Acta de Correcciones al Documento de Trabajo de Grado

Santiago de Cali, 15 de julio del 2024

Autor: Keyner Martínez Miranda

Título del Trabajo de Grado: "Pronóstico de disponibilidad de los recursos de generación de la central TermoGuajira a partir de modelos de aprendizaje automático"

Director: David Arango Londoño

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.

David Arango Londoño

_____

Firma del Director del Trabajo de Grado

Santiago de Cali, 06 de junio del 2024

Doctor
Diego Luis Linares
Director 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 "Pronóstico de la disponibilidad de los recursos de generación de la central Termoguajira a partir de modelos de aprendizaje automático", el cual fue realizado por el estudiante Keyner Hernando Martínez Miranda con código 8980308 pertenecientes a la Maestría en Ciencia de Datos, bajo la dirección de David Arango Londoño.

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,

Keyner Martínez Miranda	David Arango Londoño
C.C. 1.140.896.879 de Barranquilla	C.C. 1.130.586.950 de Cali

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

Pontificia Universidad Javeriana Cali
Facultad de Ingeniería y Ciencias
Maestría en Ciencia de Datos
Proyecto Aplicado

Pronóstico de disponibilidad de los recursos de generación de la central TermoGuajira a partir de modelos de aprendizaje automático

Keyner Martínez Miranda

Director: Dr. David Arango Londoño

15 de julio de 2024

Ficha Resumen

TÍTULO: Pronóstico de la disponibilidad de los recursos de generación de la central Termo-Guajira a partir de modelos de aprendizaje automático.

1. **ÉNFASIS:** Sistemas y Computación
2. **TIPO DE PROYECTO:** Aplicado
3. **ÁREA DE TRABAJO:** Sector eléctrico
4. **ESTUDIANTES(S):** Keyner Martínez Miranda
5. **CORREO ELECTRÓNICO:** keynerm@javerianacali.edu.co
6. **DIRECCIÓN Y TELÉFONO:** Calle 1e N° 21-89, (+57) 323 2857730
7. **DIRECTOR:** David Arango Londoño
8. **VINCULACIÓN DEL DIRECTOR (en la universidad):** Cátedra
9. **CORREO ELECTRÓNICO DEL DIRECTOR:** david.arango@javerianacali.edu.co
10. **OTROS GRUPOS O EMPRESAS:** GENERADORA Y COMERCIALIZADORA DE ENERGÍA DEL CARIBE S.A. E.S.P.
11. **PALABRAS CLAVE (al menos 5)** Ciencia de datos, Machine Learning, Sistema Eléctrico Colombiano, plantas de generación eléctrica y disponibilidad comercial.
12. **ODS QUE APLICA EL PROYECTO (Agenda 2030):** Energía asequible y no contaminante (7).
13. **FECHA DE INICIO (Desarrollo del proyecto):** 01/07/2023
14. **RESUMEN (máximo 400 palabras):** El Centro Nacional de Despacho (CND) ha identificado restricciones eléctricas en la subárea GCM del sistema eléctrico colombiano, lo cual ha llevado a declarar un estado de emergencia desde abril de 2022. En el estado actual del sistema eléctrico, la disponibilidad de los recursos de generación internos en esta subárea es crucial para garantizar la seguridad y confiabilidad del sistema eléctrico, ya que su ausencia puede desencadenar eventos no deseados y afectar a los usuarios finales. Por lo tanto, el objetivo de este proyecto es desarrollar un modelo a través de técnicas de aprendizaje automático, con el fin de implementar medidas preventivas y estrategias de contingencia que minimicen el riesgo de indisponibilidades no programadas y aseguren el suministro eléctrico confiable. El proyecto seguirá pasos metodológicos, como el análisis exploratorio de datos, el desarrollo del modelo de machine learning y la validación de las predicciones generadas.

Índice general

1. Introducción **1**

2. Definición Del Problema **3**
2.1. Planteamiento Del Problema . . . 3
2.2. Formulación Del Problema . . . 6

3. Objetivos Del Proyecto **7**
3.1. Objetivo General . . . 7
3.2. Objetivos Específicos . . . 7
3.3. Resultados Esperados . . . 7

4. Marco Teórico y Antecedentes **9**
4.1. Marco Teórico . . . 9
4.2. Antecedentes . . . 15

5. Desarrollo del proyecto **19**
5.1. Fase de planificación . . . 19
5.1.1. Entendimiento del negocio . . . 19
5.1.2. Recopilación y entendimientos de los datos . . . 21
5.2. Desarrollo del Modelo . . . 22
5.2.1. Recopilación de datos . . . 22
5.2.2. Preparación de datos . . . 24
5.2.3. Análisis Exploratorio de los datos . . . 26
5.2.4. Selección de variables . . . 28
5.2.5. Metodología de construcción del dataset . . . 30
5.2.6. Cálculo de correlaciones . . . 35
5.3. Modelamiento . . . 38
5.3.1. Selección de algoritmos . . . 38
5.3.2. Logistic Regression . . . 38
5.3.3. Decision Tree Classifier . . . 45
5.3.4. Random Forest Classifier . . . 52
5.3.5. Comparación de modelos . . . 59

6. Conclusiones y trabajos futuros **63**
6.1. Conclusiones . . . 63
6.2. Trabajos Futuros . . . 64
6.2.1. Ámbito Local . . . 64
6.2.2. Ámbito Global . . . 64

6.2.3. Ampliación de la Investigación 64

7. Anexos **67**

7.1. Repositorio Público 67

Bibliografía **69**

Índice de figuras

2.1. Sistema eléctrico Colombiano. *fuente XM [1]* . . . 3
2.2. Área Caribe, subáreas y Caribe 2. *fuente XM [1]* . . . 4

4.1. Ejemplo de fenómeno de recuperación lenta inducida de tensión ante evento. *fuente XM [1]* . . . 10
4.2. Representación de la evolución de la tensión ante un transitorio en el tiempo. *fuente XM [1]* . . . 10

5.1. Vista esquemática de un proceso de energía a vapor de agua. *fuente agora [2]* . . . 19
5.2. Parámetros técnicos de centrales de generación. *fuente propia* . . . 20
5.3. Aplicativo SINERGOX. *fuente XM [3]* . . . 21
5.4. Aplicativo PARATEC. *fuente XM [4]* . . . 22
5.5. Datos históricos aplicativo SINGERGOX. *fuente XM [3]* . . . 23
5.6. Histograma de la variable generación de energía. *fuente propia* . . . 26
5.7. Histograma de la variable disponibilidad de energía. *fuente propia* . . . 26
5.8. Consumo de combustible por tipo. *fuente propia* . . . 27
5.9. Correlación de variables. *fuente propia* . . . 27
5.10. Aplicativo SINERGOX. *fuente propia* . . . 30
5.11. Esquemático de construcción del dataset. *fuente propia* . . . 33
5.12. Boxplot de los indicadores de tiempo de operación. *fuente propia* . . . 35
5.13. Boxplot del indicador de combustible - Heat Rate. *fuente propia* . . . 35
5.14. Correlación de las variables. *fuente propia* . . . 36
5.15. Modelo Logistic Regression. *fuente propia* . . . 39
5.16. Confusion matrix del modelo Logistic Regression. *fuente propia* . . . 41
5.17. Logistic Regression Classification Report. *fuente propia* . . . 42
5.18. Feature Importance Plot - Logistic Regression. *fuente propia* . . . 43
5.19. Validation Curve for Logistic Regression. *fuente propia* . . . 44
5.20. Modelo Decision TreeClassifier . *fuente propia* . . . 45
5.21. Confusion matrix - Decision TreeClassifier. *fuente propia* . . . 47
5.22. Classification Report - Decision TreeClassifier. *fuente propia* . . . 48
5.23. Feature importance - Decision TreeClassifier. *fuente propia* . . . 50
5.24. Validation Curbe for Decision TreeClassifier. *fuente propia* . . . 51
5.25. Modelo Random ForestClassifier. *fuente propia* . . . 52
5.26. Confusion matrix - Random ForestClassifier. *fuente propia* . . . 54
5.27. Classification Report - Random ForestClassifier. *fuente propia* . . . 55
5.28. Feature importance - Random ForestClassifier. *fuente propia* . . . 57
5.29. Validation Curve for Random ForestClassifier. *fuente propia* . . . 58

7.1. Repositorio público GitHub. *fuente propia* 67

Índice de cuadros

5.1. Descripción del dataset recopilado de SINGERGOX. 23
5.2. Parámetros técnicos general de la central termoguajira 24
5.3. Disponibilidad de información por variable 25
5.4. Dataset inicial: Recopilación del cálculo de los indicadores 32
5.5. Dataset Final: Recopilación del cálculo de los indicadores 34
5.6. Comparación de modelos 59
5.7. Comparación de modelos 60

Capítulo 1

Introducción

A partir de los análisis eléctricos realizados por el Centro Nacional de Despacho (CND), se han identificado las restricciones eléctricas y operativas en cada una de las subáreas del sistema eléctrico colombiano. En particular, la subárea GCM ha sido objeto de atención debido a que, desde el 1 de abril del 2022, se declaró en estado de emergencia. Donde se ha determinado que, para garantizar condiciones seguras de operación frente a restricciones relacionadas con la necesidad de unidades para el soporte de tensión o la presencia del fenómeno de FIDVR, resulta fundamental contar con la disponibilidad de los recursos de generación internos en esta subárea. Estos recursos desempeñan un papel crucial al asegurar la estabilidad y confiabilidad del sistema eléctrico, permitiendo hacer frente de manera eficiente y oportuna a las demandas y desafíos específicos que se presenten en la subárea GCM [1].

Es importante destacar que cualquier indisponibilidad no programada de los recursos de generación internos en la subárea GCM puede tener consecuencias significativas en la operación del sistema eléctrico. Estas situaciones imprevistas pueden desencadenar una serie de eventos no deseados, como la desconexión de carga, que afecta directamente a los usuarios finales y puede ocasionar interrupciones en el suministro de energía. Además, estas indisponibilidades pueden generar eventos en cascada, donde el desequilibrio entre la oferta y la demanda de electricidad se amplifica, afectando a otras áreas del sistema y propagando el impacto negativo. Esta cadena de eventos puede llegar a comprometer la estabilidad de voltaje en la subárea GCM y, en casos extremos, incluso provocar inestabilidades más amplias en el sistema eléctrico.

En consecuencia, el objetivo principal de este proyecto fue desarrollar un modelo predictivo para determinar la disponibilidad de los recursos de generación en la subárea GCM. Esta predicción es crucial para implementar medidas preventivas y desarrollar estrategias de contingencia efectivas, con el fin de minimizar el riesgo de indisponibilidades no programadas y garantizar la seguridad y confiabilidad del suministro eléctrico en dicha subárea. El modelo se basa en una clasificación binaria de la variable disponibilidad, distinguiendo entre unidades disponibles e indisponibles. Este enfoque permite anticipar posibles problemas y tomar decisiones informadas para mantener la estabilidad del sistema eléctrico.

Para lograrlo, se siguieron una serie de pasos metodológicos. En primer lugar, se llevó a cabo un análisis exploratorio de datos para identificar las variables clave que afectan la disponibilidad de los recursos de generación en la subárea GCM. Una vez completado este análisis, se procedió al desarrollo de un modelo de machine learning diseñado específicamente para pronósticar la disponi-

bilidad de los recursos de generación mediante la clasificación binaria. Por último, se realizó una validación exhaustiva para evaluar la precisión y confiabilidad de los pronósticos generados por el modelo.

En cuanto a la estructura del presente proyecto, en las siguientes secciones se describe en detalle la problemática, los objetivos, marco de referencia, metodología para el desarrollo del proyecto, resultados y conclusiones.

Capítulo 2

Definición Del Problema

2.1. Planteamiento Del Problema

El sistema eléctrico colombiano, por sus características geográficas, topología (red) y por la ubicación de sus parques de generación, se divide en cinco áreas operativas (*ver figura* 2.1): Antioquia, Caribe, Nordeste, Oriental y Suroccidente.

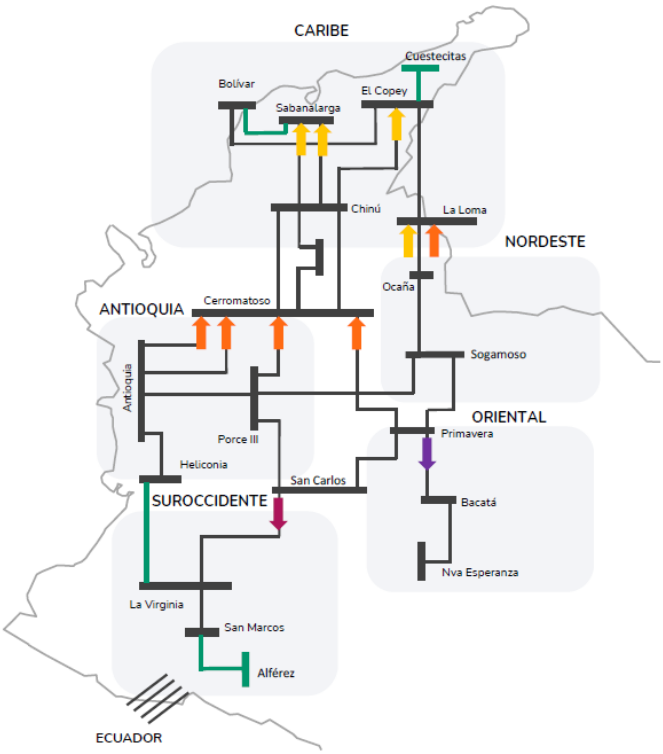

Figura 2.1: Sistema eléctrico Colombiano. *fuente XM [1]*

Actualmente, en el sistema eléctrico colombiano se han identificado un número significativo de restricciones, siendo un total de 163 restricciones reportadas en la actualidad. Es importante resaltar que en el área Caribe del país se concentra aproximadamente el 46 % de la totalidad de las restricciones del sistema. Además, se destaca que, dentro del área Caribe, el 82 % de estas restricciones

se encuentran en estado de emergencia. Esto significa que, ante la ocurrencia de una contingencia sencilla, se violan los límites de seguridad establecidos o no se puede atender la totalidad de la demanda eléctrica de manera adecuada [5].

El área Caribe se compone de las subáreas Cerromatoso y Córdoba – Sucre, y de las subáreas Atlántico, Bolívar y Guajira – Cesar – Magdalena GCM (Caribe 2), (*ver figura* 2.2) .

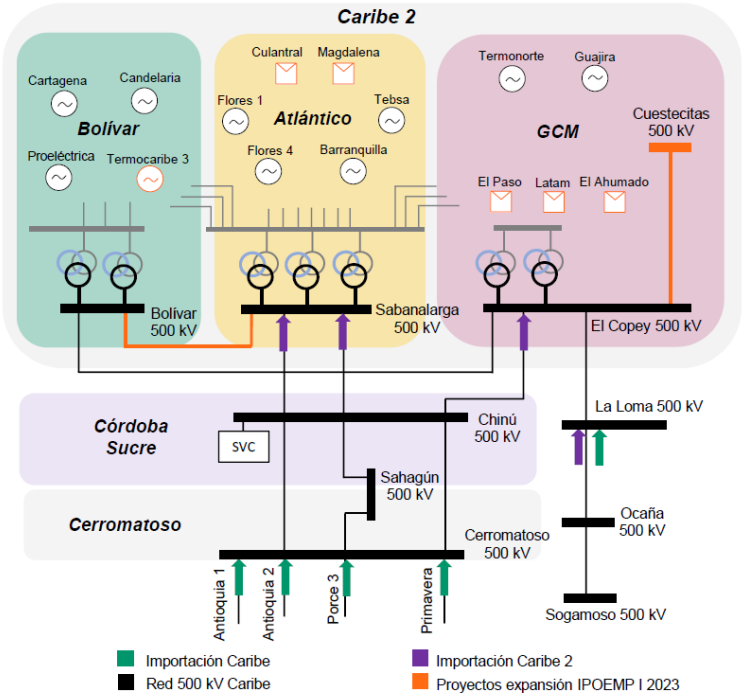

Figura 2.2: Área Caribe, subáreas y Caribe 2. *fuente XM [1]*

A partir del análisis de seguridad eléctrica del área Caribe, se identificó que, para mantener condiciones seguras de operación, se debe cumplir las siguientes recomendaciones asociadas a: i) límite seguro de importación; ii) requerimiento mínimo número de unidades en línea por seguridad; iii) operar dentro de rangos recomendados de tensión; y iv) realizar un adecuado control de potencia reactiva [1, pp. 26-35].

Desde el 1 de abril de 2022, se ha declarado el estado de emergencia en la subárea GCM debido a las evidencias de su susceptibilidad a experimentar el fenómeno de recuperación lenta inducida de tensión (FIDVR), principalmente debido a su condición de bajo nivel de cortocircuito [1, p. 34].

Se destaca la importancia de abordar el fenómeno de FIDVR mediante la programación de todos los recursos de generación síncrona disponibles dentro de la subárea, especialmente en escenarios

de demanda media y máxima. El objetivo es aumentar el nivel de cortocircuito y fortaleza de la red para reducir la profundidad del hueco de tensión en caso de producirse una perturbación en la subárea, lo que contribuye a mitigar la posible ocurrencia de este fenómeno.

Además, el análisis de seguridad eléctrica ha permitido identificar las restricciones eléctricas y operativas en la subárea GCM, donde se destaca el requerimiento de unidades para soporte de tensión.

La necesidad de unidades para brindar soporte de tensión destaca que solo las unidades conectadas internamente cumplen la función adicional de controlar las restricciones que limitan la importación de potencia activa y mejoran la robustez de la red al proporcionar una corriente efectiva de cortocircuito a los nodos de la subárea. Por lo tanto, se sugiere priorizar las unidades conectadas internamente para cumplir con los requisitos equivalentes de soporte de tensión de la subárea GCM[1, p.32].

Por lo anterior, es fundamental resaltar que cualquier indisponibilidad no programada de los recursos de generación internos en la subárea GCM puede tener consecuencias significativas en la operatividad del sistema eléctrico. Estas situaciones imprevistas pueden desencadenar una serie de eventos indeseados, como la desconexión de carga, lo cual afecta directamente a los usuarios finales y puede resultar en interrupciones del suministro de energía. Además, estas indisponibilidades pueden dar lugar a eventos en cascada, donde el desequilibrio entre la oferta y la demanda de electricidad se amplifica, afectando a otras áreas del sistema y propagando el impacto negativo. Esta secuencia de eventos puede comprometer la estabilidad del voltaje en la subárea GCM y, en casos extremos, incluso provocar inestabilidades más amplias en el sistema eléctrico.

Estos datos ponen de manifiesto la relevancia de abordar de manera efectiva el problema de la seguridad eléctrica de la subárea GCM a través de la disponibilidad de los recursos de generación internos.

Ante esta situación, la aplicación de modelos de aprendizaje automático en el contexto de la seguridad eléctrica de la subárea GCM es un enfoque altamente relevante y prometedor. Estos modelos ofrecen la capacidad de realizar pronósticos precisos y confiables sobre la disponibilidad de los recursos de generación en la subárea GCM [6].

La utilización de modelos de aprendizaje automático nos brinda la oportunidad de aprovechar la gran cantidad de datos disponibles, así como las relaciones complejas entre variables, para realizar pronósticos precisos sobre la disponibilidad de los recursos de generación. Estos modelos pueden analizar patrones históricos, condiciones operativas y otros datos relevantes para generar predicciones confiables.

La aplicación de modelos de aprendizaje automático en este contexto también ofrece beneficios

adicionales, como la capacidad de adaptarse y aprender de los nuevos datos a medida que se van generando [7]. Esto permite mejorar continuamente los pronósticos y ajustar las estrategias preventivas y de contingencia a medida que evoluciona el panorama de generación y operación.

En conclusión, al implementar medidas preventivas y estrategias de contingencia basadas en los pronósticos generados por los modelos de aprendizaje automático, es posible anticiparse a posibles indisponibilidades no programadas y tomar acciones preventivas para garantizar la seguridad y confiabilidad del suministro eléctrico en la subárea GCM.

2.2. Formulación Del Problema

El objetivo de esta investigación es responder a la pregunta general:

¿Cómo pronosticar la disponibilidad de los recursos de generación de la subárea GCM a partir de modelos de aprendizaje automático?

Para lograrlo, se plantearon las siguientes preguntas de sistematización:

1. ¿Qué variables inciden en la disponibilidad de los recursos de generación térmicos?
2. ¿Existe una correlación entre la disponibilidad de los recursos de generación con las variables predictoras?
3. ¿Las técnicas de machine learning permiten realizar predicciones de corto plazo para la disponibilidad de los recursos de generación?
4. ¿Cómo validar las predicciones de disponibilidad de los recursos de generación?

Capítulo 3

Objetivos Del Proyecto

3.1. Objetivo General

Desarrollar un modelo a través de técnicas de aprendizaje automático, con el fin de realizar un pronóstico a corto plazo de la disponibilidad de los recursos de generación de la subárea GCM.

3.2. Objetivos Específicos

Además del objetivo general, se plantean cuatro objetivos específicos:

1. Identificar las variables que inciden en la disponibilidad de los recursos de generación de la subárea GCM a través del análisis exploratorio de datos.
2. Desarrollar un análisis de correlación cruzada entre la disponibilidad de los recursos de generación con las variables predictoras.
3. Desarrollar un modelo de machine learning para la predicción de la disponibilidad de los recursos de generación de la subárea GCM.
4. Validar las predicciones de disponibilidad del modelo desarrollado a partir de validaciones cruzadas.

3.3. Resultados Esperados

Posteriormente se esperan los siguientes resultados:

1. Exploración estadística descriptiva del histórico de las variables que inciden en la disponibilidad de los recursos de generación de la planta TermoGuajira.
2. Análisis de correlación cruzada entre la disponibilidad de los recursos de generación de la planta TermoGuajira con las variables predictoras.
3. Modelo de pronóstico, basado en técnicas de aprendizaje de máquina, para la disponibilidad de la planta TermoGuajira.
4. Pronóstico de la disponibilidad de los recursos de generación de la central TermoGuajira

Capítulo 4

Marco Teórico y Antecedentes

4.1. Marco Teórico

- **Centro Nacional de despacho**
  Dependencia encarga de la planeación, supervisión y control de la operación integrada de los recursos de generación, interconexión y transmisión del sistema interconectado nacional.[8]

- **Disponibilidad**
  El tiempo total sobre un período dado, durante el cual un Activo de Uso estuvo en servicio, o disponible para el servicio. La Disponibilidad siempre estará asociada con la Capacidad Nominal del Activo, en condiciones normales de operación.[8]

- **Restricción Eléctrica**
  Limitación en el equipamiento del SIN, o de las interconexiones, tales como límites térmicos admisibles en la operación de equipos de transporte o transformación, limites en la operación del equipamiento que resulten del esquema de protecciones (locales o remotas), límites de capacidad del equipamiento o indisponibilidad de equipos.[8]

- **Restricción Operativa**
  Exigencia operativa del sistema eléctrico para garantizar la seguridad en Subáreas o Áreas Operativas, los criterios de calidad y confiabilidad, la estabilidad de tensión, la estabilidad electromecánica, los requerimientos de compensación reactiva y de regulación de frecuencia del SIN.[8]

- **Recuperación lenta inducida de tensión ante falla –FIDVR**
  En [1, pp. 112-113] el fenómeno (FIDVR por sus siglas en inglés), se refiere a un retraso en la recuperación de la tensión, luego de un evento transitorio del sistema que produzca una caída de tensión (falla, etc). El efecto de baja tensión de forma prolongada provoca altos consumos de corriente y potencia reactiva, lo que puede causar desconexión de carga por actuación de protecciones, eventos en cascada e incluso llevar a inestabilidad de voltaje.

  El estado anterior es seguido en ocasiones por una condición de sobrevoltaje en nodos de la zona de influencia (overshoot), causado por la salida de carga y la disminución de potencia reactiva demandada versus los elementos con aporte capacitivo que estaban en operación, condición que puede causar disparo por sobretensión de equipos de compensación y luego de recuperación de parte de la carga, pasar a un posible evento de sub-tensión.

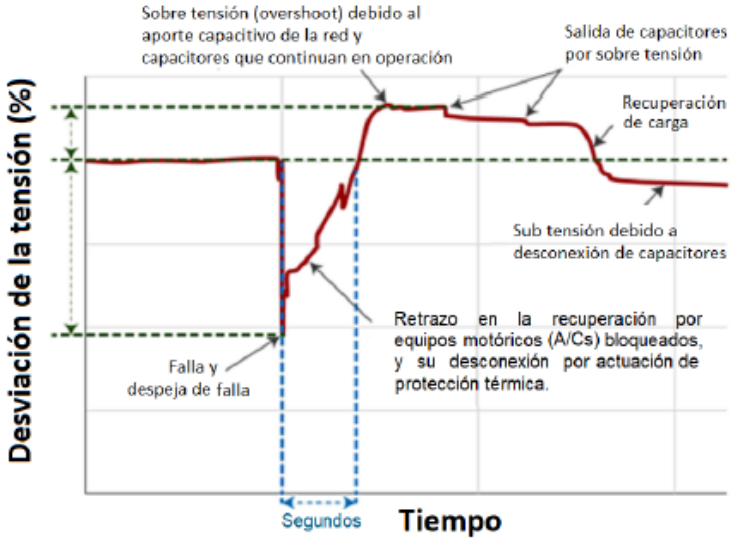

Figura 4.1: Ejemplo de fenómeno de recuperación lenta inducida de tensión ante evento. *fuente XM* [1]

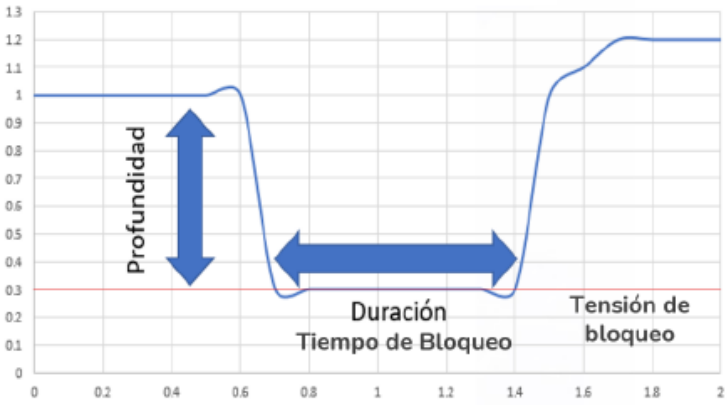

Figura 4.2: Representación de la evolución de la tensión ante un transitorio en el tiempo. *fuente XM* [1]

En un sistema eléctrico entre los principales impulsores del fenómeno (FIDVR), está el contar con alta concentración cargas motóricas impulsadas por motores de inducción monofásicos, ya que por la naturaleza de estas cargas, se caracterizan por tener un torque proporcional a la tensión en el estator, si la tensión se reduce, el flujo y el torque lo hacen de igual manera, de esta forma un motor que opere a voltaje menor al nominal tendrá cierta dificultad para llevar la carga vinculada a él como si estuviera ante un efecto de "bloqueo".

El fenómeno FIDVR se presenta cuando ante una perturbación la profundidad del hueco de tensión es inferior a la tensión de bloqueo y por un tiempo superior al tiempo de bloqueo, en este punto las cargas del tipo antes descrito se "bloquean" y consumen cantidades extremadamente altas de corriente y potencia reactiva, y continuarán haciéndolo hasta que sus protecciones lo desconecten del sistema.

- **Análisis exploratorio de los datos.**
  De acuerdo con [9],[10] el análisis exploratorio de datos se realiza con el objetivo de comprender la estructura de los datos y extraer información relevante para guiar las decisiones en etapas posteriores del proceso de análisis de datos. Algunas de las técnicas utilizadas en el análisis exploratorio incluyen la visualización de datos, el cálculo de estadísticas descriptivas, la identificación de valores atípicos, la exploración de relaciones entre variables y la búsqueda de patrones o tendencias.

  Durante el análisis exploratorio, se emplean herramientas gráficas como histogramas, diagramas de dispersión, gráficos de cajas y diagramas de barras para representar visualmente los datos y analizar su distribución, variabilidad y relaciones. Además, se pueden calcular medidas descriptivas como la media, la mediana, la desviación estándar y los cuartiles para resumir las características numéricas de los datos.

  El análisis exploratorio también puede incluir la identificación de valores atípicos o datos anómalos que se desvían significativamente de la tendencia general del conjunto de datos. Estos valores atípicos pueden ser el resultado de errores de medición, errores de entrada de datos o pueden indicar la presencia de eventos inusuales o anómalos en el fenómeno estudiado.

  Al explorar las relaciones entre variables, se pueden detectar patrones, correlaciones o dependencias que ayuden a comprender mejor los datos y proporcionen pistas sobre posibles asociaciones o influencias entre diferentes variables. Esto puede ser especialmente útil en la identificación de variables relevantes para futuros análisis o en la formulación de hipótesis iniciales.

  Es importante destacar que el análisis exploratorio de datos no es un proceso lineal y se realiza de manera iterativa a medida que se van adquiriendo nuevos conocimientos y surgen nuevas preguntas. Además, la selección de técnicas y herramientas específicas puede variar dependiendo del tipo de datos y el objetivo del análisis.

- **Modelos de Series de tiempo.**
  Los modelos de series de tiempo, según [11],[12] son herramientas estadísticas utilizadas para analizar y predecir patrones en conjuntos de datos secuenciales a lo largo del tiempo. Estos modelos consideran la dependencia temporal de los datos y capturan la estructura y características de la serie, como tendencia, estacionalidad y aleatoriedad.

  Existen varios enfoques para modelar series de tiempo, como los modelos autoregresivos (AR), los modelos de media móvil (MA), los modelos autoregresivos de media móvil (ARMA) y los

modelos autoregresivos integrados de media móvil (ARIMA). Estos modelos utilizan información pasada para realizar predicciones futuras.

Además de los modelos básicos, hay enfoques más avanzados, como los modelos de series de tiempo estacionales (SARIMA), los modelos de suavizamiento exponencial y los modelos de espacio de estado, que consideran factores adicionales para mejorar la precisión de las predicciones, como la estacionalidad y las tendencias cambiantes.

- **Modelos de regresión lineal.**
  Según [13] los modelos de regresión lineal múltiple es una técnica estadística que busca establecer una relación entre una variable dependiente continua y dos o más variables independientes. Se basa en la suposición de que la variable dependiente puede ser explicada por una combinación lineal de las variables independientes, considerando un modelo lineal en su forma general.

  En el modelo de regresión lineal múltiple, se consideran múltiples variables independientes simultáneamente.

  La ecuación general para el modelo de regresión lineal múltiple es la siguiente:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon$$

  Donde:

  - Y representa la variable dependiente que deseamos predecir.
  - $X_1, X_2, ..., X_n$ son las variables independientes que se utilizan para predecir Y.
  - $\beta_1, \beta_2, \beta_3, ..., \beta_n$ son los coeficientes de regresión que representan las contribuciones de las variables independientes al modelo.
  - ε es el término de error, que captura las diferencias entre los valores observados y los valores predichos por el modelo.

  El objetivo del modelo de regresión lineal múltiple es estimar los valores óptimos para los coeficientes de regresión $(\beta_1, \beta_2, \beta_3, ..., \beta_n)$ utilizando técnicas de estimación como el método de mínimos cuadrados, de manera que el modelo sea capaz de predecir la variable dependiente Y con la mayor precisión posible.

- **Machine Learning.**

  El aprendizaje automático (ML) es una disciplina dentro de la inteligencia artificial (IA) y la informática que se enfoca en utilizar datos y algoritmos para imitar el proceso de aprendizaje humano y mejorar su precisión con el tiempo [7]. Algunos expertos, como [6], lo describen como un conjunto de técnicas que permiten identificar patrones en los datos de manera automática y utilizar esos patrones para predecir eventos futuros, lo cual resulta útil para la toma

de decisiones en las organizaciones.

La UC Berkeley [7] divide el sistema de aprendizaje de un algoritmo de aprendizaje automático en tres partes principales:

Un proceso de decisión: En términos generales, los algoritmos de aprendizaje automático se emplean para realizar predicciones o clasificaciones. Utilizando datos de entrada, ya sea etiquetados o no, el algoritmo genera una estimación acerca de un patrón presente en los datos.

Una función de error: La función de error tiene como propósito evaluar la precisión de las predicciones realizadas por el modelo. Si se dispone de ejemplos conocidos, la función de error permite realizar una comparación y determinar qué tan precisa es la estimación realizada por el modelo.

Un proceso de optimización del modelo: Si el modelo puede ajustarse de manera más precisa a los puntos de datos del conjunto de entrenamiento, los pesos se modifican para reducir la discrepancia entre los ejemplos conocidos y las estimaciones realizadas por el modelo. El algoritmo repite este proceso de evaluación y optimización, actualizando los pesos de manera autónoma hasta alcanzar un nivel de precisión deseado.

- **Máquinas de Soporte Vectorial**

Según [14],[15] las Máquinas de Soporte Vectorial (SVM) son algoritmos de aprendizaje automático supervisado utilizados para abordar problemas de clasificación y regresión. Fueron desarrollados por Vapnik y sus colegas en los años 90 en Bell Labs y desde entonces se han convertido en una herramienta popular en el campo de la inteligencia artificial y el análisis de datos.

La idea central detrás de SVM es encontrar un hiperplano óptimo que separe los datos en diferentes clases de manera eficiente. Este hiperplano es una superficie de separación que divide el espacio de características en dos regiones, correspondientes a cada clase. La optimización se logra maximizando la margen, que es la distancia perpendicular desde el hiperplano hasta los puntos de datos más cercanos de cada clase. Estos puntos se conocen como vectores de soporte y juegan un papel crucial en la construcción del modelo SVM.

Una de las ventajas clave de SVM es su capacidad para manejar datos linealmente separables y no linealmente separables. En el caso de datos linealmente separables, se puede encontrar un hiperplano que los divida de manera precisa. Sin embargo, cuando los datos no son linealmente separables, SVM utiliza trucos de kernel para mapear los datos a un espacio de características

de mayor dimensión donde sí sean linealmente separables. Los kernels permiten capturar relaciones no lineales entre las variables y amplían la capacidad de SVM para resolver problemas más complejos.

Además de la clasificación binaria, SVM también se utiliza para la clasificación multiclase y la regresión. En la clasificación multiclase, se emplean diferentes estrategias como üno contra todosz üno contra uno"para extender SVM a más de dos clases. En la regresión, se ajusta una función lineal o no lineal a los datos para predecir valores continuos en función de las variables de entrada.

SVM tiene varias ventajas, entre las que se incluyen su capacidad para manejar conjuntos de datos de alta dimensionalidad, su resistencia al sobreajuste y su capacidad para generalizar bien a datos nuevos. Sin embargo, también tiene algunas limitaciones. Por ejemplo, SVM es sensible a la selección de parámetros, como el tipo de kernel y su configuración, lo que requiere un ajuste cuidadoso para obtener buenos resultados. Además, el entrenamiento de SVM puede ser computacionalmente costoso, especialmente cuando se trabaja con grandes conjuntos de datos. A pesar de estas limitaciones, SVM sigue siendo ampliamente utilizado en diversas aplicaciones, como reconocimiento de imágenes, procesamiento de texto, bioinformática y más.

- **Las redes neuronales.**

De acuerdo con [16], las redes Neuronales Simples, también conocidas como perceptrones, son un tipo de modelo de aprendizaje automático supervisado que se inspira en la estructura y función de las neuronas del cerebro humano. Consisten en una red interconectada de nodos o neuronas, donde cada nodo representa una unidad de procesamiento que recibe una entrada y produce una salida.

En una Red Neuronal Simple, cada conexión entre los nodos tiene un peso asociado que determina la influencia de la entrada en la salida del nodo. La salida de cada nodo se calcula utilizando una función de activación, la cual puede ser lineal o no lineal. El proceso de entrenamiento de una Red Neuronal Simple implica ajustar los pesos de las conexiones para minimizar una función de costo, que mide la discrepancia entre las salidas predichas por la red y las salidas reales. Esto se logra utilizando algoritmos de aprendizaje, como el descenso del gradiente.

La arquitectura de una Red Neuronal Simple consta de una capa de entrada, una o más capas ocultas y una capa de salida. La capa de entrada recibe los datos de entrada, mientras que la capa de salida produce la salida final de la red. Las capas ocultas proporcionan a la red la capacidad de aprender representaciones más complejas y abstractas de los datos.

Las Redes Neuronales Simples se utilizan en problemas de clasificación binaria o regresión. En el caso de la clasificación binaria, la salida de la red es una única unidad con una función de activación sigmoidea que produce una probabilidad de pertenencia a una de las dos clases. En el caso de la regresión, la salida de la red es una unidad con una función de activación lineal que produce una salida continua.

Las Redes Neuronales Simples ofrecen diversas ventajas, como la capacidad de aprender patrones complejos en los datos y el manejo de características no lineales. Sin embargo, también presentan limitaciones, como la necesidad de conjuntos de datos grandes para el entrenamiento y la dificultad de interpretar los resultados.

4.2. Antecedentes

Se destacan los siguientes trabajos de investigación con relación con el tema de análisis propuesto:

- **Nuevos modelos de predicción a corto plazo de la generación eléctrica en plantas basadas en energía solar fotovoltaica.**

  En el artículo [17], se analizan y desarrollan modelos de predicción a corto plazo de la energía eléctrica en una planta solar fotovoltaica. Los modelos desarrollados utilizan técnicas de minería de datos, redes neuronales artificiales, máquinas de vectores soporte, sistema basados en lógica borrosa y algoritmos evolutivos. Por último, se presentan técnicas de evaluación de los modelos como validación cruzada y evaluación con entrenamiento, validación y test.

  Este artículo guarda relación con nuestro proyecto actual, ya que ambos comparten el objetivo común de realizar pronósticos de energía en plantas de generación utilizando técnicas de aprendizaje automático. No obstante, existen distinciones en cuanto al tipo de tecnología de las plantas (solar y térmica) y en el propósito específico del pronóstico. Un aspecto destacado de este artículo es su aporte fundamental en términos de los fundamentos proporcionados para lograr predicciones precisas y validar los modelos. Se presentan de manera sólida los fundamentos teóricos, metodológicos y una comprensión profunda del contexto en el que trabajamos, lo que resulta fundamental para lograr pronósticos precisos y validados en los recursos de generación de la subárea GCM.

- **Modelo de predicción a corto plazo de la generación eléctrica en parques eólicos, utilizando técnicas de Machine-Learning.**

  En el artículo [18], se aborda la elaboración de un modelo de predicción de la energía eléctrica en parques eólicos, basados en el uso de redes neuronales artificiales recurrentes. En la etapa inicial se realizó una etapa de pre procesamiento de las series temporales de las variables más significativas que influyen sobre la producción de la energía eólica.

Este artículo está estrechamente relacionado con nuestro proyecto actual, ya que ambos comparten el objetivo principal de realizar pronósticos de energía en plantas de generación mediante técnicas de aprendizaje automático. Sin embargo, existen diferencias clave en cuanto al tipo de tecnología de la planta (eólico y térmica) y en el propósito específico del pronóstico. Una contribución fundamental de este artículo se centra en la metodología utilizada para el tratamiento de los datos en la etapa de preprocesamiento de las series temporales. Esta etapa es de vital importancia, ya que implica la preparación y transformación de los datos brutos de la planta de generación en un formato adecuado para su posterior análisis y pronóstico. El artículo proporciona una sólida metodología que nos permitirá abordar eficientemente este desafío en nuestro proyecto.

- **Modelo del proceso de producción de energía en centrales de generación térmica considerando el perfil de funcionamiento.**

  En el artículo [19], se desarrolló un modelo del proceso de producción de energía a través del análisis del perfil de funcionamiento y datos reales. Se identificó las variables de entrada y salida que gobiernan a las plantas térmicas.

  Este artículo guarda relación con nuestro actual proyecto, ya que ambos comparten el objetivo común de desarrollar modelos de predicción de energía en plantas de generación térmicas. Si bie nuestros objetivos son similares, es importante destacar que la distinción principal radica el tipo de central térmica y el tipo de combustibles utilizados en la operación. Una contribución fundamental de este artículo se encuentra en el análisis exploratorio de datos proporcionado. Este análisis nos permitirá identificar y comprender las variables clave que inciden en la disponibilidad de energía en las plantas de generación térmicas. El artículo nos brinda una valiosa orientación sobre cómo explorar y examinar a fondo los datos disponibles, lo que nos ayudará a seleccionar las variables relevantes y descubrir patrones o relaciones significativas. Este enfoque exploratorio nos permitirá construir modelos más precisos y confiables para nuestras predicciones de energía.

- **Modelos de predicción de caudales mensuales para el sector eléctrico colombiano.**

  En el artículo [20], se muestra la aplicación de algunos modelos (redes neuronales artificiales, redes adaptativas neuro-difusas, análisis espectral singular, modelo estructural y modelo físico) de predicción de caudales de tipo matemático-estadístico, y da unos primeros pasos en la implementación de un modelo físico-matemático de la dinámica hidrológica y climática. Estos modelos fueron aplicados en varias series de caudales pertenecientes al sistema de generación de energía eléctrica del país.

  Este artículo guarda relación parcial con nuestro proyecto, ya que ambos comparten el objetivo de realizar pronósticos utilizando técnicas de aprendizaje automático. Sin embargo, es importante destacar que la distinción principal radica en el tipo de variable que queremos predecir en cada caso. Un aporte fundamental de este artículo radica en la amplia gama de

modelos de aprendizaje automático que se presentan. Cada uno de estos modelos ofrece una perspectiva única y aporta diferentes herramientas para abordar el problema de pronóstico en nuestro proyecto. Al tener acceso a una variedad de modelos de aprendizaje automático, podremos evaluar y comparar su rendimiento en relación con la variable que queremos pronosticar. Esto nos permitirá seleccionar el enfoque más adecuado y utilizarlo como base para nuestro propio modelo de pronóstico.

Capítulo 5

Desarrollo del proyecto

5.1. Fase de planificación

5.1.1. Entendimiento del negocio

La Central TermoGuajira, bajo la operación de GECELCA SAS ESP (Generadora y Comercializadora de Energía del Caribe), es una instalación de generación de energía térmica de gran importancia en la región de La Guajira, Colombia. Esta central térmica se caracteriza por aspectos clave como tecnología, flexibilidad de combustible, relevancia operativa, entre otras, que la convierten en un activo esencial en la infraestructura energética de la región.

La central opera empleando tecnología de generación térmica. Este enfoque implica la combustión controlada de combustibles fósiles, como el gas natural o los combustibles líquidos, para producir vapor de agua que, a su vez, impulsa turbinas generadoras de electricidad, ver 5.1. Esta tecnología se caracteriza por su eficiencia y versatilidad, lo que permite una generación continua y confiable de energía.

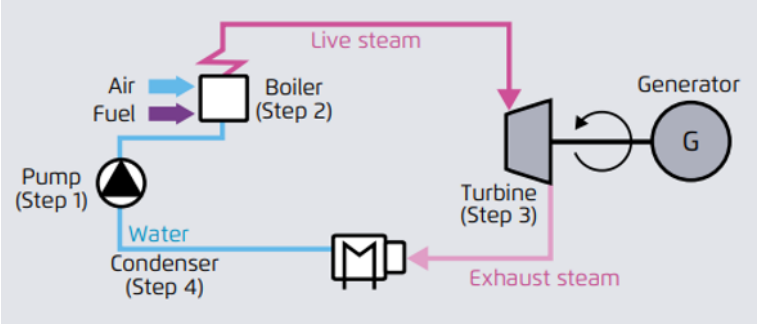

Figura 5.1: Vista esquemática de un proceso de energía a vapor de agua. *fuente agora* [2]

Un aspecto esencial de la Central TermoGuajira es su capacidad de utilizar dos tipos de combustibles para la generación de energía térmica: Gas Natural y Carbón. Esta flexibilidad le confiere la adaptabilidad necesaria para enfrentar las cambiantes condiciones del mercado de combustibles, asegurando así la disponibilidad de recursos energéticos vitales.

Los parámetros técnicos de las plantas de generación eléctrica son elementos fundamentales que

describen y cuantifican el rendimiento y la eficiencia de dichas instalaciones. Estos parámetros abarcan aspectos cruciales como capacidad nominal, cantidad máxima de electricidad que una planta puede generar en condiciones óptimas, donde la Central TermoGuajira alberga dos unidades de generación, conocidas como TermoGuajira 1 y TermoGuajira 2, ambas con una capacidad de generación eléctrica de 145 megavatios (MW). Estos parámetros técnicos no solo son indicadores clave del desempeño de una planta de generación eléctrica, sino que también orientan las decisiones estratégicas para optimizar su funcionamiento y garantizar un suministro eléctrico eficiente y sostenible.

La operación diaria de las plantas térmicas está marcada por diversos escenarios que incluyen rampas de entrada, es decir, el arranque de las unidades para la generación de energía, así como la operación a diferentes capacidades, desde el mínimo técnico hasta su capacidad nominal, e incluso paradas programadas de las unidades de generación. Además, es importante tener en cuenta que también pueden ocurrir eventos no programados, como la indisponibilidad del recurso de generación, los cuales pueden afectar significativamente la operación y el rendimiento de las unidades. Estos eventos, ver 2.1, reflejan la dinámica de funcionamiento de las unidades térmicas.

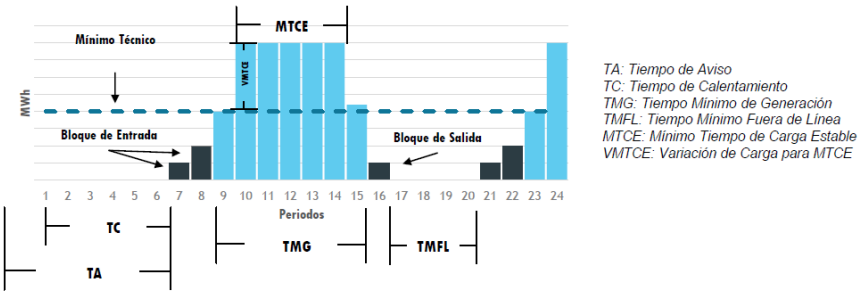

Figura 5.2: Parámetros técnicos de centrales de generación. *fuente propia*

5.1.2. Recopilación y entendimientos de los datos

La recopilación de datos relacionados con la disponibilidad de los recursos de generación se realizó mediante la extracción de información de fuentes públicas proporcionadas por el administrador del mercado XM.

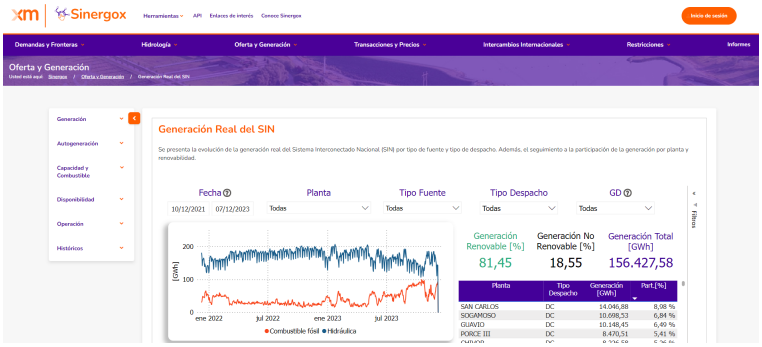

Figura 5.3: Aplicativo SINERGOX. *fuente XM [3]*

Dentro del aplicativo SINERGOX de XM, que se presenta en la sección de oferta y generación, se encuentran datos cruciales para cada recurso de generación inscrito ante el mercado eléctrica. Estos datos proporcionan una visión detallada de la operación de las unidades. A continuación, se detallan los tipos de datos disponibles:

- **Disponibilidad:** La variable de Disponibilidad es un aspecto crucial, ya que proporciona información sobre la cantidad de energía que está disponible en cada recurso de generación, medida en kilovatios-hora (kWh). Esta disponibilidad comercial es esencial para comprender la capacidad de un recurso para suministrar energía en un momento dado. No solo indica si un recurso está disponible o no, sino también cuánta energía puede generar en un período de tiempo específico.
- **Generación:** Cantidad de energía producida efectivamente por cada recurso de generación en cada hora. Este dato, también expresado en kilovatios-hora (kWh), proporciona información sobre el rendimiento real de los recursos en términos de producción de energía.
- **Combustible:** Este variable refleja la cantidad de combustible consumido por cada recurso de generación en su proceso de producción de energía. Medido en millones de unidades térmicas británicas (MBTU), este indicador es vital para evaluar la eficiencia operativa asociados con el consumo de combustible.

Dentro del aplicativo PARATEC de XM, se presentan información de parámetros técnicos de los recursos de generación que conforman el sistema interconectado naciona - SIN. Estos datos proporcionan una visión detallada de la caracteristicas técnicas de las unidades de generación. A continuación, se detallan los tipos de datos disponibles:

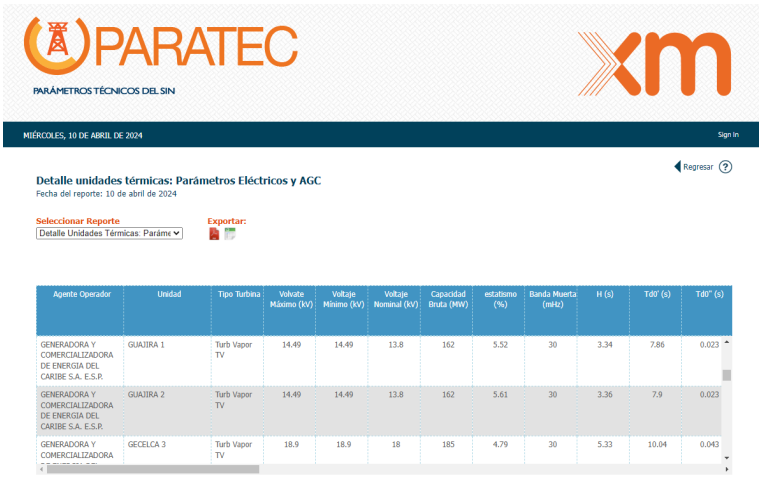

Figura 5.4: Aplicativo PARATEC. *fuente XM [4]*

- **Capacidad nominal:** Este dato representa la potencia de diseño o placa de la unidad o planta de generación. Es decir, la cantidad máxima de energía que la unidad puede producir en condiciones ideales.
- **Mínimo técnico:** Se refiere a la potencia mínima a la que puede operar la unidad o planta en condiciones normales de operación para cada configuración específica de la planta. Este parámetro es crucial para entender los límites inferiores de rendimiento de la planta y su capacidad para ajustarse a las demandas variables de energía.
- **Tipo de unidad:** Indica la clasificación de la unidad de generación, como turbina de gas tipo frame (TG), turbina de gas aeroderivada (TGA), turbina de vapor (TV), entre otros. Esta información proporciona una comprensión más detallada de la tecnología y el diseño de las unidades de generación, lo que es fundamental para la planificación y el mantenimiento adecuados.

5.2. Desarrollo del Modelo

5.2.1. Recopilación de datos

Para iniciar el proceso de recopilación de datos, se identificó la fuente pública proporcionada por el administrador de mercado XM, ver la sección de fase de planificación. Estos datos estaban disponibles en formato .csv, como se puede observar en la figura 5.5. Es importante destacar que, para la lectura del formato y el desarrollo del proyecto, se empleó la herramienta Python.

Se procedió a consolidar todos los archivos en un único archivo .csv que contenía información relacionada con la central termoguajira. El dataset compilado consta de 47,032 registros. En este conjunto de datos, el campo de fecha abarca desde el 1 de octubre de 2005. Además, el campo de

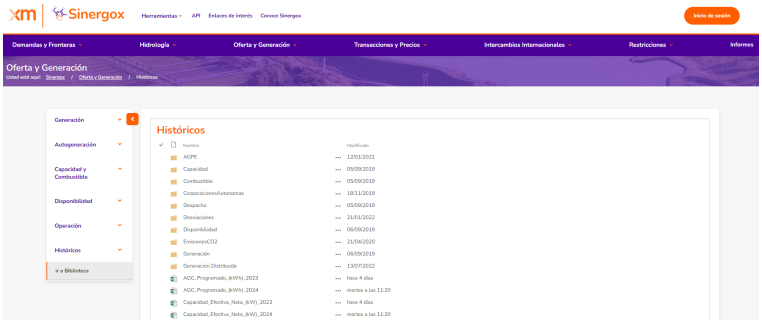

Figura 5.5: Datos históricos aplicativo SINGERGOX. *fuente XM [3]*

recurso presenta dos variables, *TGJ1* y *TGJ2*, las cuales hacen referencia a los recursos de generación de la central termoguajira.

Dentro de cada registro, se encuentran diversas variables. El campo *Código* tiene cuatro variables distintas: *DCOM*, asociado a la disponibilidad comercial; *GREA*, asociado a la generación real; *GASN*, relacionado con el combustible de gas natural utilizado para cada día de operación; y *CARB*, relacionado con el combustible de carbón utilizado para cada día de operación. Por último, las columnas *Hora01* a *Hora24* detallan los valores de cada una de las variables para cada día de operación.

A continuación, se presenta la estructura de la tabla:

Atributo	**Descripción**	**Ejemplo**
Fecha	Esta columna indica la fecha en la que se registraron los datos de disponibilidad.	1/10/2005
Recurso	Esta columna especifica el recurso del que se están registrando los datos.	TGJ1
Código	Este campo es un código identificador único para la variable a analizar, ejemplo, disponibilidad comercial.	DCOM
Hora01 - Hora24	Estas columnas representan las horas del día, desde la 1 hasta la 24, y muestran los valores de cada variable y del recurso en cada una de esas horas.	151000

Cuadro 5.1: Descripción del dataset recopilado de SINGERGOX.

Con respecto a la información recopilada de los parámetros técnicos, se presenta una descripción

detallada de los mismos para ambas unidades de la central termoguajira, donde se observa que ambas unidades de la central termoguajira presentan la misma capacidad nominal, tipo de turbina y mínimo técnico:

Recurso	Capacidad nominal	Mínimo técnico	Tipo
TGJ1	145000	72000	TV
TGJ2	145000	72000	TV

Cuadro 5.2: Parámetros técnicos general de la central termoguajira

5.2.2. Preparación de datos

Durante la etapa de preparación de datos, se llevó a cabo un proceso destinado a optimizar la calidad y la coherencia de la información para su posterior análisis y comprensión. Para ello, en primer lugar, se creó una copia del DataFrame original, una práctica para preservar la integridad de los datos originales y permitir la reversión de cambios en caso necesario.

Posteriormente, se procedió a una estandarización de los nombres de las columnas con el fin de garantizar la uniformidad y la consistencia en todo el conjunto de datos. Este proceso implicó la conversión de los nombres de las columnas a minúsculas y, específicamente para aquellas relacionadas con las horas del día, se extrajo el número de hora para su posterior uso en el análisis. De esta manera, se logró una estructura homogénea en todo el DataFrame.

A continuación, se aplicó la función melt() para pivotar la tabla, reorganizando los datos de manera que cada fila representara una combinación única de identificadores de recursos y fechas. Esta operación facilitó la manipulación y la visualización de los datos, permitiendo un análisis más eficiente y detallado.

Con el objetivo de mejorar la manipulación temporal de los datos, se convirtió la columna *fecha* al formato de fecha adecuado mediante la función pd.to datetime(). Esta conversión facilitó la realización de operaciones temporales con mayor precisión y facilidad en las etapas posteriores del análisis.

Además, se generó una nueva columna denominada *fecha periodo*, la cual integró la información de la fecha y el periodo (hora) correspondiente. Esta columna resultó de gran utilidad para análisis temporales más detallados, permitiendo identificar con precisión cada punto en el tiempo en el que se registraron los datos.

En la última fase de la preparación de datos, se realizó una revisión para verificar la consistencia y completitud de la información en todas las variables pertinentes. Se verificó que los registros de las variables *GREA*, *DCOM*, *GASN* y *CARB* comenzaran desde la misma fecha, siendo este aspecto

fundamental para asegurar la coherencia temporal en el análisis. Sin embargo, se detectó que las variables no iniciaban desde la misma fechas, por lo cual, se requirió de un ajuste para que todas las variables comenzaran desde la misma fecha, a continuación, se detalla para cada variable la fecha de inicio:

Código	**Fecha**
DCOM	2005-10-01 00:00:000
GREA	2005-10-01 00:00:000
GASN	2008-08-01 00:00:000
CARB	2008-08-01 00:00:000

Cuadro 5.3: Disponibilidad de información por variable

Para asegurar la integridad de los datos, se eliminaron las filas que contenían valores faltantes mediante el método .dropna(), garantizando que únicamente se trabajara con registros completos y fiables. Por lo cual, el set de datos para cada variable quedó finalmente con información del 01 de agosto del 2008.

En resumen, tras completar este proceso de preparación de datos, se confirmó que el dataset consta de una serie de registros completos y consistentes, lo que proporciona una base sólida y confiable para el análisis subsiguiente.

5.2.3. Análisis Exploratorio de los datos

Durante la fase de análisis exploratorio del conjunto de datos, nos adentramos en una serie de análisis gráficos con el propósito de profundizar en la identificación de patrones, detectar posibles valores atípicos y explorar las correlaciones entre las variables en estudio.

Al examinar la variable de generación real, se realizó un histograma que nos ofreció una visión reveladora: en la mayoría de las instancias, ambas unidades de generación se encontraban en reserva, lo que significa que no estaban aportando energía al sistema. Cuando sí estaban en operación, notamos que predominaba el rango de funcionamiento del mínimo técnico y máxima carga.

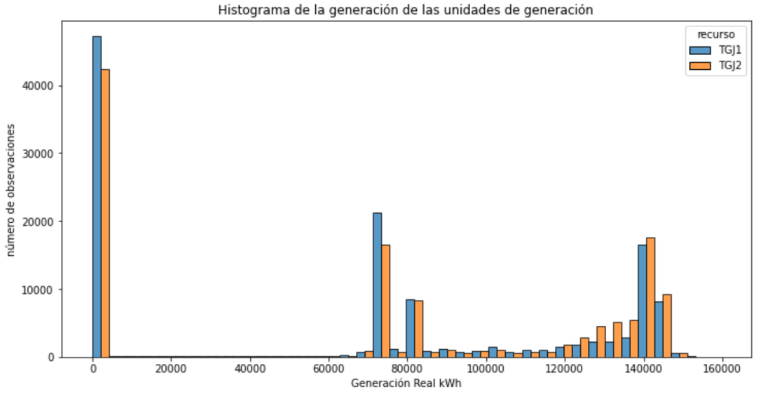

Figura 5.6: Histograma de la variable generación de energía. *fuente propia*

En cuanto a la variable de disponibilidad real, profundizamos nuestra exploración a través de otro histograma. Este análisis arrojó que, en su mayoría, ambas unidades se encontraban disponibles y operativas a su capacidad nominal. Sin embargo, llamó nuestra atención el hecho de que el estado de indisponibilidad era el siguiente más común.

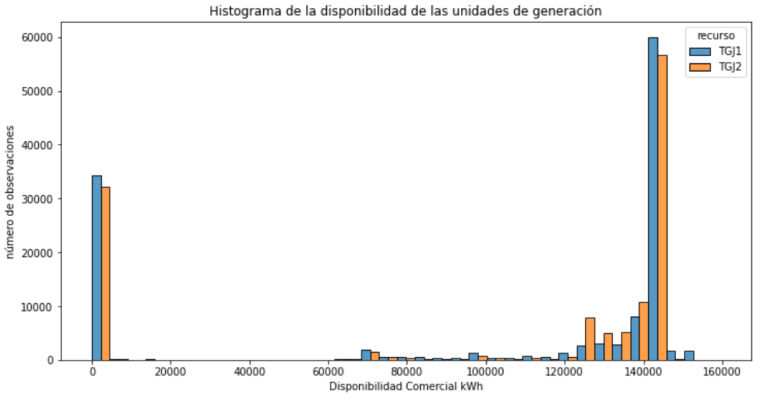

Figura 5.7: Histograma de la variable disponibilidad de energía. *fuente propia*

Para abordar el análisis de las variables de combustible, Carbón y Gas Natural, recurrimos a un

gráfico de barras apiladas. Esta representación visual nos permitió comparar de manera efectiva el uso de ambos tipos de combustible. Se destacó que el carbón era el combustible mayoritariamente utilizado, señalando posiblemente su mayor disponibilidad o eficiencia en el contexto estudiado.

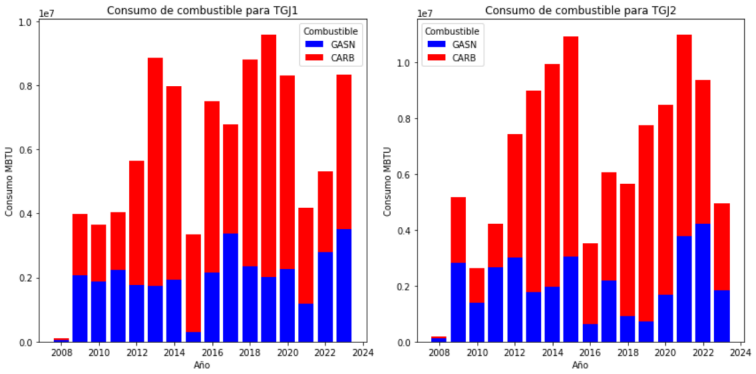

Figura 5.8: Consumo de combustible por tipo. *fuente propia*

Para completar nuestro análisis, llevamos a cabo una evaluación de correlación entre las variables predictoras (Generación Real: *GREA*, Combustible Carbón: *CARB* y Combustible Gas Natural: *GASN*) y la variable de interés principal (Disponibilidad Comercial: *DCOM*), ver 5.9. Se reveló una correlación significativa entre las variables predictoras y la disponibilidad comercial. Específicamente, encontramos que la generación real mostraba la correlación más robusta, alcanzando un coeficiente de 0.73. Este hallazgo es congruente con la lógica operativa, donde la generación de energía está intrínsecamente ligada a la disponibilidad de las unidades y al consumo de combustible.

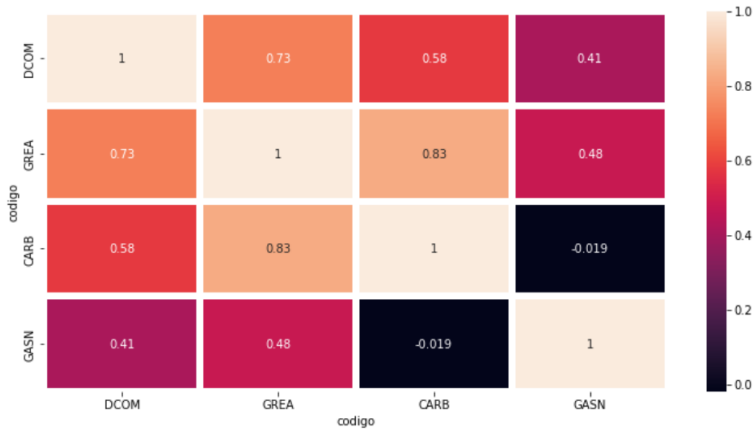

Figura 5.9: Correlación de variables. *fuente propia*

5.2.4. Selección de variables

En virtud de lo expuesto en el análisis exploratorio de los datos, más especificamente en el cálculo de las correlaciones, resulta imperativo tener en cuenta la relevancia de la independencia entre variables al construir el modelo. La alta correlación entre variables puede introducir sesgos y redundancias en el modelo, comprometiendo su capacidad predictiva y dificultando la interpretación de los resultados.

Por consiguiente, se vuelve crucial buscar un equilibrio entre la inclusión de variables relevantes y la mitigación del efecto de la multicolinealidad. Es importante considerar la interpretabilidad del modelo, ya que variables altamente correlacionadas pueden dificultar la identificación de relaciones causales y la toma de decisiones informadas.

Al mantener un conjunto de variables predictoras diverso pero no redundante, se facilita la comprensión de los factores que influyen en la disponibilidad de los recursos de generación y se promueve una mayor confianza en las predicciones del modelo. Por lo anterior, se realizó una revisión metodologíca para seleccionar las variables predictoras, priorizando aquellas que aporten información única y complementaria al modelo:

1. **Tiempo de reserva (TRE)**

   El tiempo de reserva se refiere al período durante el cual una unidad de generación está disponible pero no está generando energía. Se puede calcular utilizando la siguiente fórmula:

$$\text{TRE}[i] = \begin{cases} \text{TRE}[i-1] + 1 & \text{si DCOM}[i] \neq 0 \text{ y GREA}[i] = 0 \\ \text{TRE}[i-1] & \text{en otro caso} \end{cases}$$

2. **Tiempo de generación (TGE)**

   El tiempo de generación se refiere al período durante el cual una unidad está generando energía. Se puede calcular utilizando la siguiente fórmula:

$$\text{TGE}[i] = \begin{cases} \text{TGE}[i-1] + 1 & \text{si DCOM}[i] \neq 0 \text{ y GREA}[i] \neq 0 \\ \text{TGE}[i-1] & \text{en otro caso} \end{cases}$$

3. **Tiempo al mínimo técnico (TMT)**

   El tiempo en el que una unidad se encuentra operando a su mínimo técnico, con una franja de error del $\pm 5\,\%$. Se puede calcular utilizando la siguiente fórmula:

$$\text{TMT}[i] = \begin{cases} \text{TMT}[i-1] + 1 & \text{si DCOM}[i] \neq 0 \\ & \quad \text{si GREA}[i] \leq \text{mínimo técnico} \times 1{,}05 \\ & \quad \text{si GREA}[i] \geq \text{mínimo técnico} \times 0{,}95 \\ \text{TMT}[i-1] & \text{en otro caso} \end{cases}$$

4. **Tiempo a carga parcial (TCP)**

   El tiempo en el que una unidad se encuentra operando a una carga diferente al mínimo técnico y capacidad nominal. Se puede calcular utilizando la siguiente fórmula:

$$\text{TCP}[i] = \begin{cases} \text{TCP}[i-1] + 1 & \text{si DCOM}[i] \neq 0 \\ & \text{si GREA}[i] > \text{mínimo técnico} \times 1{,}05 \\ & \text{si GREA}[i] < \text{máxima carga} \times 0{,}95 \\ \text{TCP}[i-1] & \text{en otro caso} \end{cases}$$

5. **Tiempo a máxima carga (TMC)**

   El tiempo en el que una unidad se encuentra operando a su capacidad nominal, con una franja de error del $\pm 5\,\%$. Se puede calcular utilizando la siguiente fórmula:

$$\text{TMC}[i] = \begin{cases} \text{TMC}[i-1] + 1 & \text{si DCOM}[i] \neq 0 \\ & \text{si GREA}[i] \leq \text{máxima carga} \times 1{,}05 \\ & \text{si GREA}[i] \geq \text{máxima carga} \times 0{,}95 \\ \text{TMC}[i-1] & \text{en otro caso} \end{cases}$$

6. **Tiempo de arranque y parada (TAP)**

   El tiempo en el que una unidad se encuentra en proceso de arranque o parada programada. Se puede calcular utilizando la siguiente fórmula:

$$\text{TAP}[i] = \begin{cases} \text{TAP}[i-1] + 1 & \text{si DCOM}[i] \neq 0 \\ & \text{si GREA}[i] < \text{mínimo técnico} \times 0{,}95 \\ \text{TAP}[i-1] & \text{en otro caso} \end{cases}$$

7. **Heat Rate (HRT):**

   El Heat Rate representa el consumo promedio de combustible en relación con la generación de energía durante la operación de la unidad. Se puede calcular utilizando la siguiente fórmula:

$$\text{HRT}[i] = \begin{cases} (\text{HRT}[i-1] + \text{HRT}[i])/2 & \text{si DCOM}[i] \neq 0 \text{ y si GREA}[i] \neq 0 \\ \text{HRT}[i-1] & \text{en otro caso} \end{cases}$$

5.2.5. Metodología de construcción del dataset

En esta sección, se profundiza en la metodología empleada para el cálculo de los indicadores del dataset recopilado. Se seleccionó un fragmento ejemplar de la operación de los recursos de generación, donde se registran períodos de disponibilidad e indisponibilidad de la unidad, como se muestra a continuación:

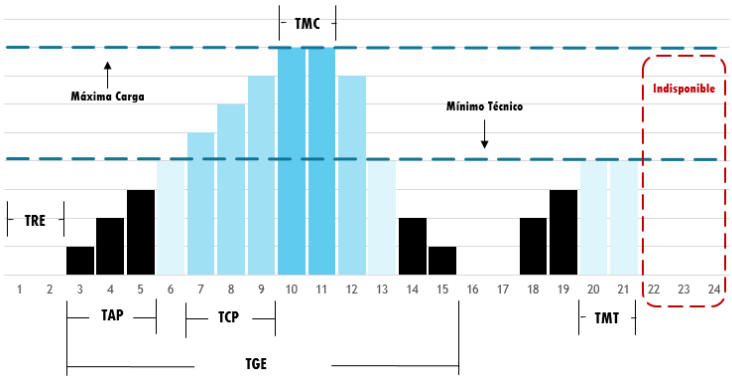

Figura 5.10: Aplicativo SINERGOX. *fuente propia*

En el diagrama presentado 5.10, se observa el proceso de arranque de la unidad durante el período 3, con un incremento gradual de la carga hasta alcanzar el mínimo técnico en el período 6 y, posteriormente, la carga máxima en el período 10. Asi mismo, se observa una disminución progresiva a partir del período 12, culminando en una parada programada en el período 15. La unidad permanece en reserva durante 2 períodos y luego arranca nuevamente, alcanzando el mínimo técnico. En el período 22, la unidad experimenta una indisponibilidad.

El modelo se diseñó con una premisa fundamental: **toda indisponibilidad surge como resultado de la operación de la unidad**. Esta premisa se fundamenta en el hecho de que las unidades térmicas están concebidas para operar a cargas estables. Sin embargo, en el contexto actual del sistema eléctrico colombiano, las plantas térmicas no son consideradas como base en el despacho para satisfacer la demanda, lo que implica que estas unidades sean sometidas a ciclos de carga, es decir, a fluctuaciones frecuentes en la carga. Estos ciclos de carga generan un estrés adicional en los recursos de generación térmica, lo que se traduce directamente en un aumento de la indisponibilidad de las unidades.

En consecuencia, en este proyecto se estableció una relación directa entre cada incidencia de indisponibilidad y su correspondiente operación. Esto significa que se identifica y registra cada evento de indisponibilidad en función de las operaciones específicas que precedieron o acompañaron a dicho evento. Esta metodología permite comprender mejor cómo las condiciones operativas influyen en la disponibilidad de las unidades térmicas.

Así mismo, durante esta etapa, se llevó a cabo una modificación crucial en el conjunto de datos: **Clasificación binaria de la variable objetivo**. se reemplazó la variable *DCOM* por la variable *disp*. Este cambio fue necesario porque nuestro objetivo no es predecir la cantidad de energía disponible, sino determinar si una unidad estará disponible o no. Es decir, se implementó una clasificación binaria para esta predicción.

La nueva variable, denominada *disp*, representa la disponibilidad de los recursos de generación en términos binarios: disponible o indisponible. Esta variable se generó a partir de la información proporcionada por *DCOM*, que originalmente indicaba la cantidad de energía disponible en kilovatio-hora (kWh) de las unidades. El uso de una clasificación binaria simplifica el modelo y permite enfocarse en predecir la ocurrencia de indisponibilidades, facilitando así la toma de decisiones para la gestión de los recursos de generación.

Implementar esta clasificación binaria es esencial para anticipar problemas de disponibilidad y desarrollar estrategias de contingencia adecuadas, mejorando la capacidad del modelo para proporcionar predicciones precisas y útiles. La variable *disp* se convierte así en la variable objetivo del modelo de machine learning, permitiendo predecir de manera efectiva si una unidad estará operativa en un momento determinado, lo cual es fundamental para la planificación y operación segura del sistema eléctrico.

Con este fundamento y teniendo en cuenta los indicadores descritos en la sección anterior, se procedió a recopilar el dataset conforme al escenario operativo mencionado:

periodo	disp	tre	tge	tmt	tcp	tmc	tap	hrt
1	1	1	0	0	0	0	0	0.00
2	1	2	0	0	0	0	0	0.00
3	1	2	1	0	0	0	1	11.50
4	1	2	2	0	0	0	2	11.45
5	1	2	3	0	0	0	3	11.44
6	1	2	4	1	0	0	3	11.43
7	1	2	5	1	1	0	3	11.30
8	1	2	6	1	2	0	3	11.25
9	1	2	7	1	3	0	3	11.20
10	1	2	8	1	3	1	3	11.00
11	1	2	9	1	3	2	3	11.00
12	1	2	10	1	4	2	3	11.20
13	1	2	11	2	4	2	3	11.30
14	1	2	12	2	4	2	4	11.33
15	1	2	13	2	4	2	5	11.35
16	1	3	13	2	4	2	5	11.35
17	1	4	13	2	4	2	5	11.35
18	1	4	14	2	4	2	6	11.40
19	1	4	15	2	4	2	7	11.39
20	1	4	16	3	4	2	7	11.35
21	1	4	17	4	4	2	7	11.35
22	0	4	17	4	4	2	7	11.35
23	0	4	17	4	4	2	7	11.35
24	0	4	17	4	4	2	7	11.35

Cuadro 5.4: Dataset inicial: Recopilación del cálculo de los indicadores

La Tabla 5.4 ilustra un ejemplo representativo de los datos recopilados, destacando los siguientes aspectos:

- Los datos correspondientes a cada punto de operación se acumulan, con el propósito de registrar el desgaste progresivo de los elementos térmicos como resultado de la actividad operativa. Esta acumulación permite capturar de manera precisa el impacto acumulativo de la operación en el estado de los recursos de generación térmica.
- Durante los períodos de indisponibilidad, se conservan los datos acumulados de la operación previa. Esta estrategia se implementa para respaldar la hipótesis de que la indisponibilidad está intrínsecamente ligada a la operación de la unidad. Mantener estos datos acumulados durante los períodos de indisponibilidad proporciona una pista crucial para que el modelo pueda asociar claramente la indisponibilidad con la actividad operativa previa.
- Cuando la unidad se encuentra nuevamente disponible, es decir, al finalizar el período de

indisponibilidad, los datos se reinician a cero con el fin de acumular los nuevos registros de la nueva operación. Esto asegura que se pueda evaluar de manera efectiva el rendimiento y desgaste en cada ciclo operativo, sin la influencia de datos históricos de indisponibilidad.

Tras revisar el conjunto de datos mencionado anteriormente, se identificó la necesidad de realizar ajustes adicionales:

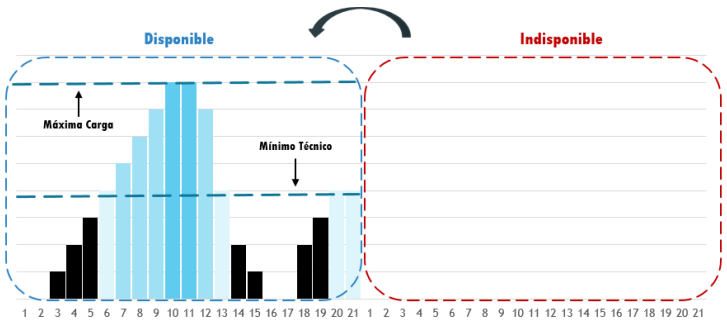

Figura 5.11: Esquemático de construcción del dataset. *fuente propia*

- La descripción de la operación de la unidad abarca 21 períodos, mientras que la indisponibilidad se registra únicamente durante 3 períodos. Esta discrepancia en la duración de las clases de operación y de indisponibilidad puede generar un desequilibrio en el conjunto de datos, ya que los datos para la indisponibilidad son idénticos a los del último período de disponibilidad. Además, si la unidad permanece en reserva, es probable que haya múltiples períodos con los mismos indicadores, lo que podría dificultar que los modelos identifiquen claramente los períodos de indisponibilidad. Por tanto, se llevó a cabo un proceso de balanceo de clases para abordar esta discrepancia.

- Para lograr un equilibrio en la clase de indisponibilidad, se replicaron los períodos de disponibilidad, como se muestra en la Figura 5.11. Este proceso garantiza que la clase de indisponibilidad capture adecuadamente el desgaste acumulado de la unidad como resultado de toda su operación. La replicación de los períodos de disponibilidad permite que la clase de indisponibilidad refleje de manera precisa el efecto acumulativo de la operación en la unidad, facilitando así la identificación y el análisis de los períodos de indisponibilidad por parte de los modelos.

- Además, se consideró que los recursos de generación TGJ1 y TGJ2 comparten parámetros técnicos idénticos. Por lo tanto, se llevó a cabo una integración de los conjuntos de datos de ambas unidades con el propósito de enriquecer y fortalecer la información disponible para el entrenamiento del modelo. Esta acción no solo aumenta la cantidad de datos disponibles, sino que también mejora la representatividad del conjunto de entrenamiento al incorporar muestras de múltiples fuentes con características técnicas similares.

Por lo anterior, se presenta el datasate final recopilado para este ejemplo:

periodo	disp	tre	tge	tmt	tcp	tmc	tap	hrt
1	1	1	0	0	0	0	0	0.00
2	1	2	0	0	0	0	0	0.00
3	1	2	1	0	0	0	1	11.50
4	1	2	2	0	0	0	2	11.45
5	1	2	3	0	0	0	3	11.44
6	1	2	4	1	0	0	3	11.43
7	1	2	5	1	1	0	3	11.30
8	1	2	6	1	2	0	3	11.25
9	1	2	7	1	3	0	3	11.20
10	1	2	8	1	3	1	3	11.00
11	1	2	9	1	3	2	3	11.00
12	1	2	10	1	4	2	3	11.20
13	1	2	11	2	4	2	3	11.30
14	1	2	12	2	4	2	4	11.33
15	1	2	13	2	4	2	5	11.35
16	1	3	13	2	4	2	5	11.35
17	1	4	13	2	4	2	5	11.35
18	1	4	14	2	4	2	6	11.40
19	1	4	15	2	4	2	7	11.39
20	1	4	16	3	4	2	7	11.35
21	1	4	17	4	4	2	7	11.35
1	0	4	17	4	4	2	7	11.35
2	0	4	17	4	4	2	7	11.35
3	0	4	17	4	4	2	7	11.35
4	0	4	17	4	4	2	7	11.35
5	0	4	17	4	4	2	7	11.35
..	..	..	..	..	..	..	..	..
21	0	4	17	4	4	2	7	11.35

Cuadro 5.5: Dataset Final: Recopilación del cálculo de los indicadores

5.2.6. Cálculo de correlaciones

Una vez completada la recopilación del dataset final, procedimos a examinar las correlaciones entre las variables seleccionadas y la variable de interés, la disponibilidad.

Se llevó a cabo un análisis utilizando gráficos de boxplot para estudiar la distribución de los tiempos de operación en diferentes estados (disponible e indisponible):

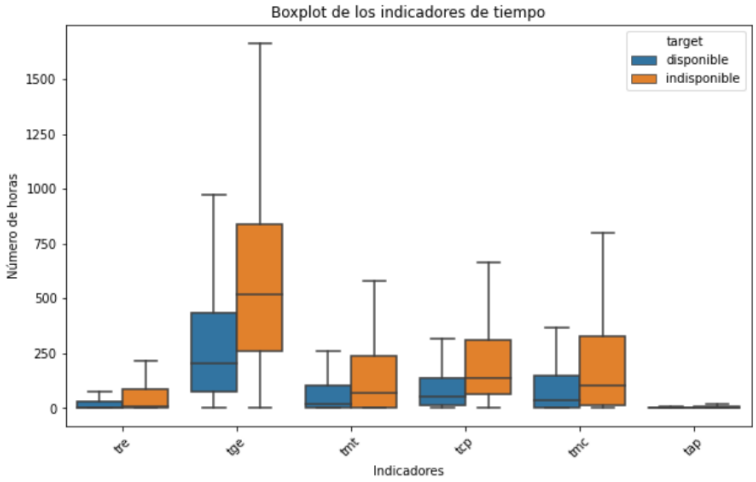

Figura 5.12: Boxplot de los indicadores de tiempo de operación. *fuente propia*

Para los estados de disponibilidad e indisponibilidad, se observó que los tiempos de generación (TGE) mostraban una mediana más alta y un rango intercuartílico más amplio, lo que indica una mayor variabilidad en los tiempos de generación.

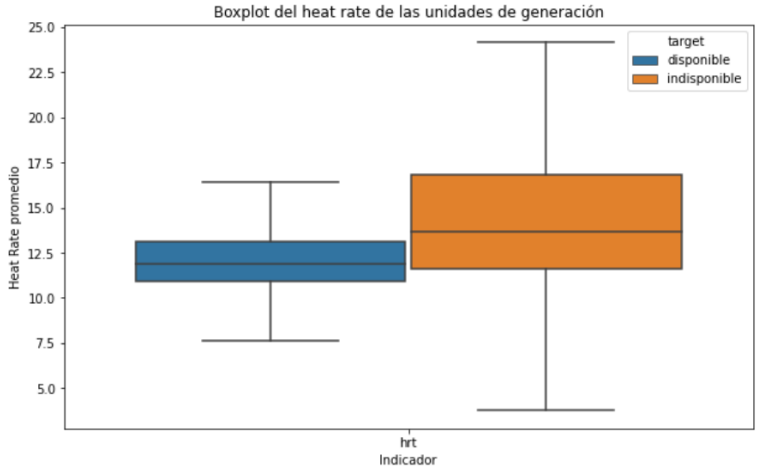

Figura 5.13: Boxplot del indicador de combustible - Heat Rate. *fuente propia*

Asi mismo, se analizaron los tiempos de carga máxima, tiempos de carga mínima y tiempos de carga parcial, observándose medianas y rangos intercuartílicos similares para ambos estados. Esto sugiere una dispersión moderada de los datos en estos aspectos. Sin embargo, es importante señalar que, durante los períodos de indisponibilidad, los valores de las variables se mantienen constantes. Esta situación puede dar lugar a valores promedio más altos en comparación con los períodos de operación disponible, lo que se refleja en boxplots que muestran alturas superiores para los períodos indisponibles. Este fenómeno resalta la diferencia en el comportamiento de las variables entre ambos estados operativos.

Además, se examinó la relación entre los combustibles y la disponibilidad mediante gráficos de boxplot para los diferentes estados de las unidades, ya sea disponibles o indisponibles. Se observó que, en general, los estados de indisponibilidad presentaban los valores más altos de *Heat Rate*, lo que indica una mayor ineficiencia de la unidad durante esos períodos:

Finalmente, se calculó y analizó las correlaciones entre las variables predictoras y la variable de interés:

Figura 5.14: Correlación de las variables. *fuente propia*

- La variable predictora *tre* tiene un coeficiente de 0,12, lo que indica una correlación postivia moderada con *disp*. Esto suiere que a medida que aumenta el valor de *tre*, también tiende a aumentar el valor de *disp*.
- La variable predictora *tge* tiene un coeficiente de 0.33, lo que sugiere una correlación positiva más fuerte con *disp*. Esto significa que *tge* tiene una asociación más estrecha y postiviia con la variable objetivo *disp* en comparación con *tre*.
- Las variables *tmt, tcp, tmc y tap* tienen coeficientes entre 0.26 y 0.2, indicando una correlación positiva más débil con *disp*. Los cambios en *tmt, tcp, tmc y tap* están menos relacionados con

disp en comparación con *tge*.

- La variable *hrt* tiene un coeficiente de -0.003, sugiriendo una correlación negativa muy débil con *disp*.

5.3. Modelamiento

5.3.1. Selección de algoritmos

En el proceso de selección de modelos, se optó inicialmente por dos algoritmos ampliamente reconocidos por su eficacia en tareas de clasificación: la regresión logística y el árbol de decisión. Estos modelos fueron escogidos debido a varias ventajas. En primer lugar, ambos son capaces de manejar tanto datos categóricos como continuos, lo que los hace versátiles en una amplia gama de aplicaciones. Además, ofrecen un alto grado de interpretabilidad, facilitando la comprensión de cómo se toman las decisiones. Finalmente, su implementación es relativamente sencilla, permitiendo una configuración y evaluación rápidas.

Tras evaluar el rendimiento inicial de estos modelos, se observó que el árbol de decisión ofrecía un desempeño prometedor. Este modelo no solo proporcionaba buenos resultados, sino que también permitía identificar de manera clara las características más relevantes para la clasificación. Debido a estos beneficios, se decidió explorar una versión más avanzada de los algoritmos basados en árboles de decisión: el modelo de Random Forest.

A continuación, se presentará un análisis detallado del proceso de implementación, evaluación y optimización de los modelos de *Logistic Regression, Decision Tree Classifier, Random Forest Classifier*, por sus siglas en ingles, destacando sus ventajas y limitaciones en comparación con los modelos iniciales. Por último, se incluió una sección de comparación de modelos para determinar el más adecuado.

5.3.2. Logistic Regression

5.3.2.1. Creación del modelo

Para la creación del modelo, se llevó a cabo varios pasos con el fin de asegurar la calidad y efectividada del proceso. A continuación, se detalla las acciones realizadas:

1. **Normalización de variables predictoras:**
   - Se seleccionaron las variables predictoras *tre, tge, tmt, tcp, tmc, tap y hrt* para el análisis.
   - Estas variables se normalizaron utilizando el métodopreprocessing.scale() de la librería-scikit learn.
   - El proceso de normalización consiste en reescalar los valores de cada variable para que tengan una media de 0 y una desviación estándar de 1.
   - Esta normalización es importante cuando se trabaja con variables que tienen diferentes escalas, ya que permite que el modelo otorgue un peso similar a cada variable durante el entrenamiento.
2. **División de datos en conjuntos de entrenamientos y prueba:**

- Se separó el conjunto de datos en dos partes: conjunto de entrenamiento (Xtrain, ytrain) y conjunto de prueba (Xtest, ytest).
- Para esta división se utilizó la función train test split() de scikit-learn, estableciendo un tamaño de prueba del 30 % (testsize=0.3).
- El conjunto de entrenamiento se utilizará para ajustar los parámetros del modelo, mientras que el conjunto de prueba servirá para evaluar el desempeño del modelo en datos que no fueron utilizados durante el entrenamiento.

3. **Entrenamiento del modelo de regresión logística:**

- Se seleccionó un modelo de Regresión Logística de la clase LogisticRegression() de scikit-learn.
- El modelo se ajustó a los datos de entrenamiento utilizando el método fit(Xtrain, ytrain).
- Durante el entrenamiento, el modelo aprende a relacionar las variables predictoras (Xtrain) con la variable objetivo *disp* (ytrain), encontrando los coeficientes y el intercepto que mejor explican esta relación.

Figura 5.15: Modelo Logistic Regression. *fuente propia*

Este proceso de normalización, división de datos y entrenamiento del modelo de regresión logística es una práctica común en el desarrollo de modelos predictivos, ver 5.15. La normalización de las variables predictoras ayuda a que el modelo asigne pesos más equilibrados a cada variable durante el entrenamiento. La división en conjuntos de entrenamiento y prueba permite evaluar el desempeño del modelo en datos nuevos, evitando el sobreajuste. Finalmente, la elección del modelo

de regresión logística se basa en que es adecuado para problemas de clasificación binaria, como en este caso predecir la variable *disp*.

5.3.2.2. Evaluación del modelo

En el proceso de evaluación del modelo, se emplearon dos herramientas fundamentales: la matriz de confusión y el informe de clasificación. Estas técnicas proporcionan una evaluación exhaustiva del desempeño del modelo, ofreciendo métricas clave como precisión, exhaustividad, puntuación F1 y soporte.

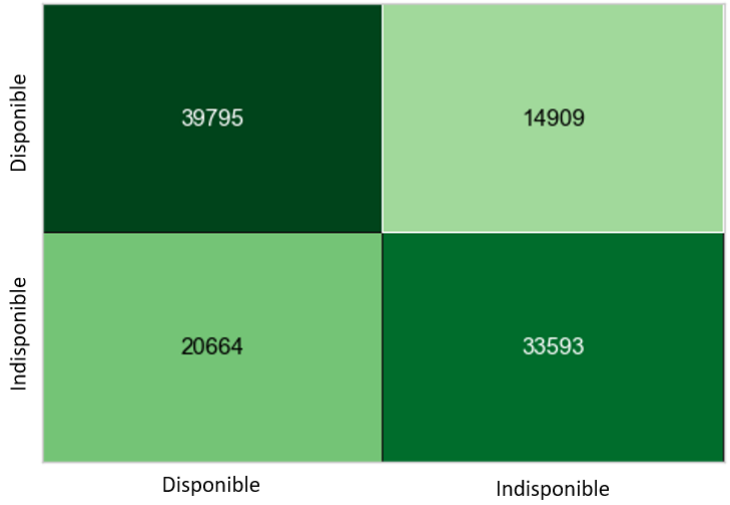

Figura 5.16: Confusion matrix del modelo Logistic Regression. *fuente propia*

Al analizar la matriz de confusión, se obtuvieron los siguientes hallazgos:

- **Clasificación de la disponibilidad:** El modelo acertó en la clasificación de 39,795 registros como disponibles, sin embargo, erróneamente clasificó 20,664 registros.
- **Clasificación de la indisponibilidad:** Se observó que el modelo logró clasificar correctamente 33,593 registros como indisponibles, pero cometió errores en la clasificación de 14,909 registros.

Por otro lado, al revisar el informe de clasificación, se identificaron los siguientes resultados detallados:

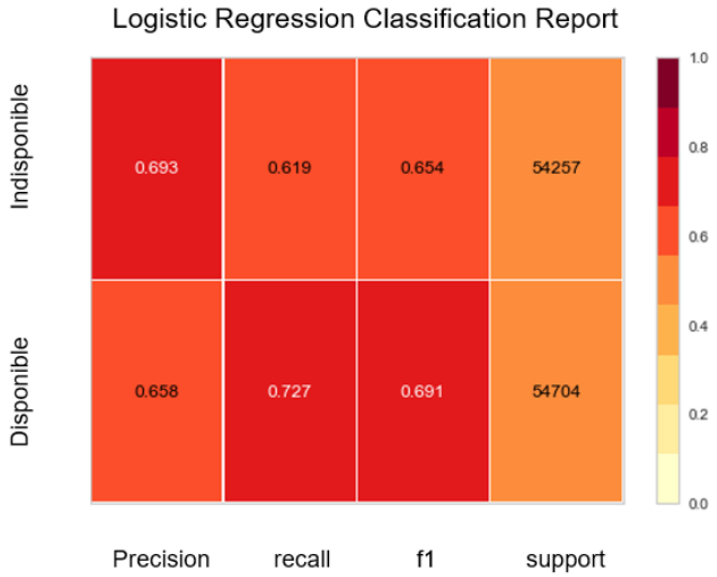

Figura 5.17: Logistic Regression Classification Report. *fuente propia*

- **Clasificación de la disponibilidad:** El modelo exhibió una precisión del 65.8 %, lo que indica que de todas las instancias clasificadas como disponibles, el 65.8 % realmente lo eran. Además, el modelo alcanzó una exhaustividad (recall) del 72.7 %, lo que significa que logró identificar el 72.7 % de todas las instancias realmente disponibles. La puntuación F1, que combina precisión y exhaustividad, fue de 0.691, y el soporte total de la clase de disponibilidad fue de 54,481 instancias.
- **Clasificación de la indisponibilidad:** En cuanto a la clasificación de la indisponibilidad, el modelo mostró una precisión del 69.3 %, lo que indica que el 69.3 % de todas las instancias clasificadas como indisponibles realmente lo eran. La exhaustividad fue del 61.9 %, lo que significa que logró identificar correctamente el 61.9 % de todas las instancias realmente indisponibles. La puntuación F1 para esta clase fue de 0.654, y el soporte total fue de 54,257 instancias.

Estos resultados proporcionan una visión holística del rendimiento del modelo de regresión logística en la predicción de la disponibilidad de recursos de generación, destacando tanto sus fortalezas como sus áreas de mejora potencial.

5.3.2.3. Resultados del modelo

La primera imagen muestra un gráfico de importancia de características, el cual brinda información sobre la importancia relativa de diferentes variables en el modelo. El eje x representa la importancia de la variable, mientras que el eje y lista las diversas características:

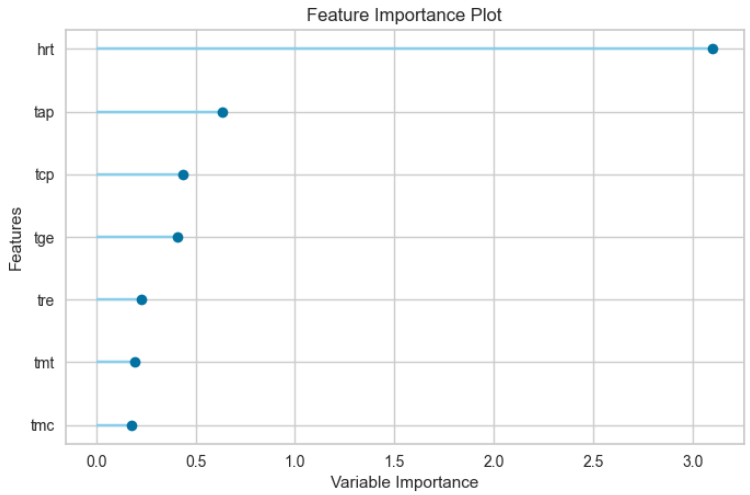

Figura 5.18: Feature Importance Plot - Logistic Regression. *fuente propia*

Según el gráfico, la característica más importante parece ser *hrt*, seguida por *tap*. Esto sugiere que estas características tienen el mayor impacto en las predicciones del modelo.

La segunda imagen muestra un gráfico de Validación Cruzada, que es una técnica utilizada para evaluar el rendimiento de un modelo de aprendizaje automático. El eje y representa la métrica que se está evaluando , accuracy, mientras que el eje x muestra los diferentes pliegues o iteraciones del proceso de validación cruzada, ver 5.19.

El gráfico indica un rendimiento relativamente estable en los diferentes pliegues, con fluctuaciones menores. Esto sugiere que el modelo es capaz de generalizar bien y de mantener un rendimiento consistente en datos no vistos, lo cual es un aspecto importante de la evaluación del modelo.

Basándonos en estos dos análisis, podemos proporcionar las siguientes conclusiones:

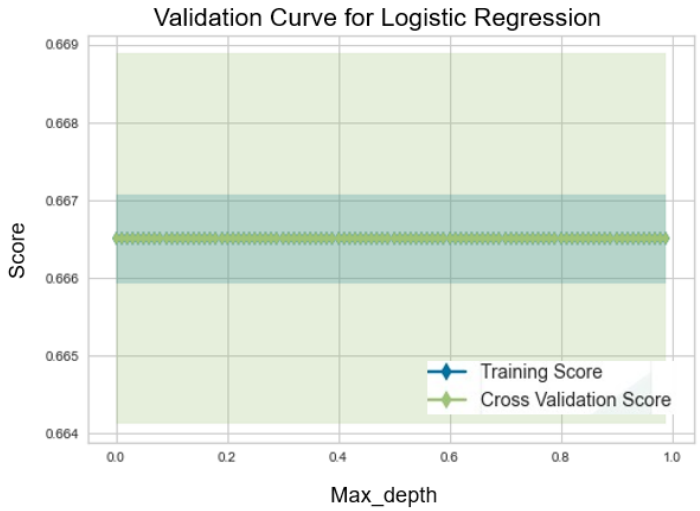

Figura 5.19: Validation Curve for Logistic Regression. *fuente propia*

1. Importancia de las Características:
   - Las características más importantes en el modelo son *hrt* y *tap*.
   - Estas características tienen el mayor impacto en las predicciones del modelo y deberían ser consideradas con mayor atención durante la selección de características y la optimización del modelo.
   - Es importante comprender el significado y la relación de estas características con la variable objetivo para garantizar que el modelo esté capturando los patrones relevantes en los datos.
2. Validación Cruzada:
   - El rendimiento estable en los pliegues de validación cruzada sugiere que el modelo es capaz de generalizar bien y de mantener un rendimiento consistente en datos no vistos.
   - Los resultados de la validación cruzada también pueden utilizarse para estimar el rendimiento esperado del modelo en datos futuros, fuera de la muestra, lo cual es crucial para evaluar la eficacia general del modelo.

En general, la combinación de análisis de importancia de características y validación cruzada proporciona una comprensión integral de las fortalezas, debilidades y áreas de optimización potencial del modelo. Esta información puede ser valiosa para refinar el modelo, seleccionar las características más relevantes y garantizar que el rendimiento del modelo sea confiable y generalizable.

5.3.3. Decision Tree Classifier

5.3.3.1. Creación del modelo

En este caso, se ha optado por implementar un modelo de Árbol de Decisión para la predicción de la variable *disp*, ver 5.20.

Arbol de decisión

```
# arbol de decisión
from sklearn.tree import DecisionTreeClassifier

# creación del modelo
tree = DecisionTreeClassifier(max_depth=3)
tree.fit(X_train,y_train)
y_train_pred = tree.predict(X_train)
y_test_pred = tree.predict(X_test)

print(metrics.accuracy_score(y_train,y_train_pred))
print(metrics.accuracy_score(y_test,y_test_pred))
```

#Pronostico#Disponibilidad

Figura 5.20: Modelo Decision TreeClassifier . *fuente propia*

1. **Importación del modelo de árbol de decisión:**
   - Se importa la clase DecisionTreeClassifier del módulo sklearn.tree.
   - Este tipo de modelo de árbol de decisión es adecuado para problemas de clasificación binaria, como es el caso de predecir la variable *disp*.
2. **Creación y entrenamiento del modelo de árbol de decisión:**
   - Se crea una instancia del modelo DecisionTreeClassifier con un parámetro max depth=3.
   - El parámetro max depth controla la complejidad del árbol de decisión, limitando la profundidad máxima del árbol a 3 niveles.
   - El modelo se entrena utilizando el método fit(Xtrain, ytrain), donde se proporcionan los datos de entrenamiento (Xtrain y ytrain).
   - Durante el entrenamiento, el algoritmo de árbol de decisión aprende a crear un modelo de clasificación que mejor se ajusta a los datos de entrenamiento.

3. **Realizar predicciones:**
    - Una vez entrenado el modelo, se utilizan los métodos predict(*Xtrain*) y predict(*Xtest*) para generar las predicciones en los conjuntos de entrenamiento y prueba, respectivamente.
    - Las predicciones para el conjunto de entrenamiento (*ytrainpred*) y el conjunto de prueba (*ytestpred*) se almacenan.

En comparación con el modelo de Regresión Logística, el uso de un Árbol de Decisión ofrece algunas ventajas:

- Capacidad de capturar relaciones no lineales entre las variables predictoras y la variable objetivo.
- Interpretabilidad del modelo, ya que los árboles de decisión son más fáciles de entender y explicar.
- Tolerancia a la presencia de variables predictoras irrelevantes o ruidosas.
- Capacidad de trabajar con variables tanto numéricas como categóricas.

5.3.3.2. Evaluación del modelo

En el proceso de evaluación del modelo, se emplearon dos herramientas fundamentales: la matriz de confusión y el informe de clasificación. Estas técnicas proporcionan una evaluación exhaustiva del desempeño del modelo, ofreciendo métricas clave como precisión, exhaustividad, puntuación F1 y soporte.

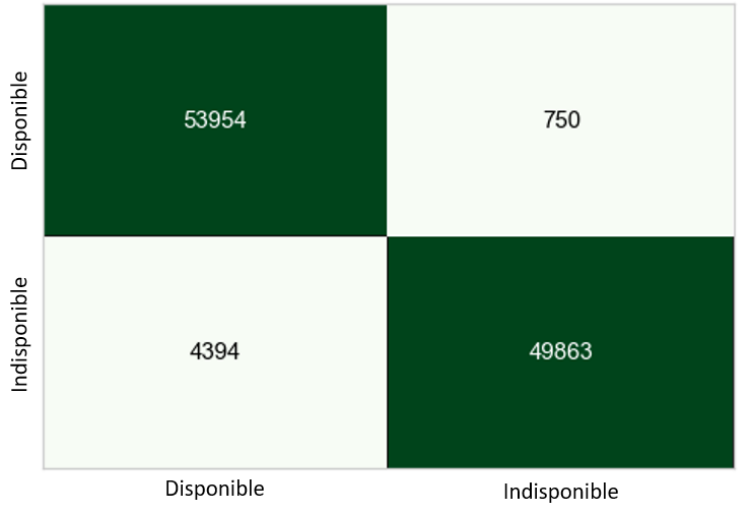

Figura 5.21: Confusion matrix - Decision TreeClassifier. *fuente propia*

Al analizar la matriz de confusión, se obtuvieron los siguientes hallazgos:

- **Clasificación de la disponibilidad:** El modelo acertó en la clasificación de 53,954 registros como disponibles, sin embargo, erróneamente clasificó 4,394 registros.
- **Clasificación de la indisponibilidad:** Se observó que el modelo logró clasificar correctamente 49,863 registros como indisponibles, pero cometió errores en la clasificación de 750 registros.

Por otro lado, al revisar el informe de clasificación, se identificaron los siguientes resultados detallados:

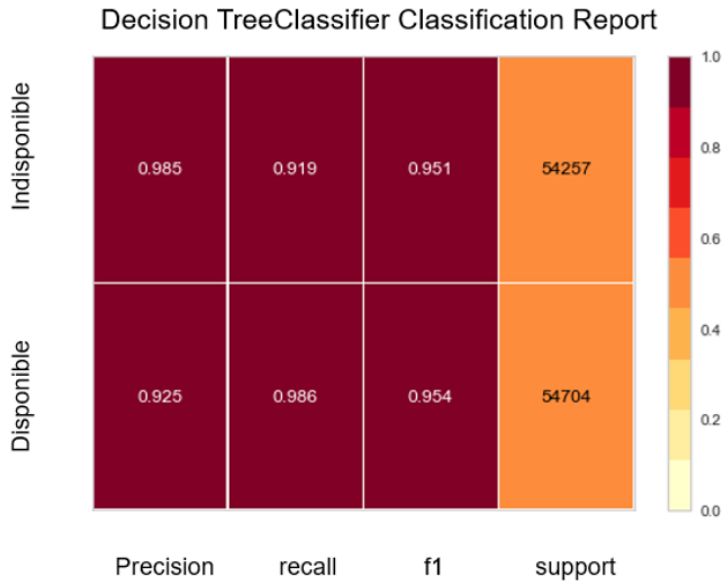

Figura 5.22: Classification Report - Decision TreeClassifier. *fuente propia*

- **Clasificación de la disponibilidad:** el modelo mostró una precisión del 92.5 %, lo que significa que del total de instancias clasificadas como disponibles, el 92.5 % realmente lo eran. Además, alcanzó una exhaustividad del 98.6 %, lo que indica que identificó correctamente el 98.6 % de todas las instancias realmente disponibles. La puntuación F1, que combina precisión y exhaustividad, fue de 0.954. El soporte total de la clase de disponibilidad fue de 54,704 instancias.
- **Clasificación de la indisponibilidad:** el modelo exhibió una precisión del 98.5 %, lo que significa que el 98.5 % de las instancias clasificadas como indisponibles realmente lo eran. La exhaustividad fue del 91.9 %, lo que indica que identificó correctamente el 91.9 % de todas las instancias realmente indisponibles. La puntuación F1 para esta clase fue de 0.951. El soporte total fue de 54,257 instancias.

A partir de la evaluación utilizando la matriz de confusión y el informe de clasificación, podemos concluir lo siguiente:

- **Rendimiento del modelo:** El árbol de decisión logró una precisión destacada en la clasificación tanto de la disponibilidad como de la indisponibilidad de recursos de generación. Esto sugiere que el modelo es capaz de distinguir entre períodos de funcionamiento normal y situaciones de fallo con una precisión considerable.
- **Precisión y exhaustividad:** Se observa que el modelo alcanza altos niveles de precisión y exhaustividad para ambas clases de disponibilidad e indisponibilidad. Esto indica que el

modelo es capaz de identificar correctamente la mayoría de los casos tanto positivos como negativos.

- **Interpretación del modelo:** Además de su rendimiento predictivo, el árbol de decisión ofrece la ventaja de ser interpretable. Esto significa que podemos comprender fácilmente cómo se toman las decisiones del modelo y qué características son más importantes para la predicción de la disponibilidad de recursos de generación.

5.3.3.3. Resultados del modelo

La primera imagen muestra un Gráfico de Importancia de Características para un modelo de árbol de decisión. Este gráfico brinda información sobre la importancia relativa de las diferentes características o variables utilizadas en el modelo.

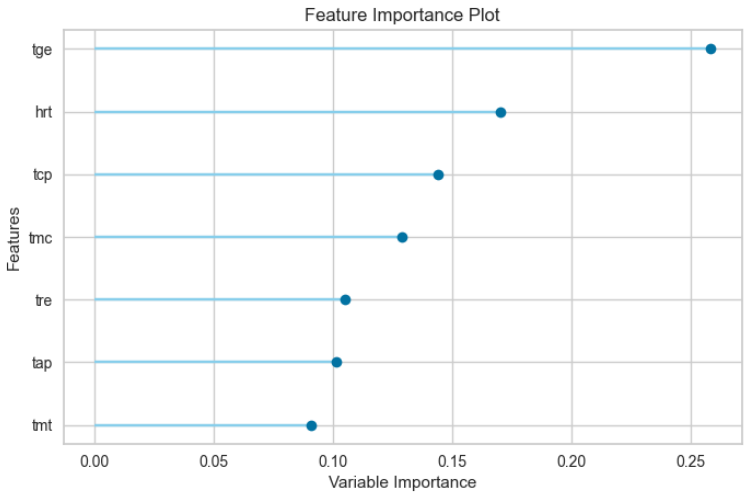

Figura 5.23: Feature importance - Decision TreeClassifier. *fuente propia*

Según el gráfico, la característica más importante parece ser *tge*, seguida de *hrt, tcp y tmc*. Esto sugiere que el modelo depende en gran medida de estas características para realizar sus predicciones, y es probable que tengan el mayor impacto en el rendimiento general del modelo.

La segunda imagen muestra un gráfico de Validación Cruzada. El eje y representa la métrica que se está evaluando, precisión.

El aumento constante en la métrica de rendimiento a lo largo del eje x, que representa los diferentes pliegues o iteraciones del proceso de validación cruzada, sugiere que el modelo puede generalizar bien y mantener un rendimiento consistente en datos no vistos. Esta es una indicación positiva de que el modelo es robusto y se puede esperar que funcione de manera confiable en aplicaciones del mundo real.

Basándonos en estos dos análisis, podemos proporcionar las siguientes conclusiones:

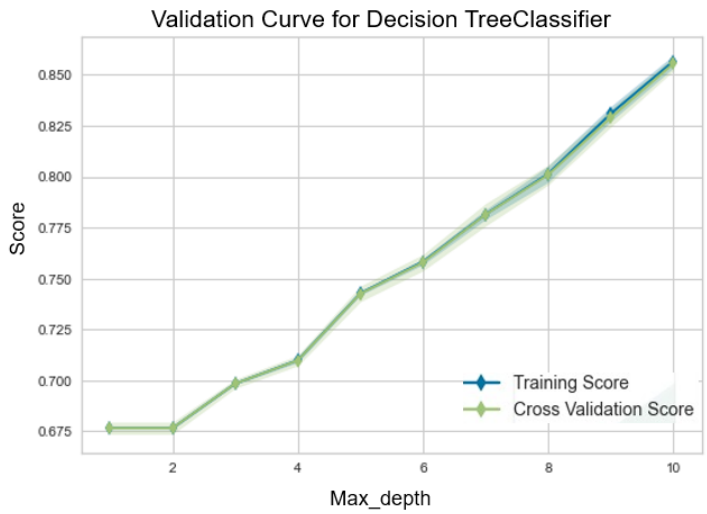

Figura 5.24: Validation Curbe for Decision TreeClassifier. *fuente propia*

1. Importancia de las Características:
   - Las características más importantes en el modelo de árbol de decisión son *tge, hrt, tcp y tmc*.
   - Estas características tienen el mayor impacto en las predicciones del modelo y deben ser consideradas con mayor atención durante la selección de características y la optimización del modelo.
   - Comprender las relaciones e interacciones entre estas características clave y la variable objetivo puede proporcionar información valiosa sobre los patrones y factores subyacentes en los datos.
2. Validación Cruzada:
   - El aumento constante en la métrica de rendimiento a lo largo de los pliegues de validación cruzada indica que el modelo puede generalizar bien y mantener un rendimiento consistente en datos no vistos.
   - Esto sugiere que es probable que el modelo sea robusto y confiable en aplicaciones del mundo real, ya que puede aplicar consistentemente los patrones aprendidos a instancias nuevas y no vistas.

En general, la combinación de la importancia de las características y los análisis de validación cruzada proporciona una comprensión integral de las fortalezas, debilidades y áreas de optimización potencial del modelo de árbol de decisión.

5.3.4. Random Forest Classifier

5.3.4.1. Creación del modelo

En este caso, se ha implementado un modelo de Random Forest para la predicción de la variable *disp*. El Random Forest es un algoritmo de aprendizaje automático basado en el ensamble de múltiples árboles de decisión.

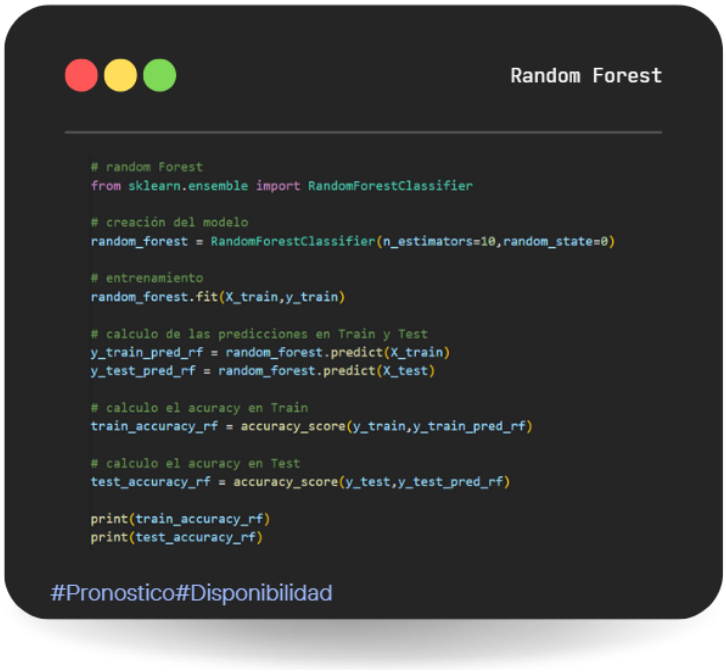

Figura 5.25: Modelo Random ForestClassifier. *fuente propia*

1. **Importación del modelo de Random Forest:**
   - Se importa la clase RandomForestClassifier del módulo sklearn.ensemble.
   - Este modelo de Random Forest es adecuado para problemas de clasificación binaria, como es el caso de predecir la variable *disp*.
2. **Creación y entrenamiento del modelo de Random Forest:**
   - Se crea una instancia del modelo RandomForestClassifier con los parametros nestimators=10: Indica que se utilizarán 10 árboles de decisión en el modelo y randomstate=0: Establece una semilla aleatoria para asegurar la reproducibilidad de los resultados.
   - El modelo se entrena utilizando el método fit(Xtrain, ytrain), donde se proporcionan los datos de entrenamiento (Xtrain y ytrain).

- Durante el entrenamiento, el algoritmo de Random Forest genera múltiples árboles de decisión utilizando diferentes subconjuntos aleatorios de las variables predictoras y de los registros de entrenamiento. Esto ayuda a crear un modelo más robusto y menos propenso al sobreajuste.

3. **Realizar predicciones:**

   - Una vez entrenado el modelo de Random Forest, se utilizan los métodos predict(Xtrain) y predict(Xtest) para generar las predicciones en los conjuntos de entrenamiento y prueba, respectivamente.
   - Las predicciones para el conjunto de entrenamiento (*ytrainpred*) y el conjunto de prueba (*ytestpred*) se almacenan.

El modelo de Random Forest ofrece varias ventajas en comparación con modelos individuales como el árbol de decisión:

- Mayor robustez y capacidad de generalización: Al combinar múltiples árboles de decisión, el modelo de Random Forest es menos susceptible al sobreajuste y puede manejar mejor la presencia de variables irrelevantes o ruidosas.
- Mejor rendimiento en general: Suele superar el desempeño de los árboles de decisión individuales, especialmente en conjuntos de datos complejos.
- Capacidad de capturar relaciones no lineales: Al igual que los árboles de decisión, el Random Forest puede modelar relaciones no lineales entre las variables predictoras y la variable objetivo.
- Importancia de las variables: El modelo de Random Forest proporciona una medida de la importancia de cada variable predictora, lo que puede ser útil para la selección de variables y la interpretación del modelo.

5.3.4.2. Evaluación del modelo

En el proceso de evaluación del modelo, se emplearon dos herramientas fundamentales: la matriz de confusión y el informe de clasificación. Estas técnicas proporcionan una evaluación exhaustiva del desempeño del modelo, ofreciendo métricas clave como precisión, exhaustividad, puntuación F1 y soporte.

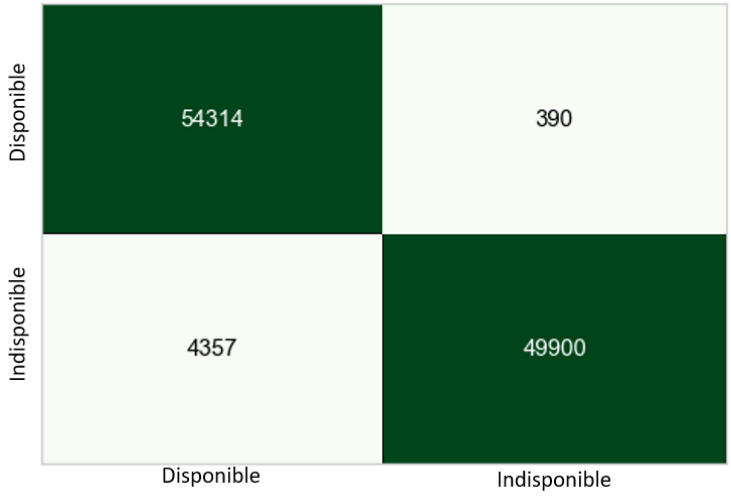

Figura 5.26: Confusion matrix - Random ForestClassifier. *fuente propia*

Al analizar la matriz de confusión, se obtuvieron los siguientes hallazgos:

- **Clasificación de la disponibilidad:** El modelo acertó en la clasificación de 54,314 registros como disponibles, sin embargo, erróneamente clasificó 4,357 registros.
- **Clasificación de la indisponibilidad:** Se observó que el modelo logró clasificar correctamente 49,900 registros como indisponibles, pero cometió errores en la clasificación de 390 registros.

Por otro lado, al revisar el informe de clasificación, se identificaron los siguientes resultados detallados:

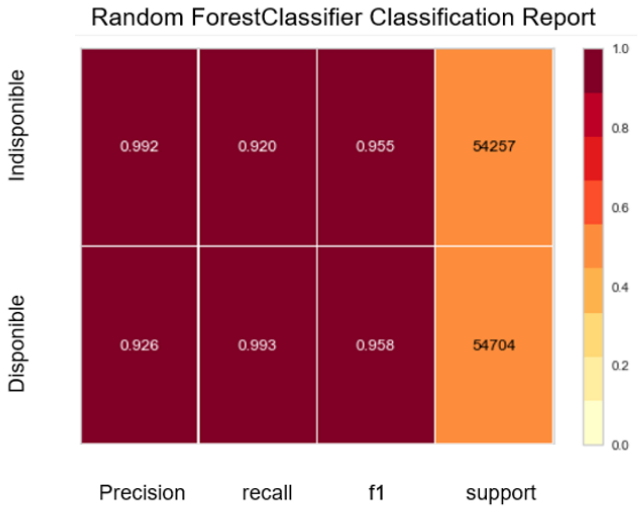

Figura 5.27: Classification Report - Random ForestClassifier. *fuente propia*

- **Clasificación de la disponibilidad:** el modelo mostró una precisión del 92.6 %, lo que significa que del total de instancias clasificadas como disponibles, el 92.6 % realmente lo eran. Además, alcanzó una exhaustividad del 99.3 %, lo que indica que identificó correctamente el 99.3 % de todas las instancias realmente disponibles. La puntuación F1, que combina precisión y exhaustividad, fue de 0.958. El soporte total de la clase de disponibilidad fue de 54,704 instancias.
- **Clasificación de la indisponibilidad:** el modelo exhibió una precisión del 99.2 %, lo que significa que el 99.2 % de las instancias clasificadas como indisponibles realmente lo eran. La exhaustividad fue del 92.0 %, lo que indica que identificó correctamente el 92.0 % de todas las instancias realmente indisponibles. La puntuación F1 para esta clase fue de 0.955. El soporte total fue de 54,257 instancias.

En conclusión, el modelo de árbol de decisión ha demostrado ser una herramienta eficaz para predecir la disponibilidad de recursos de generación. A partir de la evaluación exhaustiva utilizando la matriz de confusión y el informe de clasificación, podemos concluir lo siguiente:

- **Rendimiento del modelo:** El modelo de Random Forest exhibe un rendimiento sobresaliente en la clasificación tanto de la disponibilidad como de la indisponibilidad de recursos de generación. Esta capacidad sugiere que el modelo puede diferenciar eficazmente entre períodos de funcionamiento normal y situaciones de fallo con una precisión considerable.
- **Precisión y exhaustividad:** Los resultados revelan que el modelo logra niveles elevados tanto de precisión como de exhaustividad para ambas clases de disponibilidad e indisponibilidad.

Esto implica que el modelo es capaz de identificar correctamente la mayoría de los casos tanto positivos como negativos, lo que lo hace altamente confiable en la predicción de la disponibilidad de recursos de generación.

- **Interpretación del modelo:** Además de su rendimiento predictivo excepcional, el modelo de Random Forest ofrece la ventaja de ser interpretable. Esto significa que podemos comprender fácilmente cómo se toman las decisiones del modelo y qué características son más relevantes para la predicción de la disponibilidad de recursos de generación. La capacidad de interpretar el modelo no solo brinda información valiosa sobre el proceso de toma de decisiones del modelo, sino que también puede ayudar en la identificación de los factores clave que influyen en la disponibilidad de los recursos de generación, lo que puede ser fundamental para mejorar la eficiencia y la gestión operativa en el sector energético.

5.3.4.3. Resultados del modelo

La primera imagen muestra un Gráfico de Importancia de Características para un modelo de árbol de decisión. Este gráfico brinda información sobre la importancia relativa de las diferentes características o variables utilizadas en el modelo.

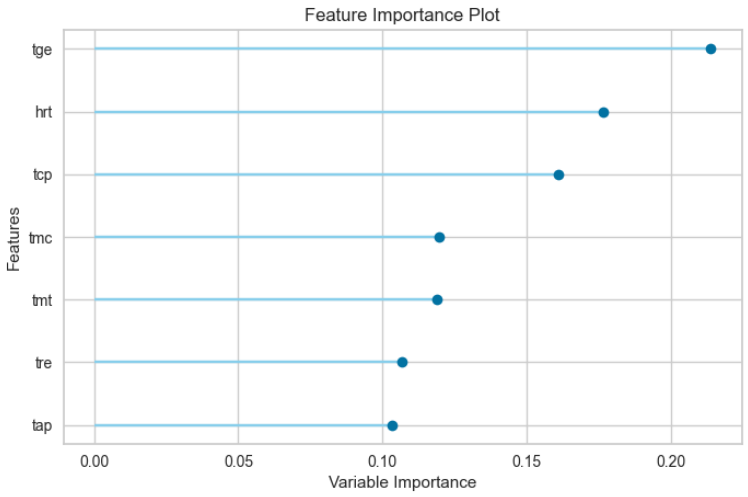

Figura 5.28: Feature importance - Random ForestClassifier. *fuente propia*

Según el gráfico, la característica más importante parece ser *tge*, seguida de *hrt, tcp, y tmc*. Esto sugiere que el modelo depende en gran medida de estas características para realizar sus predicciones, y es probable que tengan el mayor impacto en el rendimiento general del modelo.

La segunda imagen muestra un gráfico de Validación Cruzada. El eje y representa la métrica que se está evaluando, precisión.

El aumento constante en la métrica de rendimiento a lo largo del eje x, que representa los diferentes pliegues o iteraciones del proceso de validación cruzada, sugiere que el modelo puede generalizar bien y mantener un rendimiento consistente en datos no vistos. Esta es una indicación positiva de que el modelo es robusto y se puede esperar que funcione de manera confiable en aplicaciones del mundo real.

Basándonos en estos dos análisis, podemos proporcionar las siguientes conclusiones:

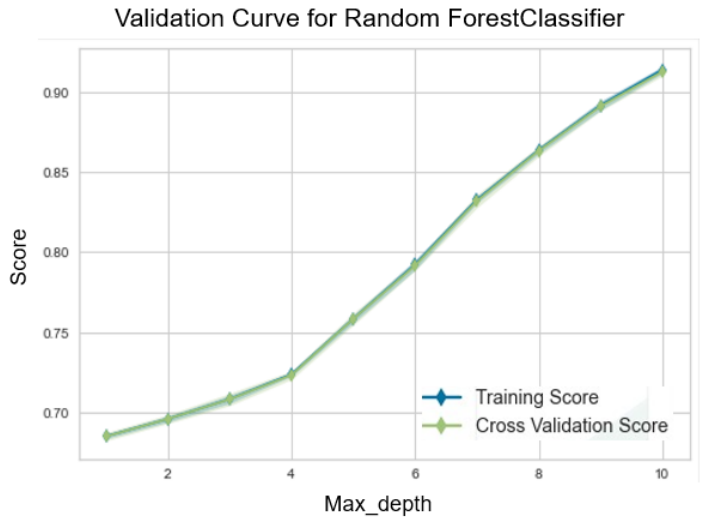

Figura 5.29: Validation Curve for Random ForestClassifier. *fuente propia*

1. Importancia de las Características:
   - Las características más importantes en el modelo de árbol de decisión son *tge, hrt, tcp, tmc*.
   - Estas características tienen el mayor impacto en las predicciones del modelo y deben ser consideradas con mayor atención durante la selección de características y la optimización del modelo.
   - Comprender las relaciones e interacciones entre estas características clave y la variable objetivo puede proporcionar información valiosa sobre los patrones y factores subyacentes en los datos.
2. Validación Cruzada:
   - El aumento constante en la métrica de rendimiento a lo largo de los pliegues de validación cruzada indica que el modelo puede generalizar bien y mantener un rendimiento consistente en datos no vistos.
   - Esto sugiere que es probable que el modelo sea robusto y confiable en aplicaciones del mundo real, ya que puede aplicar consistentemente los patrones aprendidos a instancias nuevas y no vistas.

En general, la combinación de la importancia de las características y los análisis de validación cruzada proporciona una comprensión integral de las fortalezas, debilidades y áreas de optimización potencial del modelo de árbol de decisión.

5.3.5. Comparación de modelos

En esta sección, se llevó a cabo una comparación exhaustiva de varios modelos de clasificación utilizando el conjunto de datos proporcionado. Se utilizó la biblioteca PyCaret para facilitar la creación y evaluación de los modelos.

Se realizó una configuración inicial del experimento utilizando PyCaret, donde se estableció el objetivo de la predicción (disp) y se aplicaron ajustes adicionales, como la corrección del desbalance de clases y la eliminación de valores atípicos. Se dividió el conjunto de datos en un 70 % para entrenamiento y un 30 % para pruebas.

Se compararon varios modelos de clasificación utilizando la función comparemodels() de PyCaret. Esta función evaluó cada modelo en términos de precisión (Accuracy), Área bajo la curva ROC (AUC), Recall, Precisión, F1-score, Kappa, y coeficiente de correlación de Matthews (MCC). A continuación se presentan los resultados de la comparación:

Model	Classifier	Accuracy	AUC	Recall	Prec.
et	Extra Trees Classifier	0.9943	0.9967	0.9954	0.9928
rf	Random Forest Classifier	0.9942	0.9971	0.9953	0.9927
dt	Decision Tree Classifier	0.9931	0.9955	0.9954	0.9905
knn	K Neighbors Classifier	0.9906	0.9966	0.9947	0.9860
lightgbm	Light Gradient Boosting Machine	0.9411	0.9844	0.9264	0.9513
gbc	Gradient Boosting Classifier	0.8017	0.8748	0.7802	0.8059
ada	Ada Boost Classifier	0.7313	0.8118	0.7204	0.7255
ridge	Ridge Classifier	0.6674	0.0000	0.5731	0.6905
lda	Linear Discriminant Analysis	0.6674	0.7336	0.5731	0.6905
lr	Logistic Regression	0.6665	0.7339	0.5861	0.6833
svm	SVM - Linear Kernel	0.6630	0.0000	0.5427	0.6973
nb	Naive Bayes	0.6293	0.7282	0.3952	0.7152
qda	Quadratic Discriminant Analysis	0.6205	0.7254	0.3698	0.7116
dummy	Dummy Classifier	0.5136	0.5000	0.0000	0.0000

Cuadro 5.6: Comparación de modelos

Model	Classifier	F1	Kappa	MCC	TT (Sec)
et	Extra Trees Classifier	0.9941	0.9885	0.9885	6.4530
rf	Random Forest Classifier	0.9940	0.9883	0.9884	9.3990
dt	Decision Tree Classifier	0.9929	0.9862	0.9862	1.3780
knn	K Neighbors Classifier	0.9903	0.9811	0.9811	6.5680
lightgbm	Light Gradient Boosting Machine	0.9387	0.8820	0.8823	2.3320
gbc	Gradient Boosting Classifier	0.7928	0.6027	0.6030	11.4340
ada	Ada Boost Classifier	0.7229	0.4622	0.4623	3.6630
ridge	Ridge Classifier	0.6263	0.3312	0.3359	1.5770
lda	Linear Discriminant Analysis	0.6263	0.3312	0.3359	1.3950
lr	Logistic Regression	0.6309	0.3300	0.3332	2.7580
svm	SVM - Linear Kernel	0.6101	0.3215	0.3294	1.3780
nb	Naive Bayes	0.5090	0.2491	0.2776	1.1720
qda	Quadratic Discriminant Analysis	0.4865	0.2307	0.2619	1.3730
dummy	Dummy Classifier	0.0000	0.0000	0.0000	1.3900

Cuadro 5.7: Comparación de modelos

Los resultados muestran que los modelos basados en árboles de decisión, como Extra Trees Classifier, Random Forest Classifier y Decision Tree Classifier, obtuvieron los puntajes más altos en términos de precisión y otras métricas de rendimiento. Estos modelos lograron una precisión superior al 99 % y un alto AUC, lo que indica una excelente capacidad para distinguir entre clases.

Uno de los aspectos cruciales es la adecuación y optimización de los hiperparámetros de los modelos, lo que incide directamente en su rendimiento y capacidad predictiva. Para ello, es fundamental entender y ajustar los parámetros específicos de cada algoritmo utilizado. A continuación, se detallan los hiperparámetros empleados en el entrenamiento de cada modelo:

1. **Extra Trees Classifier (ET)**

   En el caso específico del modelo Extra Trees Classifier (ET), se utilizó un total de 100 estimadores. La profundidad máxima del árbol se estableció sin restricción, lo que permitió una flexibilidad óptima para la adaptación del modelo a los datos. Durante el proceso de entrenamiento, se observó que la profundidad promedio de los árboles en el bosque fue de aproximadamente 52.46, con una profundidad máxima y mínima registrada de 65 y 42, respectivamente. Además, se aplicó un criterio de división basado en el índice de Gini, con un mínimo de 2 muestras requeridas para dividir un nodo y 1 muestra mínima en las hojas. La configuración automática de características en cada división garantizó una exploración exhaustiva de las variables disponibles, maximizando así la capacidad del modelo para capturar patrones complejos en los datos.

   - Número de estimadores (n estimators): 100

- Profundidad máxima del árbol (max depth): No hay restricción (None)
- Mínimo de muestras para dividir un nodo (min samples split): 2
- Mínimo de muestras en las hojas (min samples leaf): 1
- Máximo de características a considerar en cada división (max features): Automático (auto)
- Criterio de división (criterion): Gini
- Mínima disminución de impureza requerida para dividir un nodo (min impurity decrease): 0.0
- Fracción mínima de muestras requerida en un nodo para considerarse una hoja (min weight fraction leaf): 0.0

2. **Random Forest Classifier (RF)**

   En el caso específico del modelo Random Forest Classifier (RF), se utilizó también un total de 100 estimadores para construir el bosque de árboles. La profundidad máxima del árbol se dejó sin restricción, lo que permite que el algoritmo se adapte dinámicamente a la complejidad de los datos. Durante el entrenamiento, se observó que la profundidad promedio de los árboles en el bosque fue de aproximadamente 36.7, con una profundidad mínima registrada de 31. Al igual que en el modelo ET, se aplicó un criterio de división basado en el índice de Gini, con un mínimo de 2 muestras requeridas para dividir un nodo y 1 muestra mínima en las hojas. Además, se configuró el modelo para que automáticamente seleccione las características más relevantes en cada división, lo que contribuye a una exploración exhaustiva y eficiente del espacio de características. Este enfoque holístico garantiza que el modelo RF pueda capturar tanto la complejidad inherente como los patrones sutiles presentes en los datos de manera óptima.

   - Número de estimadores (n estimators): 100
   - Profundidad máxima del árbol (max depth): No hay restricción (None)
   - Mínimo de muestras para dividir un nodo (min samples split): 2
   - Mínimo de muestras en las hojas (min samples leaf): 1
   - Máximo de características a considerar en cada división (max features): Automático (auto)
   - Criterio de división (criterion): Gini
   - Mínima disminución de impureza requerida para dividir un nodo (min impurity decrease): 0.0
   - Fracción mínima de muestras requerida en un nodo para considerarse una hoja (min weight fraction leaf): 0.0

3. **Decision Tree Classifier (DT)**

   Para el modelo Decision Tree Classifier (DT), se configuró la profundidad máxima del árbol sin restricciones, lo que permitió una exploración exhaustiva de las posibles ramificaciones. Durante el proceso de entrenamiento, se observó que tanto la profundidad máxima como la mínima del árbol fueron de 36, lo que indica una complejidad intermedia en la estructura del árbol resultante. Los criterios de división se basaron en el índice de Gini, lo que implica una evaluación de la impureza de los nodos para determinar la mejor división en cada paso. Además, se estableció un mínimo de 2 muestras para dividir los nodos y 1 muestra en las hojas, lo que garantiza una exploración precisa del espacio de características.

   - Profundidad máxima del árbol (max depth): Sin restricción (None)
   - Mínimo de muestras para dividir un nodo (min samples split): 2
   - Mínimo de muestras en las hojas (min samples leaf): 1
   - Criterio de división (criterion): Gini
   - Mínima disminución de impureza requerida para dividir un nodo (min impurity decrease): 0.0

En conjunto, estos resultados destacan el sobresaliente desempeño de los modelos basados en árboles de decisión, subrayando su capacidad para realizar clasificaciones precisas y distintivas en una amplia variedad de escenarios. Estos modelos no solo mostraron una alta precisión, sino que también demostraron ser robustos ante diferentes conjuntos de datos y condiciones de prueba.

Los modelos de Gradient Boosting y K Neighbors también mostraron un rendimiento notable. Aunque su desempeño fue ligeramente inferior al de los modelos basados en árboles de decisión, estos algoritmos aún proporcionaron resultados competitivos y consistentes. Su capacidad para manejar grandes volúmenes de datos y complejidades en los patrones subyacentes los hace valiosos en aplicaciones donde la precisión es crucial.

En contraste, los modelos Ridge Classifier, Linear Discriminant Analysis y Logistic Regression presentaron un rendimiento inferior en comparación con los modelos mencionados anteriormente. Aunque estos modelos tienen sus fortalezas en términos de simplicidad y eficiencia computacional, su capacidad para manejar problemas de clasificación más complejos y diversos fue limitada en este estudio.

Los resultados de esta comparación indican claramente que los modelos basados en árboles de decisión son los más adecuados para este problema de clasificación en particular. Su superioridad en términos de precisión y adaptabilidad los posiciona como la mejor opción para futuras aplicaciones y estudios en este dominio. No obstante, los modelos de Gradient Boosting y K Neighbors también deben considerarse como alternativas viables, especialmente en casos donde se necesite un equilibrio entre precisión y eficiencia computacional.

Capítulo 6

Conclusiones y trabajos futuros

6.1. Conclusiones

En la investigación propuesta, fue posible identificar variables clave relacionadas con la operación de los recursos de generación de la central termoeléctrica Termoguajira, las cuales se encuentran vinculadas con la disponibilidad de dichos recursos. Estas variables incluyen los tiempos de generación, tiempos a mínimo técnico, tiempos a carga parcial, tiempos en reserva, tiempos de arranque-parada, tiempos a máxima carga y el heat rate. Esta información permitió seleccionar los mejores atributos para el entrenamiento de modelos de aprendizaje supervisado, con el objetivo de detectar los estados de disponibilidad o indisponibilidad de los recursos de generación térmica de la central Termoguajira.

Se realizó una comparación de diferentes modelos, donde se identificó que los modelos basados en árboles de decisión son los más adecuados para este problema de clasificación en particular. Los modelos que presentaron los mejores rendimientos fueron Extra Trees Classifier, Random Forest Classifier y Decision Tree Classifier, con un promedio de precisión por encima del 95 %.

A partir del conjunto de datos utilizado para este proyecto, se pudo establecer que es posible realizar predicciones de disponibilidad de las centrales de generación tomando como referencia únicamente la operación diaria de dichas unidades. Es decir, la dinámica de funcionamiento de las unidades térmicas influye directamente en la disponibilidad de estos recursos.

Con la investigación realizada, se pudo desarrollar un modelo capaz de predecir la disponibilidad de la central Termoguajira a partir de la operación diaria de los recursos de generación, con un alto porcentaje de exactitud. Esto sienta las bases para construir un modelo de predicción más robusto, que pueda aplicarse a todas las unidades de generación conectadas al sistema interconectado nacional. Esto ayudaría al área de planificación de la operación diaria a mejorar la coordinación de la operación, evitando indisponibilidades no planificadas y penalizaciones económicas.

Estos hallazgos no solo respaldan la idoneidad del modelo para pronosticar la disponibilidad de los recursos de generación en la subárea GCM, sino que también proporcionan una base sólida para la implementación de medidas preventivas y estrategias de contingencia. La capacidad del modelo para realizar predicciones precisas abre la puerta a la mejora continua de la seguridad y confiabilidad del suministro eléctrico en la mencionada subárea.

6.2. Trabajos Futuros

6.2.1. Ámbito Local

Para el ámbito local, se propone una implementación inmediata del modelo, tomando en cuenta el despacho programado publicado por el administrador del mercado antes del día de operación. Esto permitirá anticiparse a una posible indisponibilidad de la Central Termoguajira y establecer estrategias comerciales más realistas basadas en la disponibilidad proyectada de las unidades.

A mediano plazo, se sugiere incorporar este modelo dentro de los procesos internos de operación y mantenimiento de las unidades. Esto incluiría medidas preventivas basadas en la probabilidad de ocurrencia de indisponibilidades, permitiendo una mejor planificación y gestión de los recursos.

Adicionalmente, se plantea la integración de datos adicionales en el modelo, como mantenimientos programados, causas específicas de las indisponibilidades (por ejemplo, rotura de caldera, fallas en los molinos, etc.), y datos de sensores de temperatura y combustión. Esto ayudará a identificar con mayor precisión las causas de las fallas y a implementar medidas preventivas en componentes específicos de las unidades de generación, anticipando y mitigando posibles fallas.

6.2.2. Ámbito Global

En el ámbito global, se sugiere integrar las probabilidades de disponibilidad de las unidades en los modelos de despacho del sistema administrado por XM. Actualmente, el sistema de reserva se basa en la demanda pronosticada, pero la inclusión de este tipo de modelos podría aumentar la eficiencia económica para el usuario final al optimizar el uso de los recursos disponibles.

Además, se recomienda continuar con el estudio detallado de los atributos de cada planta térmica de generación dentro del Sistema Interconectado Nacional (SIN). Este análisis puede contribuir significativamente a la construcción de alertas tempranas en los modelos de despacho, mejorando la respuesta ante posibles indisponibilidades y aumentando la fiabilidad del sistema.

6.2.3. Ampliación de la Investigación

Para fortalecer el alcance de este estudio, se sugiere la realización de simulaciones y pruebas piloto que validen el modelo en diferentes escenarios operativos. También es relevante explorar la incorporación de técnicas avanzadas de análisis de datos y aprendizaje automático para mejorar la precisión de las predicciones de disponibilidad y optimizar el mantenimiento predictivo.

Por último, se podría considerar la colaboración con otros operadores de sistemas eléctricos a nivel internacional para intercambiar conocimientos y mejores prácticas. Esto no solo enriquecería

el modelo propuesto sino que también permitiría adaptarlo a diferentes contextos y realidades operativas, potenciando su aplicabilidad y efectividad en distintos entornos.

Con estos trabajos futuros, se espera no solo mejorar la gestión de la Central Termoguajira sino también aportar al desarrollo de modelos más eficientes y precisos para el sector energético del país.

Capítulo 7

Anexos

7.1. Repositorio Público

En este anexo se proporciona información detallada sobre el repositorio público creado para almacenar el material utilizado en el desarrollo del proyecto de pronóstico de la disponibilidad de los recursos de generación de la central TermoGuajira. El repositorio incluye el código fuente (notebook), la base de datos y las implementaciones detalladas.

Para acceder al repositorio público completo y visualizar el código desarrollado, incluyendo las implementaciones detalladas y el análisis realizado, puede seguir el siguiente enlace: Pronóstico Disponibilidad Termoguajira

Este enlace le permitirá ver todo el contenido del proyecto, revisar el código fuente, las visualizaciones y los resultados obtenidos. Además, encontrará comentarios y explicaciones adicionales que facilitan la comprensión de cada paso del proceso.

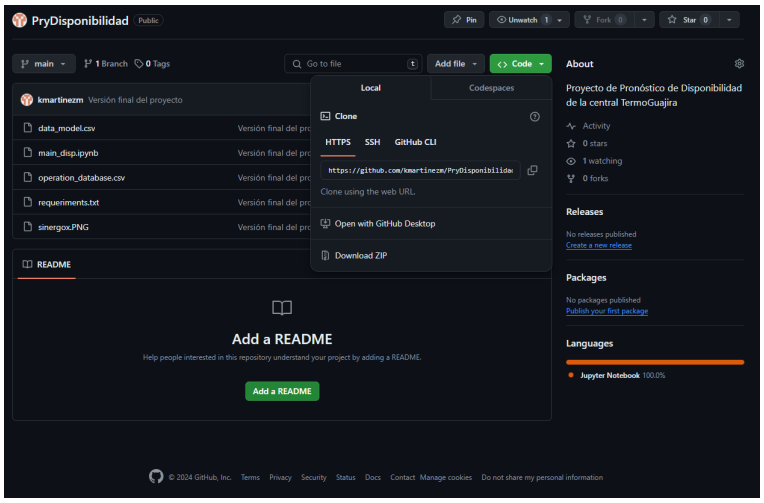

Figura 7.1: Repositorio público GitHub. *fuente propia*

Bibliografía

[1] CND, *Informe de Planeamiento Operativo Eléctrico de Mediano Plazo*, XM, mar 2023.

[2] P. Graichen, C. Redl, M. Ragwitz, E. Reiter, C. Gerbaulet, F. Kunz, J. Diekmann, and W.-P. Schill, "Flexibility in thermal power plants - with a focus on existing coal-fired power plants," Agora Energiewende, Tech. Rep., 2017. [Online]. Available: https://www.agora-energiewende.de/fileadmin/Projekte/2017/Flexibility_in_thermal_plants/115_flexibility-report-WEB.pdf

[3] XM S.A. E.S.P, "Sinergox," 2024. [Online]. Available: https://sinergox.xm.com.co/Paginas/Home.aspx

[4] XM SA E.S.P, "Paratec," 2024. [Online]. Available: https://paratec.xm.com.co/paratec/SitePages/Default.aspx

[5] C. N. DESPACHO, *Informe Trimestral de Evaluación de Restricciones*, XM, apr 2023.

[6] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[7] IBM. (2022) What is machine learning? [Online]. Available: https://www.ibm.com/topics/machine-learning

[8] C. de Regulación de Energía y Gas CREG. Glosario de términos. [Online]. Available: https://www.creg.gov.co/glosario-de-terminos-1

[9] J. B. Kadane, "Descriptive statistics: Exploratory data analysis. john w. tukey. addison-wesley, reading, mass., 1977. xvi, 688 pp., illus. $17.95." *Science*, vol. 200, no. 4338, pp. 195–195, 1978.

[10] W. S. Cleveland, *The elements of graphing data*. Wadsworth Publ. Co., 1985.

[11] P. J. Brockwell and R. A. Davis, *Introduction to time series and forecasting*. Springer, 2002.

[12] R. De Arce and R. Mahía, "Modelos arima," *Programa CITUS: Técnicas de Variables Financieras*, pp. 5–6, 2003.

[13] G. C. Canavos and E. G. U. Medal, *Probabilidad y estadística*. McGraw Hill México, 1987.

[14] A. C. Müller and S. Guido, *Introduction to machine learning with Python: a guide for data scientists*. .º'Reilly Media, Inc.", 2016.

[15] E. A. Galindo, J. A. Perdomo, and J. C. Figueroa-García, "Estudio comparativo entre máquinas de soporte vectorial multiclase, redes neuronales artificiales y sistema de inferencia neuro-difuso auto organizado para problemas de clasificación," *Información tecnológica*, vol. 31, no. 1, pp. 273–286, 2020.

[16] J. L. Sarmiento-Ramos, "Aplicaciones de las redes neuronales y el deep learning a la ingeniería biomédica," *Revista UIS Ingenierías*, vol. 19, no. 4, pp. 1–18, 2020.

[17] M. S. T. Olarte, "Nuevos modelos de predicción a corto plazo de la generación eléctrica en plantas basadas en energía solar fotovoltaica," Ph.D. dissertation, Universidad de La Rioja, 2017.

[18] D. Murti Baer, "Modelo de predicción a corto plazo de la generación eléctrica en parques eólicos, utilizando técnicas de machine-learning," 2020.

[19] S. T. Quilligana and G. A. Aviles, "Modelo del proceso de producción de energía en centrales de generación térmica considerando el perfil de funcionamiento," *Ciencia Latina Revista Científica Multidisciplinar*, vol. 6, no. 4, pp. 5541–5560, 2022.

[20] R. A. Smith, J. I. Vélez, J. D. Velásquez, A. Ceballos, P. L. Correa, C. Góez, O. O. Hernández, L. F. Salazar, and E. C. Zapata, "Modelos de predicción de caudales mensuales para el sector eléctrico colombiano," *Avances en Recursos Hidráulicos*, no. 11, pp. 91–102, 2004.

[21] XM S.A E.S.P, "xm," 2024. [Online]. Available: https://www.xm.com.co/