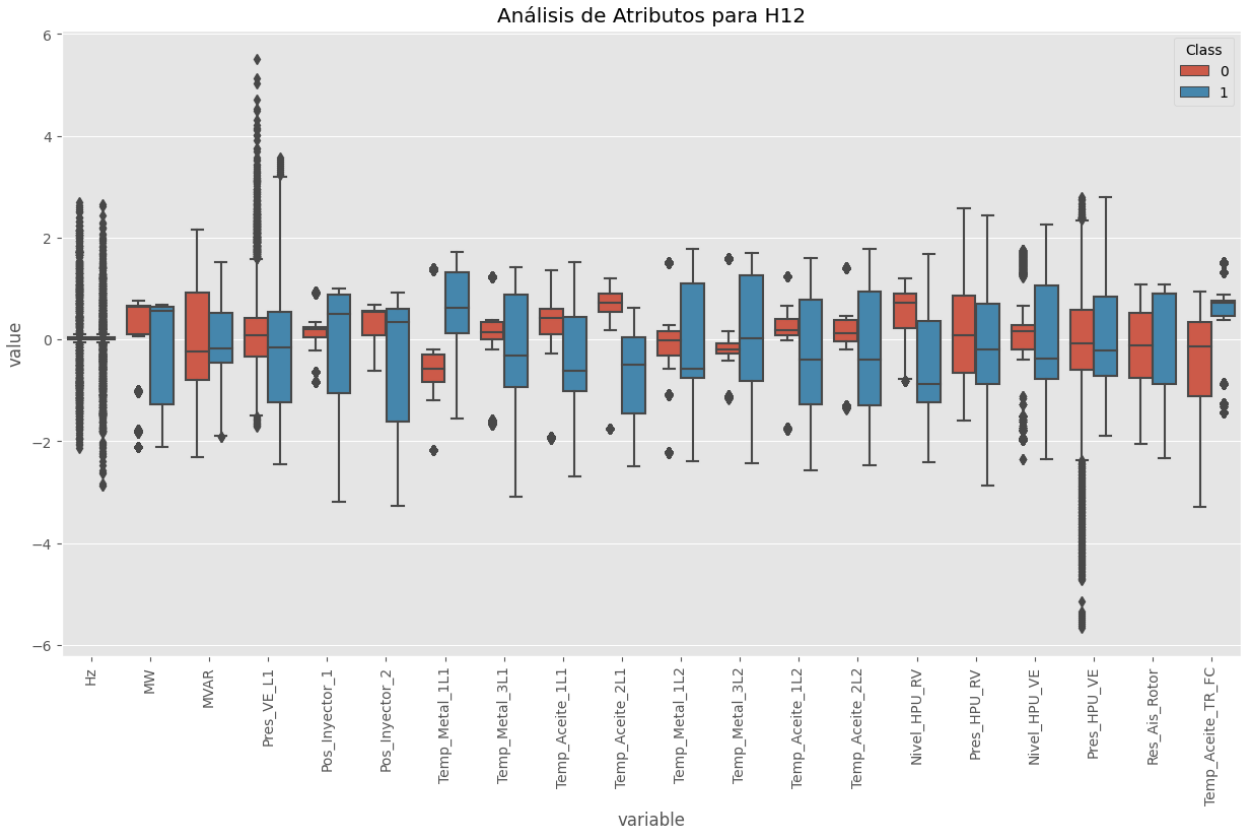


Fig. 46. Análisis de atributos para H12. Fuente: Elaboración propia.

Para H12, aunque se mantiene el mismo comportamiento de los periodos de tiempo anteriores, se observa que la diferencia entre los rangos de las variables de Temperatura Aceite 2 Lado 1, Temperatura Metal 1 Lado1 y Temperatura Aceite Transformador Fase C se ha disminuido.

Con los resultados expuestos para los periodos de tiempo de 9 a 12 horas se puede establecer que al menos una de las razones por las cuales se están clasificando correctamente todas las observaciones es por el rango de valores que toma la variable de Temperatura Aceite 2 Lado 1, ya se ha encontrado una diferencia amplia entre estos valores durante un evento de falla y durante espacios de tiempo que no se registraron fallas.

Con todo lo anterior, se pudo entender la distribución de los árboles de decisión creados para cada una de las ventanas de tiempo, de dos a cinco y de nueve a doce horas. En la Fig. 49, Fig. 50, Fig. 53 y Fig. 54, fue posible apreciar que la complejidad de árbol aumenta, mientras que para la Fig. 47, Fig. 48, Fig. 51 y Fig. 52, el árbol es mucho más sencillo.

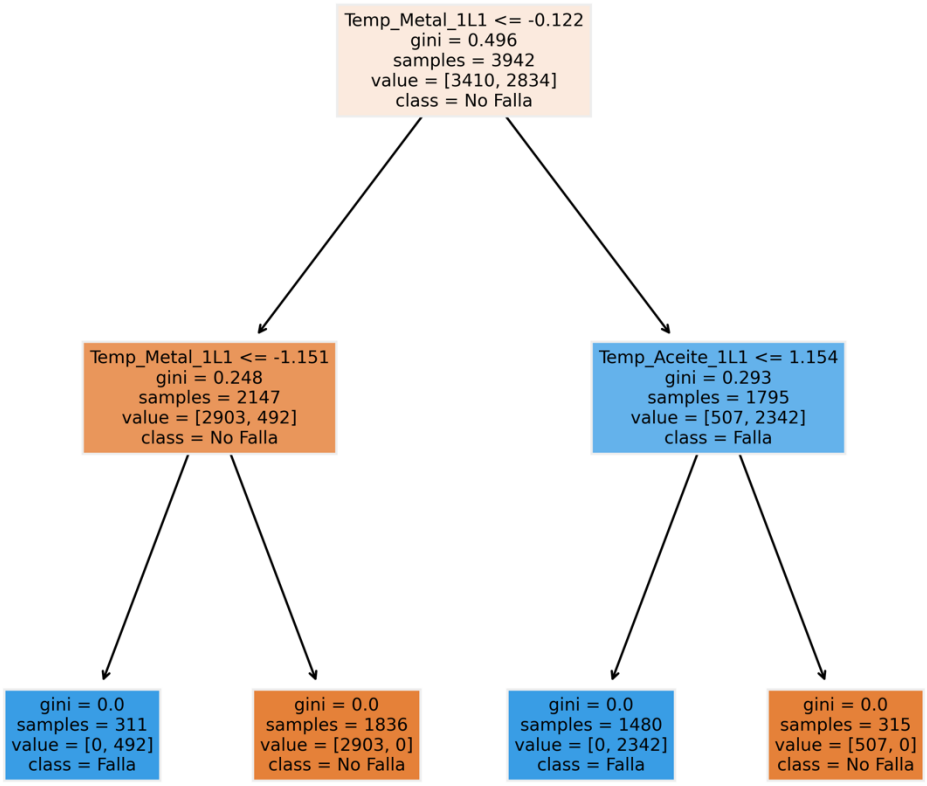


Fig. 47. Árbol de decisión para H2. Fuente: Elaboración propia.

Se observa que en la raíz del árbol está el atributo Temperatura Metal 1L1, la rama izquierda vuelve a validar el mismo atributo mientras la rama derecha valida el atributo Temperatura Aceite 1L1.

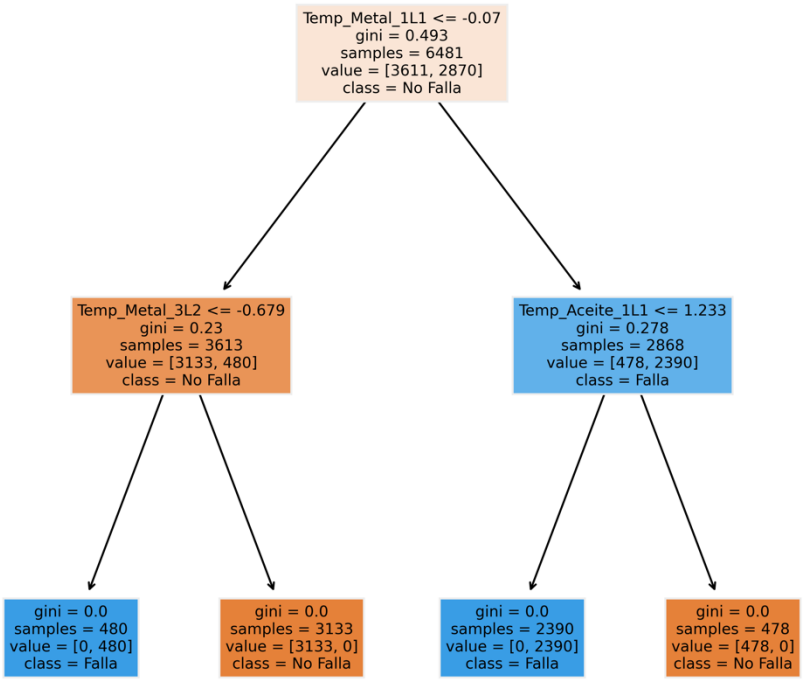


Fig. 48. Árbol de decisión para H3. Fuente: Elaboración propia.

Al igual que el árbol de decisión para H2, el árbol de decisión para H3 solo involucra atributos de temperaturas como Temperatura Metal 1L1, Temperatura Metal 3L2 y Temperatura Aceite 1L1.

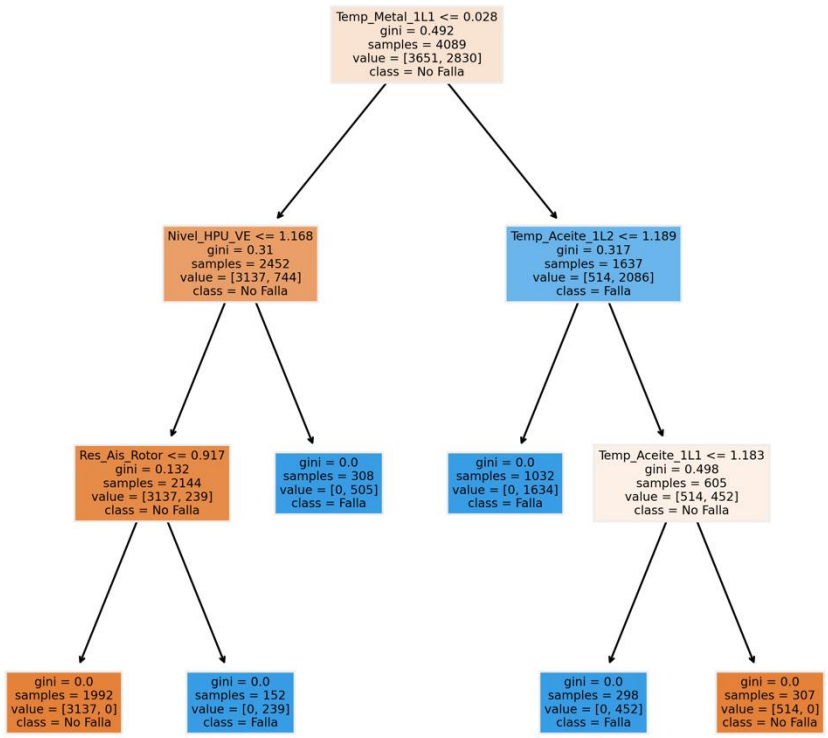


Fig. 49. Árbol de decisión para H4. Fuente: Elaboración propia.

Se observa que este árbol de decisión es más grande que los anteriores, mientras que las ramas del lado derecho solo involucran atributos de temperaturas, en las ramas del lado izquierdo aparecen dos nuevos atributos, Nivel HPU Válvula Esférica y Resistencia Aislamiento Rotor.

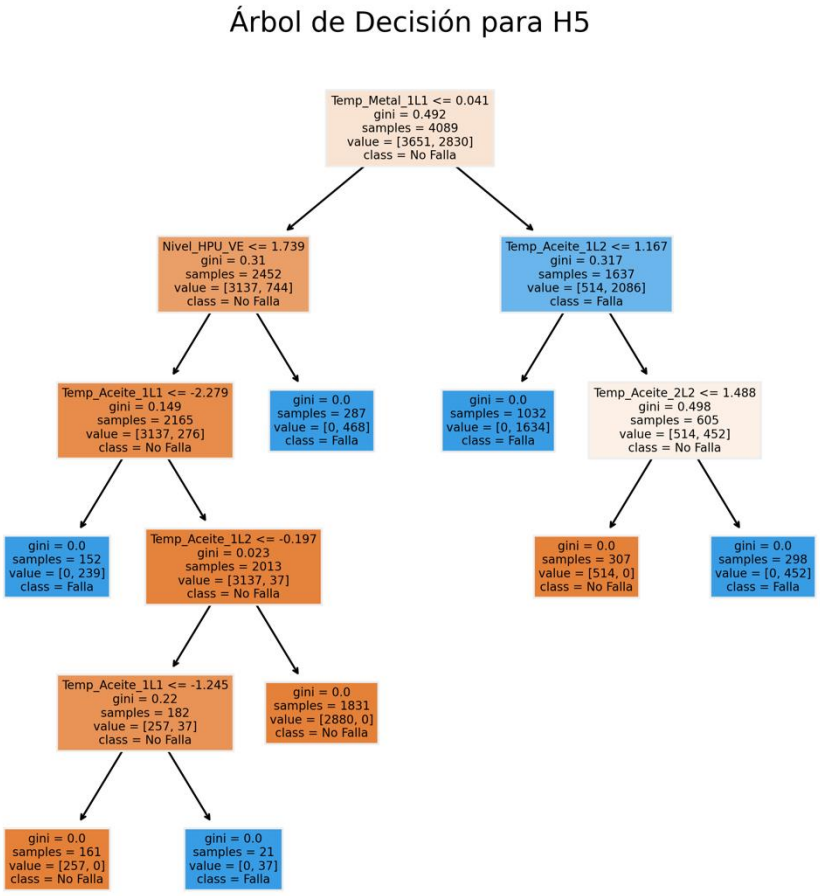


Fig. 50. Árbol de decisión para H5. Fuente: Elaboración propia.

Se observa que este árbol de decisión es más grande que los anteriores, las ramas del lado derecho involucran atributos de temperaturas, en las ramas del lado izquierdo aparte de los atributos de temperatura, también tiene el atributo Nivel HPU Válvula Esférica.

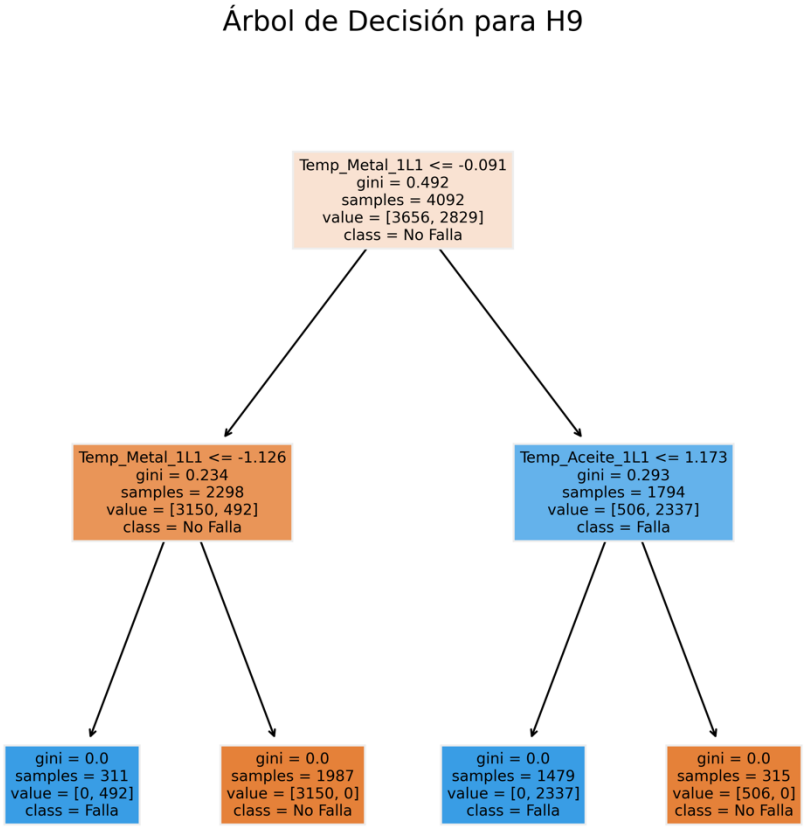


Fig. 51. Árbol de decisión para H9. Fuente: Elaboración propia.

Se observa que este árbol un árbol más compacto, parecido a los de H2 y H3, donde se evalúan solo los valores de temperaturas de metal 1 y aceite 1 del cojinete lado 1.

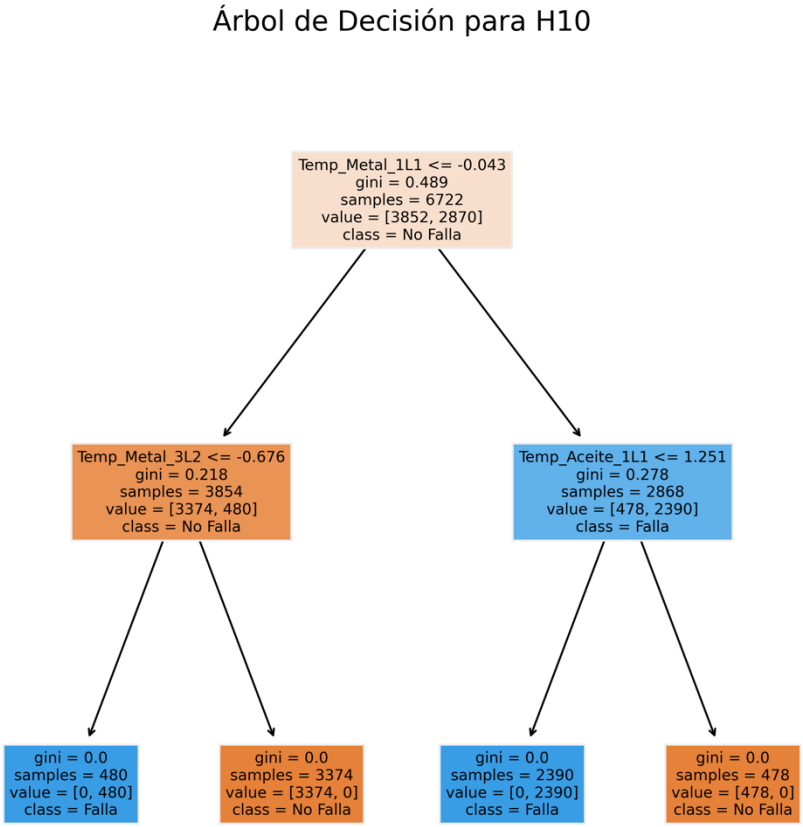


Fig. 52. Árbol de decisión para H10. Fuente: Elaboración propia.

En la figura de árbol de decisión anterior se observa un comportamiento similar al anteriormente señalado, se evalúan solo temperaturas de metal y aceite.

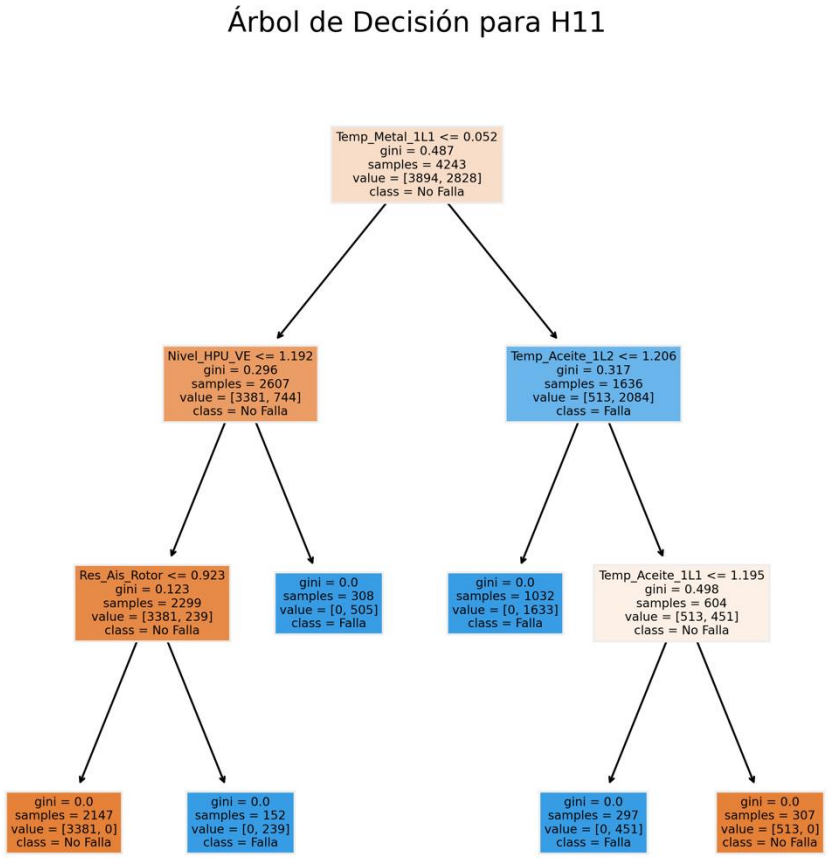


Fig. 53. Árbol de decisión para H11. Fuente: Elaboración propia.

En la Fig. 53 se observa la evaluación de algunas condiciones de temperaturas de metal y aceite en ambos cojinetes, 1 y 2. En la parte superior luego de la primera condición, se evalúa el nivel de la HPU de la válvula, mientras que del lado derecho se evalúa la temperatura de aceite 1 lado 2.

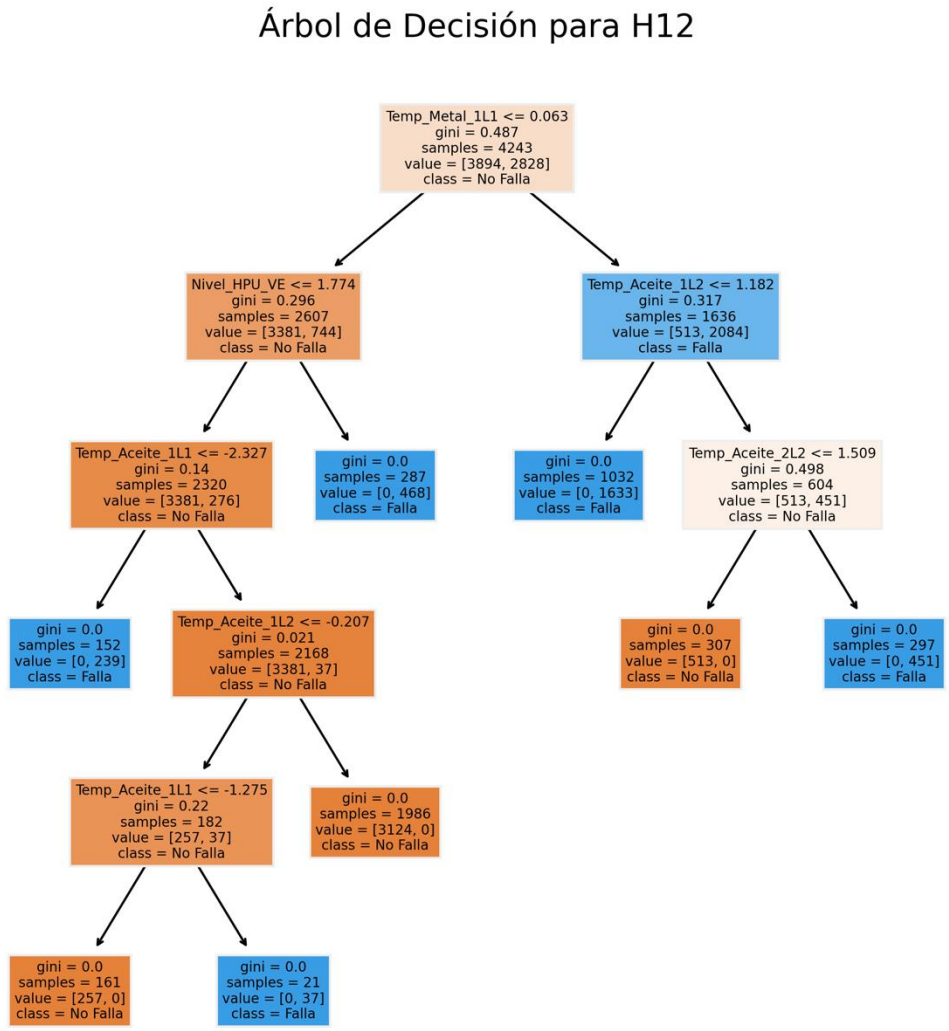


Fig. 54. Árbol de decisión para H12. Fuente: Elaboración propia.

Se observa un árbol parecido al de H5, pero a diferencia de árboles anteriores se observa un poco más de complejidad. Se observan evaluaciones de temperaturas de metal y aceite de cojinetes tanto del lado 1 como del lado 2 y niveles de la HPU de la válvula esférica.

En lo que corresponde al escenario donde se tomó solo una observación por periodo de tiempo, tanto para eventos de falla y eventos de no falla, se pudo identificar que las métricas obtenidas para cada algoritmo son similares con excepción de XGBOOST que tuvo un rendimiento menor; es decir, para el conjunto de datos con que se desarrolló el presente proyecto no fue estrictamente necesario tomar una ventana de tiempo, ya que una sola observación representó todo el periodo de tiempo definido desde dos horas antes hasta cinco horas antes de un evento.

Dicho esto, se realizó un gráfico para el atributo que los modelos de predicción identificaron como de mayor relevancia durante de distintos periodos de tiempo para observar su variación durante la ventana de tiempo de cinco (5) y doce (12) horas:

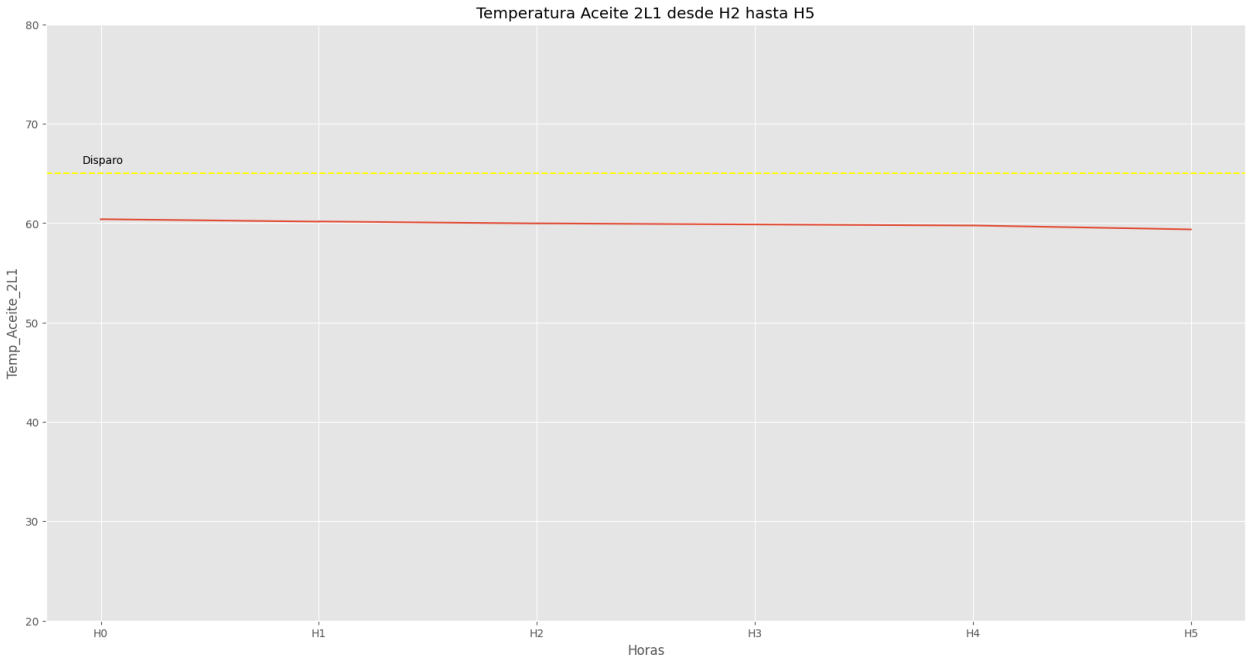


Fig. 55. Temperatura de Aceite 2L1 desde H2 hasta H5. Fuente: Elaboración propia.

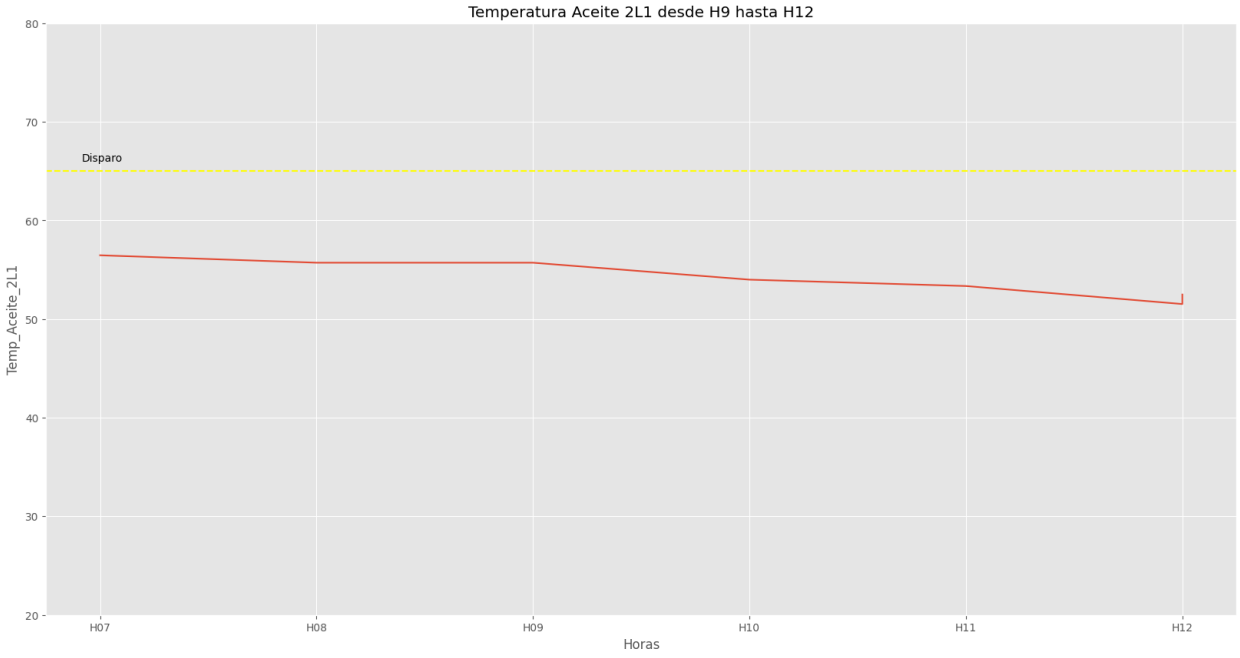


Fig. 56. Temperatura de Aceite 2L1 desde H2 hasta H5. Fuente: Elaboración propia.

De acuerdo con las dos figuras anteriores se pudo establecer que la variación de el atributo mencionado durante periodos de tiempo entre dos (2) y doce (12) horas es poca; con esto se puede explicar por qué con una sola observación se representaron los datos recopilados durante dicha ventana y se obtuvieron resultados similares, esto sin la carga a nivel de hardware necesaria para ejecutar las simulaciones.

6. CONCLUSIONES Y TRABAJOS FUTUROS

6.1. CONCLUSIONES

En la investigación propuesta fue posible identificar las variables más importantes en las turbinas de la central de generación en estudio por medio de la comparación del registro de indisponibilidades con el registro de eventos de estado de la unidad y el análisis exploratorio de los datos descargados del sistema SCADA, esto permitió seleccionar los mejores atributos para el entrenamiento del modelo de aprendizaje supervisado para la detección de fallas.

Fue posible identificar los atributos más representativos para cada modelo como: Temperatura de Aceite 2 Lado 1, Temperatura de Aceite Transformador Fase C, Temperatura de Aceite 2 Lado 2, Posición de Inyectores y Presión de la Válvula Esférica Lado 1 y analizar su comportamiento y comparación entre el conjunto de datos de fallas con el conjunto de datos de no fallas.

Tres modelos predictivos fueros seleccionados con el apoyo de la literatura consultada, y entrenados con los datos descargados de las unidades de generación, los tres modelos fueron: Random Forest, Support Vector Machine y XGBoost; los cuales presentaron buenos resultados en la evaluación de desempeño al momento de predecir fallos en las unidades de generación con porcentajes del indicador de desempeño seleccionado F2-score, por encima de 80%.

Para periodos de tiempo entre nueve y doce horas antes de un evento de falla se calcularon identificadores de desempeño F2-score con valor de uno (1); esto se atribuye a que los valores de *la Temperatura de Aceite 2 Lado 1,* a medida que el tiempo de observación aumenta de dos a doce horas, los valores que este atributo toma para eventos de falla y eventos de no falla se distancian el uno del otro, a pesar de que se encuentra dentro de los valores normales de operación.

A partir del conjunto de datos utilizados para el presente proyecto se pudo establecer que es posible realizar predicciones de fallas en generadores de centrales hidroeléctricas tomando solo una observación por evento, esto, debido que la variación que se identificó de los atributos es poca durante un periodo de tiempo de cuatro minutos.

Con la investigación realizada se pudo desarrollar un modelo capaz de predecir fallas en generadores de centrales hidroeléctricas con un porcentaje alto de exactitud que permitirá establecer las bases para construir un modelo de predicción más robusto que ayude al área operativa a mejorar la coordinación de la operación para evitar indisponibilidades no planeadas y penalizaciones económicas del operador de mercado eléctrico XM. Lo anterior, con la salvedad que para objeto del presente documento se consolidó información referente a pocas unidades de

generación y un número de eventos de falla reducido; además, se considera viable que se puedan generar alertas a partir de variaciones en ciertas medidas que, aun estando en rangos de operación normal pueden ser indicadores de un posible evento más adelante.

6.2. TRABAJOS FUTUROS

Como trabajos futuros se propone por un lado ajustar los sistemas de sensores de las unidades de generación para mejorar la calidad de la información que es recibida por el sistema SCADA ya que se encontraron muchas variables con datos erróneos que tuvieron que eliminarse del análisis.

Por otro lado, también se propone para un trabajo futuro integrar la información en tiempo real del sistema SCADA con un sistema como el de la plataforma ElasticSearch [19] para realizar ETL (extracción, transformación y carga) sobre los datos en tiempo real y posteriormente utilizar los modelos propuestos en este trabajo para generar alertas de predicción de indisponibilidades en tiempo real, permitiendo a los operadores de la central, realizar mantenimientos predictivos y evitar fallas que generen indisponibilidades en la generación y por ende multas para la empresa.

Finalmente, se considera que continuar con el estudio de los atributos de los sistemas cubiertos en el presente documento puede representar un beneficio para la construcción de alertas tempranas, pero siendo necesario obtener muchos más datos relacionados con las fallas de los generadores.

7. REFERENCIAS BIBLIOGRÁFICAS

[1] CREG, “Resolución 24 de 1995 CREG”. https://gestornormativo.creg.gov.co/gestor/entorno/docs/resolucion_creg_0024_1995.htm# Inicio (consultado el 28 de septiembre de 2022).

[2] XM, “¿Quiénes somos en XM ?” https://www.xm.com.co/nuestra-empresa/nosotros/quienes-somos (consultado el 28 de septiembre de 2022).

[3] A. Rodríguez Penin, *Sistemas SCADA*. 2011.

[4] UPME, “Primer Atlas hidroenergético revela gran potencial en Colombia”. https://www1.upme.gov.co/Paginas/Primer-Atlas-hidroenergetico-revela-gran-potencial-en-Colombia.aspx (consultado el 28 de septiembre de 2022).

[5] J. M. Molina López, A. Berlanga, M. Á. Patricio Guisado, Á. L. Bustamante, y W. R. Padilla, *Ciencia de datos : técnicas analíticas y aprendizaje estadístico, un enfoque práctico*. Bogotá: Alfaomega Colombiana , 2018.

[6] L. Judith Sandoval, “ALGORITMOS DE APRENDIZAJE AUTOMÁTICO PARA ANÁLISIS Y PREDICCIÓN DE DATOS”, *REVISTA TECNOLÓGICA*, vol. 11, 2018.

[7] P. Bangert, *Machine learning and data science in the power generation industry*.

[8] J. Mulongo, M. Atemkeng, T. Ansah-Narh, R. Rockefeller, G. M. Nguegnang, y M. A. Garuti, “Anomaly Detection in Power Generation Plants Using Machine Learning and Neural Networks”, *Applied Artificial Intelligence*, vol. 34, núm. 1, pp. 64–79, ene. 2020, doi: 10.1080/08839514.2019.1691839.

[9] S. X. Anucha Promwungkwa y K. Ngamsanroaj, “Application of Machine Learning for Predictive Maintenance Cooling System inNam Ngum-1 Hydropower Plant”, *Sixteenth International Conference on ICT and Knowledge Engineering*, 2018.

[10] A. R. de A. Vallim Filho, D. Farina Moraes, M. V. Bhering de Aguiar Vallim, L. S. da Silva, y L. A. da Silva, “A Machine Learning Modeling Framework for Predictive Maintenance Based on Equipment Load Cycle: An Application in a Real World Case”, *Energies (Basel)*, vol. 15, núm. 10, may 2022, doi: 10.3390/en15103724.

[11] A. Betti, E. Crisostomi, G. Paolinelli, A. Piazzi, F. Ruffini, y M. Tucci, “Condition monitoring and predictive maintenance methodologies for hydropower plants equipment”, *Renew Energy*, vol. 171, pp. 246–253, jun. 2021, doi: 10.1016/j.renene.2021.02.102.

[12] P. Calvo-Bascones, M. A. Sanz-Bobi, y T. M. Welte, “Anomaly detection method based on the deep knowledge behind behavior patterns in industrial components. Application to a hydropower plant”, *Comput Ind*, vol. 125, feb. 2021, doi: 10.1016/j.compind.2020.103376.

[13] D. Cao, Y. Zhou, y L. Pan, “Statistics analysis and alarm threshold setting for monitoring data of hydropower unit”, *The Proceedings of the International Conference on Power Engineering (ICOPE)*, vol. 2015.12, núm. 0, p. _ICOPE-15--_ICOPE-15-, 2015, doi: 10.1299/jsmeicope.2015.12._ICOPE-15-_163.

[14] A. R. de A. Vallim Filho, D. Farina Moraes, M. V. Bhering de Aguiar Vallim, L. S. da Silva, y L. A. da Silva, “A Machine Learning Modeling Framework for Predictive Maintenance Based on Equipment Load Cycle: An Application in a Real World Case”, *Energies (Basel)*, vol. 15, núm. 10, may 2022, doi: 10.3390/en15103724.

[15] D. Adhya, S. Chatterjee, y A. K. Chakraborty, “Performance assessment of selective

machine learning techniques for improved PV array fault diagnosis", *Sustainable Energy, Grids and Networks*, vol. 29, mar. 2022, doi: 10.1016/j.segan.2021.100582.

[16] J. L. Speiser, M. E. Miller, J. Tooze, y E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling", *Expert Systems with Applications*, vol. 134. Elsevier Ltd, pp. 93–101, el 15 de noviembre de 2019. doi: 10.1016/j.eswa.2019.05.028.

[17] M. Kubat, *An Introduction to Machine Learning*. Springer International Publishing, 2017. doi: 10.1007/978-3-319-63913-0.

[18] "sklearn.metrics.precision_recall_fscore_support — scikit-learn 1.2.2 documentation". https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.html (consultado el 27 de mayo de 2023).

[19] "Elasticsearch". https://www.elastic.co/es/ (consultado el 6 de marzo de 2023).