ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería De Ciencias de la Tierra

ANÁLISIS DE FACTORES DE CORROSIÓN EN DUCTOS DE PETRÓLEO MEDIANTE ALGORITMOS DE APRENDIZAJE AUTOMÁTICO

PROYECTO INTEGRADOR

Previo la obtención del Título de:

Ingeniero de Petróleos

Presentado por:

Bryan Anthony Coque González

Tutor de Tesis

Ing. Kenny Escobar & Ing. Xavier Vargas

GUAYAQUIL - ECUADOR

I PAO 2023

DEDICATORIA

Este proyecto se lo dedico a mi hermano y mi padre que siempre me han apoyado y alentado en el día a día, en los momentos más difíciles de esta etapa en mi vida. Así también quiero dedicárselo a mi tía por siempre estar ahí presente tanto para lo bueno como para lo malo. Y, por último, a mi madre que es lo que más quiero en este mundo.

Agradecimientos

Mi más sincero agradecimiento a cada uno de los profesores de la carrera ingeniería en petróleos por el tiempo dedicado a impartir y forjar conocimiento durante mi etapa universitaria. Por otro lado, agradecer al director de tesis Fernando Sagnay, y a mis tutores, los ingenieros Xavier Vargas y Kenny Escobar.

Declaración Expresa

“Los derechos de titularidad y explotación, me corresponde conforme al reglamento de propiedad intelectual de la institución; *Bryan Anthony Coque González* doy mi consentimiento para que la ESPOL realice la comunicación pública de la obra por cualquier medio con el fin de promover la consulta, difusión y uso público de la producción intelectual"

Bryan Anthony Coque González

Evaluadores

Firmado electrónicamente por:
XAVIER ERNESTO VARGAS GUTIERREZ

.....

Xavier Vargas

PROFESOR TUTOR

Firmado electrónicamente por:
KENNY FERNANDO ESCOBAR SEGOVIA

.....

Kenny Escobar

PROFESOR TUTOR

Firmado electrónicamente por:
FERNANDO JAVIER SAGNAY SARES

.....

Fernando Sagnay

PROFESOR DE LA MATERIA

Resumen

El presente trabajo de investigación pretende establecer los rangos más significativos de las variables que, a priori, afectan la corrosión por picaduras en ductos que transportan petróleo con respecto a la proporción de corrosión. Además, se determinan qué variables resultan más importantes y predecir mediante algoritmos de aprendizaje automático, la existencia o no de corrosión en el ducto X80. Para ello se desarrolló la metodología CRISP-DM, lo que permitió, mediante un conjunto de datos representativo del fenómeno de corrosión, realizar el análisis univariado y de correlación de las 26 variables dependientes, así como, a través de la ingeniería de características, averiguar las variables más importantes del conjunto de datos. En esta metodología se aplicó la etapa de modelado que consistió en la predicción del mejor estimador sobre la existencia o no, de corrosión por picaduras. El estudio del análisis univariado de los factores que afectan a la corrosión reveló que para un rango de valores de Co2 entre 1400-4200kpa, se tiene una mayor proporción de corrosión, con un 67% de probabilidad. Se determinó también que las variables más importantes en este estudio fueron: el azufre(S), el cobre(Cu), el dióxido de carbono(Co2) y el ácido sulfúrico (So4). El estudio proporciona información sobre las condiciones en las que la corrosión es más pronunciada y permitir identificar rangos específicos de factores de corrosión asociados a una mayor tasa de corrosión, a partir del cual se recomienda realizar un estudio detallado con la finalidad de comprender como estas variables interactúan entre sí y como sus efectos se potencian o contrarrestan mutuamente.

.

Palabras Clave: Corrosión, análisis, características, aprendizaje automático

Abstract

This research work aims to establish the most significant ranges of the variables that, a priori, affect pitting corrosion in pipelines that transport oil with respect to the proportion of corrosion. In addition, the variables that are most important are determined and predicted through automatic learning algorithms, the existence or not of corrosion in the X80 pipeline. For this, the CRISP-DM methodology was developed, which allowed, through a representative data set of the corrosion phenomenon, to carry out the univariate and correlation analysis of the 26 dependent variables, as well as, through characteristic engineering, to find out the most important variables in the data set. In this methodology, the modeling stage was applied, which consisted of the prediction of the best estimator on the existence or not of pitting corrosion. The study of the univariate analysis of the factors that affect corrosion revealed that for a range of Co2 values between 1400-4200kpa, there is a higher proportion of corrosion, with a 67% probability. It was also determined that the most important variables in this study were: sulfur (S), copper (Cu), carbon dioxide (Co2) and sulfuric acid (So4). The study provides information on the conditions in which corrosion is more pronounced and allows the identification of specific ranges of corrosion factors associated with a higher corrosion rate, from which it is recommended to carry out a detailed study in order to understand how these variables interact with each other and how their effects enhance or counteract each other.

Keywords: Corrosion, analysis, features, machine learning

Contenido

ILUSTRACIONES

TABLAS

CAPÍTULO 1

1. Introducción

Los ductos de producción son tubos de acero o polietileno que sirven para transportar cantidades industriales de petróleo y derivados de crudo refinado. Estos tubos de acero mueven alrededor de 2/3 del petróleo en todo el mundo, por lo que su uso y mantenimiento hace posible que los usuarios tengan un acceso a estos productos tan necesarios para generar energía (DXP IFS integrated flow solutions, 2019).

Hoy en día, debido que la infraestructura global opta por utilizar aceros con mayor resistencia, esto permite tener una mayor rentabilidad económica debido a que una alta resistencia de este material, provoca que se pueda obtener una mayor presión operativa, y por ende, mayor rédito económico. No obstante, existe una desventaja al utilizar este tipo de aceros más resistentes y es su vulnerabilidad frente a las corrosiones.

Los materiales metálicos se obtienen a través del procesamiento de minerales (su estado natural), lo que conduce a un estado de mayor energía. En este procesamiento, se proporcionan más electrones a los compuestos metálicos del mineral por lo que se lleva a un estado inestable a nivel termodinámico. En otras palabras, se rompen los enlaces químicos, por lo que el oxígeno, agua y aniones son removidos y hace que el metal requiera de una cantidad de energía grande para mantener su estado. Entonces a partir de aquí, se puede definir a la corrosión como la reversión química cuando un metal ya refinado, vuelve a su estado natural más estable. Esto provoca cambios en las propiedades del metal, lo que se traduce en deterioro de las funciones del metal (Javaherdashti, 2008).

Entender los factores que llevan a cabo la corrosión en oleoductos es la clave para la prevención y mitigación de los problemas de corrosión. Al analizar estos factores, las compañías pueden optar por medidas que garanticen la integridad de los oleoductos y así evitar accidentes, reducir costos operacionales debido a interrupciones de la producción, reparaciones costosas entre otros inconvenientes.

Debido a la gran cantidad de variables que se tienen en los análisis de corrosión, el aprendizaje automático es una herramienta muy útil para poder extraer información de las variables más importantes, que reflejan el estado del acero. El aprendizaje automático permite mediante algoritmos, que el sistema aprenda y mejore ciertos parámetros de forma automática sin necesidad de tenerlo programado explícitamente.

1.2 Planteamiento del problema

La existencia de corrosión en zonas específicas del oleoducto, son complejas de modelar, debido a la dependencia de múltiples factores que afectan el material metálico. Entre los factores se encuentran, parámetros ambientales, parámetros de solución y la composición del acero. Como consecuencia, a nivel macroscópico, las corrosiones suelen tener una alta aleatoriedad, por lo que coherente utilizar modelos estadísticos que permitan cuantificar los valores de corrosión (Vajo, 2002).

Existen muchas investigaciones que han sentado bases sobre el daño que puede provocar la corrosión local en la vida útil de los oleoductos, sin embargo, estos estudios se centran únicamente en un proceso o parámetro individual (Jun hu, 2014).

La ciencia de la corrosión ha implementado diferentes tecnologías para predecir los factores de corrosión a futuro, dejando de lado los modelos estadísticos.

1.3 Alcance del problema

¿Cuáles son los factores de corrosión que más afectan a la prolongación de la corrosión en ductos de petróleo SM 80SS, x52, 100 S y API X65?

Para el caso de estudio, cada muestra analizada está compuesta de 13 elementos químicos del acero (Carbono, Silicio, Manganeso, Potasio, Azufre, Cromo, Níquel, Cobre, Molibdeno, Titanio, Niobio, Aluminio y Vanadio), 4 características ambientales (H2s, Co2, Temperatura, tiempo) y 8 factores de solución (Velocidad del fluido, cloruro, bicarbonato de sódico, cátodo de magnesio, cátodo de sodio, cátodo de calcio, ánodo de sulfato).

En este caso no se analizará el diámetro de tubería y efectos de mercurio.

1.4 Justificación del problema

Debido a las ventajas que tiene el aprendizaje automático para analizar características multidimensionales, el análisis de corrosión de tuberías de petróleo pretende no sólo establecer una relación entre los factores que intervienen, sino que también enfocarse en información importante de los factores que afecten de manera considerable al estado del acero. Predicción de la corrosión y el desempeño de los modelos que se escoja para realizar el aprendizaje automático.

La aplicación de algoritmos de aprendizaje automático presenta un enfoque avanzado para afrontar los problemas por corrosión. Esta investigación contribuye a la implementación de tecnologías más sofisticadas para prevenir los problemas por corrosión en tuberías de producción de petróleo. También da lugar a una mejora las áreas con mayor riesgo para que se puedan tomar medidas preventivas adecuadas. Por último, se puede lograr la optimización de recursos debido a que este modelo calcula si existe o no corrosión a futuro.

1.5 Objetivos

1.5.1 Objetivo General

Analizar los factores que afectan la corrosión local en la tubería de producción de hidrocarburos mediante algoritmos de aprendizaje automático, estableciendo límites operativos que ayuden en la confiabilidad en los ductos.

1.5.2 Objetivos Específicos

1. Identificar los factores de mayor importancia, asegurando la capacidad de generación del modelo.
2. Predecir la existencia de corrosión en tuberías de producción del campo Espol, mediante las características del ducto X80.
3. Evaluar el desempeño del modelo, mediante medios armonizados de precisión.

1.6 Marco teórico

1.6.1 Definición de Corrosión

La corrosión se define como la reacción química de un metal con su medio adyacente y su resultado consiguiente es el deterioro de sus propiedades. Se puede entender de otra forma como la tendencia que tienen los metales a tener un estado de energía interna menor, también conocido como óxido (Barquín, 2014).

Se presenta dos reacciones necesarias para darse el fenómeno de corrosión: La reacción por oxidación o anódica que provoca la pérdida del metal y la reacción por reducción o catódica que absorbe los electrones liberados. A su vez, para el desarrollo de estas reacciones, se necesitan tres elementos: unos electrodos (ánodo y cátodo), una solución acuosa (electrolito) y una conexión eléctrica con los electrodos (ECCA, 2011)

- El ánodo corresponde a la superficie del metal expuesta a la corrosión. El metal pierde electrones y se convierte en un ion metálico positivo.
- El cátodo es el sitio donde ocurre la ganancia de electrones.

1.6.2 Tipos de corrosión

Existen varios tipos de corrosión:

1.6.2.1 Corrosión generalizada

Representa la pérdida del material en toda la superficie expuesta al ambiente. Se presenta una relación directa entre la pérdida del material, disminución del espesor y gravedad del fenómeno (Huerta, 1997). Este tipo de corrosión, es relativamente fácil de predecir y controlar ya que resulta de un mal seleccionado de un determinado material utilizado por lo que no se lo considera para el objeto de estudio.

1.6.2.2 Corrosión localizada

Se produce en zonas específicas del material por lo que representa un gran reto debido a su difícil detección. Existen mayores tasas de penetración del electrolito en la zona en cuestión mientras el resto del material permanece intacto (Aja, 2014). Los procesos de corrosión localizada que se detectan con mayor frecuencia son: por tensión, por fatiga, galvánica, en resquicio, intergranular y por picaduras. En este estudio se analiza la corrosión por picaduras más a detalle ya que, será el tipo de corrosión analizado en este trabajo.

1.6.2.3 Corrosión por picaduras

En la figura 1 se ilustra un ejemplo propagación de corrosión por picaduras. En esta imagen ya se ha formado una cavidad debido a que un anión, en este caso, el cloruro, ha roto la capa metálica pasiva, es decir, el hierro. Existen dos zonas: una próxima a la cavidad, donde se encuentran las capas pasivas. Esta zona tendrá un comportamiento catódico, ya que recibe más oxígeno, por ende, no es atacada por lo aniones de cloruro. La otra zona se encuentra

dentro de la cavidad, el cual, tendrá un comportamiento anódico, ya que se tiene menos oxígeno. El anión se coloca en esta zona ya que existe una diferencia de potencial electroquímico.

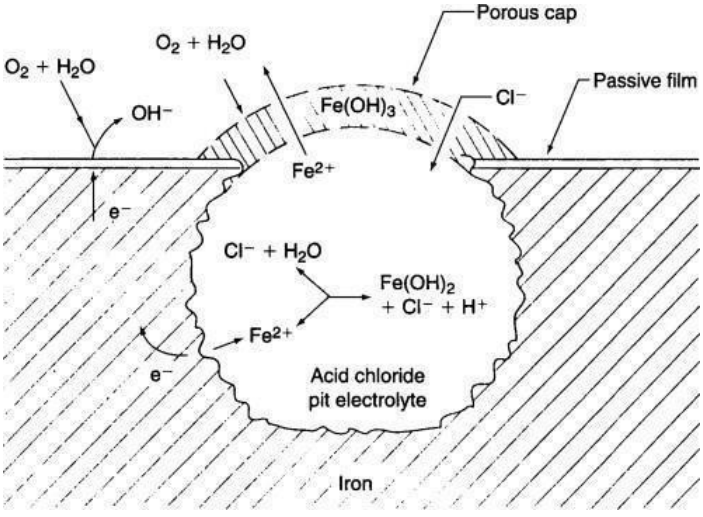
En la zona catódica cuando existe oxígeno más agua, forma aniones hidróxido (OH-). OH- entran dentro de la cavidad. Por ende, el metal se oxida y libera, cationes de metal. Lo que se transforma en una película de hidróxido, que provocan una especie de cúpula el cual aísla a la cavidad.

Dentro de la cavidad, los cationes de metal (Fe2+) tienden a rodearse de moléculas de agua. Es aquí donde se liberan protones (H+), por lo que el PH disminuye. El pH provoca que las capas pasivas rotas, no se puedan regenerar ya que existe un clima ácido.

Entonces en la cavidad, se tiene: exceso de protones (H+), cationes metálicos (Fe2+) y aniones de Cl-, Lo que agrava el efecto de corrosión de la cavidad (Ehrnstén, 2020).

Figura 1

Corrosión por Picaduras



Nota. Corrosión por picaduras en tuberías de metal*. Ehrnstén (2020).*

1.6.3 Factores que intervienen en la corrosión

En los procesos de transporte de petróleo, muchas veces conllevan cambios en la composición de los fluidos, cambios de presiones y temperaturas que hacen un entorno corrosivo para la composición del acero. A continuación, se presentan los factores de corrosión más relevantes:

Composición química del acero del acero. El carbono, el silicio, el manganeso, el azufre y el fósforo son los principales elementos de impureza en el arrabio y el acero al carbono, comúnmente conocidos como "cinco elementos". Debido a que tienen un gran impacto en el desempeño del acero, el análisis general requiere la determinación de los mismos.

Los factores ambientales, como la concentración de iones cloruro, la composición química del ambiente corrosivo, la temperatura, el pH, la presión y los agentes oxidantes, también son importantes para determinar la resistencia y seleccionar el material adecuado.

1.6.4 Aprendizaje automático

El aprendizaje automático es la capacidad que tienen las máquinas de generar modelos o algoritmos sin necesidad de programarlos de forma explícita. Lo que se logra es que los algoritmos puedan identificar patrones en los datos, y a partir de estos patrones, crear modelos para realizar predicciones. Esto es una gran ventaja para el desarrollador de un proyecto de ciencia de datos, ya que no tendrá la necesidad de programar por horas teniendo en cuenta todos los escenarios posibles (Sandoval, 2018).

Existen dos tipos de aprendizajes automáticos empleados en este trabajo: aprendizaje supervisado y aprendizaje no supervisado.

1.6.4.1 Aprendizaje Supervisado

En el aprendizaje supervisado, los algoritmos trabajan con elementos clave: características y etiquetas.

Figura 2.

Esquema de un modelo de aprendizaje automático.

Los datos de entrada son características o también denominadas en inglés como “features”. Estos datos, junto a sus respectivas etiquetas, se proporciona al sistema de aprendizaje automático o algoritmo para entrenarlo y así generar respuestas (predicciones).

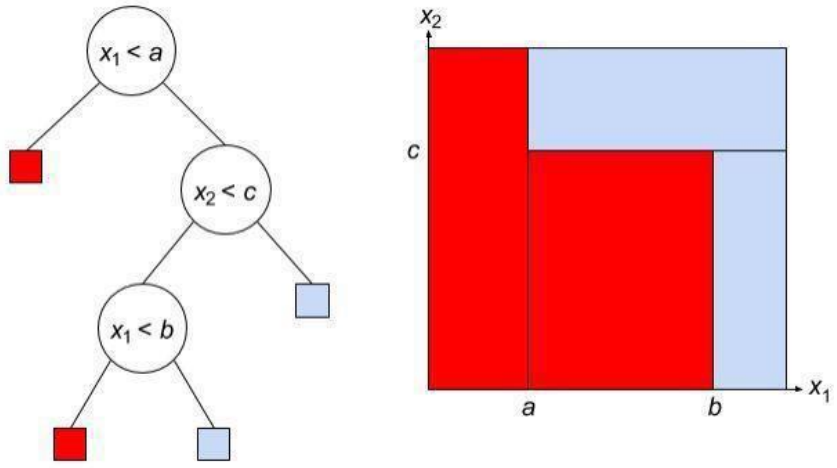
Las etiquetas pueden ser tanto cuantitativas como cualitativas. Dentro de las etiquetas cuantitativas existen número reales, matrices o vectores. En las variables categóricas se encuentran las variables normales y ordinales.

1.6.4.2 Algoritmo de bosques aleatorios

El algoritmo de bosques aleatorios es un método basado en árbol de decisión entrenados con la finalidad de disminuir la correlación entre ellos. Un árbol de decisión es un modelo de predicción basado en un conjunto de condicionales que se aplican a un conjunto de características de un dataset.

Figura 3.

Árbol de decisiones esquemático



Nota. Árbol de decisiones esquemático y llevado a un plano cartesiano con dos características*. Tomado de: Nasiriany (2019)*

Como se puede observar en la ilustración, en un árbol de decisión se reconocen:

- Conjunto de condicionales o nodos.
- Conjunto de etiquetas o clases.
- Características

Llevándolo al plano cartesiano, se grafican dos características x1 y x2. Se puede definir un espacio o frontera de decisión donde los datos que queden a un lado u otro de la frontera, pertenecen a una clase u otra. Esta idea se puede extender a n características.

En resumen, se consideran las características x1, x2, x3…xn de un elemento y se comparan mediante condicionales que llegan a definir el conjunto de etiquetas a la cual pertenece dicha característica. (Nasiriany, 2019)

Para el entrenamiento adecuado del árbol de decisión se definen los siguientes conceptos:

- Estructura del árbol: Para formar la estructura del árbol se puede limitar la profundidad de este.
- Evaluación de la calidad de divisiones mediante la minimización de la entropía o minimización factor de Gini.

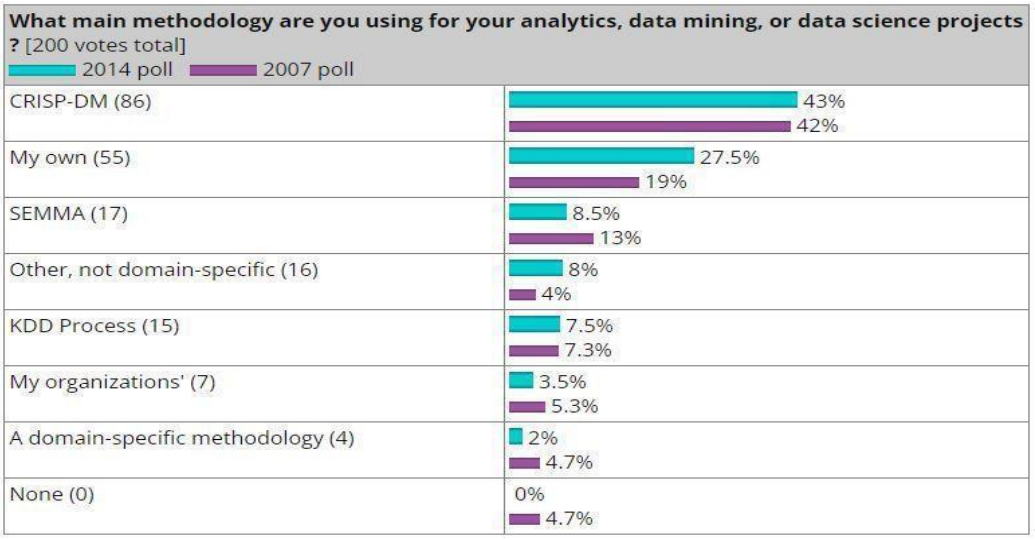
CAPITULO 2

2. METODOLOGÍA

Para este proyecto, se utilizó la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) dado que, tanto en fuentes de considerable prestigio, por ejemplo: la revista IEEE transacciones sobre conocimiento en ingeniería de datos, como diversas encuestas realizadas a profesionales de la ciencia de datos, respaldan su aplicabilidad en el proceso de descubrimiento de nuevo conocimiento en proyectos de manipulación de datos. (Haya, 2021)

Figura 4.

Metodología CRISP-DM



Nota: Encuesta sobre la metodología empleada para el desarrollo de ciencia de datos realizada a expertos del área*. Tomado de: kdnuggets (2014).*

La metodología CRISP-DM se llevó a cabo mediante una serie de procesos que permitió alcanzar una gama de objetivos.

Entendimiento del problema. Consiste en definir el problema a resolver desde una perspectiva global donde una persona con escasos conocimientos técnicos pueda entenderlo. Esta fase es muy determinante ya que asegura que el problema esté alineado con los objetivos del proyecto. Una comprensión del problema puede ayudar a identificar y mitigar problemas potenciales.

Comprensión de los datos. Se empleó cuatro tareas a realizar. A) Recolección de data para adaptarlo a las necesidades del proyecto. B) Descripción del significado de los atributos. C) Exploración de data aplicando estadística descriptiva, identificación de valores nulos o valores fuera de rango que puedan provocar ruido al modelado. (Álvarez, 2021)

Modelado. Una vez comprendido los datos, se procedió a entrenar los modelos de aprendizaje automático para predecir el mejor estimador sobre la existencia o no, de corrosión por picaduras.

2.1 Entendimiento del Problema

Se ha demostrado que, la mayoría de los incidentes en tuberías tiene que ver con el fenómeno de corrosión. Evaluando los factores que inician este fenómeno y el tipo de corrosión, se podrían determinar acciones de prevención dependiendo de los factores fundamentales que causan dicho problema. Con este análisis se consiguió evitar la progresividad de los accidentes por fuga en tuberías y por ende evitar el déficit económico en las empresas productoras de petróleo ya que evitaría la paralización de la productividad. No obstante, como se especificó en el capítulo 1, debido a la multitud de factores que afectan la corrosión dependiendo de su tipología, el aprendizaje automático puede ser una herramienta capaz de calcular la existencia o no de corrosión en el futuro, además de determinar los factores más influyentes de este fenómeno.

Un aspecto de los algoritmos de aprendizaje automático a tener en cuenta, es que, como todo software, se deben realizar suposiciones que puede provocar incertidumbre en los datos, por lo que es necesario que los modelos se evalúen mediante diferentes métricas para optimizar su desempeño y así, disminuir la incertidumbre de los resultados. (Zukhrufany, 2018). A continuación, se detalla el conjunto de datos en el que se va a realizar la metodología CRISP-DM.

Tabla 1

Esquema General del Conjunto de Datos de los Factores que Afectan la Corrosión por Picaduras en Ductos de Acero

Tipo	Variable	Descripción	Unidad
Composición química del acero	C	Carbono	% en peso
	Si	Silicio	
	Mn	Manganeso	
	P	Potasio	
	S	Azufre	
	Cr	Cromo	
	Ni	Níquel	
	Cu	Cobre	
	Mo	Molibdeno	
	Ti	Titanio	
	Nb	Niobio	
	Al	Aluminio	
	V	Vanadio	

Tabla 2.

Conjunto de Datos de tipo solución de los Factores que Afectan la Corrosión por Picaduras en Ductos de Acero.

Tipo	Variable	Descripción	Unidad
Solución	Vs	Velocidad del fluido	m/s
	Sal	Salinidad	mg/L
	Cl-	Ion cloruro	
	HCO3-	Ion Bicarbonato	
	Ca+2	Ion Calcio	
	Mg+2	Ion Magnesio	
	Na+	Ion sodio	
	SO4-2	Ion sulfato	

Tabla 3.

Conjunto de Datos de tipo ambiental de los Factores que Afectan la Corrosión por Picaduras en Ductos de Acero

Tipo	Variable	Descripción	Unidad
Ambiental	T	Temperatura	Grados C

	H2S	Sulfuro hidrógeno	Kpa
	Co2	Dióxido de carbono	Kpa
	Tiempo	Tiempo	Horas
	Pitting	Corrosión	Si/No

2.2 Comprensión de los datos

Los datos seleccionados debieron estar relacionados con las causas que conducen la corrosión por picaduras. Dado que la corrosión por picaduras es un tipo de corrosión interna, tiene que ver con la reacción entre las paredes internas de la tubería, y el producto transportado, ya sea petróleo o gas. Las principales fuentes de corrosión debido a la interacción entre el crudo y la tubería de acero, generalmente tiene que ver con la presencia de dióxido de carbono (Co2), ácido sulfhídrico (Hs2), la temperatura del fluido, la velocidad de flujo, la química del agua que contiene el crudo, y el estado de la superficie del metal.

En caso de tuberías bajo tierra, los factores que más condicionan el efecto de corrosión son: la conductividad o resistividad eléctrica, sales disueltas, el pH del suelo. Cada uno de estos factores afectan a las características de polarización anódica y catódica de un metal en el suelo.

Para este proyecto, se encontró un conjunto de 100 datos del artículo científico denominado: “Pitting Judgment Model Base On Machine Learning and Feature Optimization Methods” (Zhihao Qu et al., 2021), que mide la mayoría de los factores tanto ambientales, como superficiales y de solución. 40 de estos datos, fueron extraídos de los siguientes artículos: “Corrosion Behavior of SM 80ss Tube Steel in Simulant Solution Containing H2S and CO2” (Yin, Zhao et al., 2008); “Effects of Chloride content on Co2 Corrosion of Carbon Steel in

Simulated Oil and Gas Well Environments”; “Corrosion Behavior of 110S Tube steel in environments of high hs2 and co2 content”; “Comparison of corrosion behavior of low-alloy pipeline steel exposed to H2s/Co2 saturated brine and vapour-saturated”; “The Electrolyte renewal effect on the corrosion mechanisms of API X65 carbon steel under sweet and sour environments”.

La data restante fueron experimentos de laboratorio de simuladores de corrosión del primer artículo mencionado. A continuación, se muestra la visualización de las cinco primeras filas del conjunto de datos y el análisis estadístico descriptivo de alguna de las variables:

Figura 5.

Cinco Primeras Filas del Conjunto de Datos

	num	C	Si	Mn	P	S	Cr	Ni	Cu	Mo	...	Salinity (Sal.)	Cl-	HCO-	Ca2+	Mg2+	Na+	SO42-	pH	Time	pit
0	unit	Wt%	Wt%	Wt%	Wt%	Wt%	Wt%	Wt%	Wt%	Wt%	...	mg/L	mg/L	mg/L	mg/L	mg/L	mg/L	mg/L	\	h	NaN
1	1	0.17	0.32	1.26	0.014	0.01	0.024	0.017	0	0	...	0	0	0	0	0	0	0	3.39	2	No
2	2	0.17	0.32	1.26	0.014	0.01	0.024	0.017	0	0	...	0	0	0	0	0	0	0	3.39	6	No
3	3	0.17	0.32	1.26	0.014	0.01	0.024	0.017	0	0	...	0	0	0	0	0	0	0	3.39	18	No
4	4	0.17	0.32	1.26	0.014	0.01	0.024	0.017	0	0	...	0	0	0	0	0	0	0	3.39	24	No

Figura 6.

Estadística Descriptiva del Conjunto de Datos

	num	C	Si	Mn	P	S	Cr	Ni	Cu	Mo	...	Fluid velocity (Vs)	Salinity (Sal.)	Cl-	HCO-
count	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	...	100.000000	100.000000	100.000000	100.00000
mean	50.500000	0.184700	0.264300	1.078500	0.010480	0.007188	0.242040	0.020442	0.011200	0.140040	...	0.330000	93860.418000	57244.712000	288.53200
std	29.011492	0.075203	0.040508	0.392563	0.002607	0.004659	0.401934	0.052263	0.028378	0.243008	...	0.518448	78489.939943	48461.460788	1113.41474
min	1.000000	0.070000	0.220000	0.410000	0.006000	0.001000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.00000
25%	25.750000	0.140000	0.230000	0.690000	0.010000	0.005000	0.000000	0.000000	0.000000	0.000000	...	0.000000	28570.000000	15000.000000	0.00000
50%	50.500000	0.170000	0.270000	1.290000	0.011000	0.006000	0.012000	0.006600	0.000000	0.000000	...	0.000000	62097.000000	39425.000000	54.60000
75%	75.250000	0.260000	0.280000	1.380000	0.013000	0.010000	0.510000	0.017000	0.000000	0.290000	...	1.000000	176470.600000	107089.000000	195.00000
max	100.000000	0.260000	0.410000	1.450000	0.014000	0.015000	1.270000	0.270000	0.087000	0.720000	...	1.500000	211510.200000	129880.000000	10000.00000

Figura 7.

Estadística Descriptiva del Conjunto de Datos

Ca2+	Mg2+	Na+	SO42-	pH	Time
100.000000	100.000000	100.000000	100.000000	100.000000	100.000000
4624.920000	825.040000	30813.694000	63.520000	3.774600	130.480000
6113.797612	1253.071039	28402.198876	77.359515	0.679128	62.712504
0.000000	0.000000	0.000000	0.000000	2.970000	2.000000
0.000000	0.000000	8797.000000	0.000000	3.380000	72.000000
300.000000	200.000000	18353.150000	15.000000	3.550000	120.000000
0732.000000	1464.000000	69133.000000	100.000000	3.857500	168.000000
8200.000000	5000.000000	69381.600000	192.000000	6.000000	240.000000

Una vez descrito el significado de los atributos, se procedió a realizar la limpieza de los datos. Los datos deben ser de calidad para que exista calidad en los resultados, es por eso que, tal análisis podría ser fundamental ya que solucionaba diferentes problemas que podían cambiar el comportamiento de los modelos. Dicha limpieza consistía en tratar datos como: datos incompletos y valores ruidosos (outliers). Para el tratamiento de datos incompletos se eliminaron filas o columnas dependiendo: 1) si la mayoría de las características de cierta medición eran valores nulos, es decir, la desviación estándar de la característica era igual a 0 o 2) si la característica en particular no representaba valores para la mayoría de las mediciones. En este caso en particular, como se puede apreciar en la siguiente figura, las variables C, Si, Mn, P, S, Cr, Ni, Cu, Mo, Ti, Nb, Al, V y velocidad del fluido (Vs) tuvieron valores de desviación estándar cercanos a 0 pero en ningún caso igual a 0 por lo que se dejó el conjunto de datos, tal y como estaba.

Figura 8.

Valores de Desviación Estándar (std) de las Diferentes Variables para la Eliminación de Columnas Irrelevantes

index	num	C	Si	Mn	P	S	Cr	Ni	Cu
count	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.
mean	50.5	0.18470000000000006	0.2643000000000001	1.0784999999999996	0.010479999999999995	0.007188	0.24204	0.020442000000000005	0.011199999999999999
std	29.011491975882016	0.07520282339423597	0.04050826078615559	0.39256312623679795	0.0026072160933792495	0.0046586355135903096	0.40193350673759937	0.052262950509510984	0.02837839553839 03
min	1.0	0.07	0.22	0.41	0.006	0.001	0.0	0.0	0.
25%	25.75	0.14	0.23	0.69	0.01	0.005	0.0	0.0	0.
50%	50.5	0.17	0.27	1.29	0.011	0.006	0.012	0.0066	0.
75%	75.25	0.26	0.28	1.38	0.013	0.01	0.51	0.017	0.
max	100.0	0.26	0.41	1.45	0.014	0.015	1.27	0.27	0.08

Figura 9.

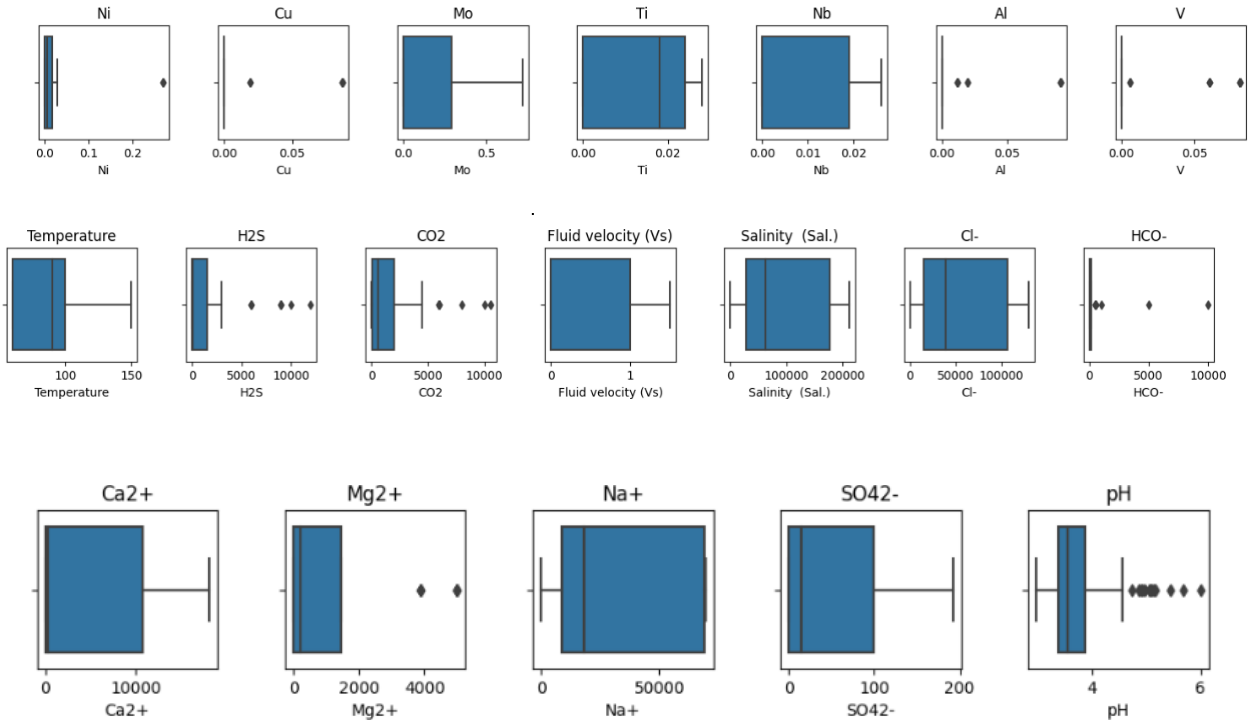
Valores de Desviación Estándar (std) de las Diferentes Variables para la Eliminación de Columnas Irrelevantes

Mo	Ti	Nb	Al	V	Temperature	H2S	CO2	Fluid velocity (Vs)
100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
0.14004000000000003	0.01362	0.0087	0.01256	0.01476	86.7	1280.1084799999999	1728.5929999999998	0.33
0.243008164638162	0.011461644425181447	0.01116225602626815	0.02950854010949948	0.029454680134372222	23.42212542866167	2634.4602795449646	2551.7587990514235	0.5184475660931187
0.0	0.0	0.0	0.0	0.0	60.0	0.048	0.0	0.0
0.0	0.0	0.0	0.0	0.0	60.0	10.0	73.65	0.0
0.0	0.018	0.0	0.0	0.0	90.0	50.0	569.0	0.0
0.29	0.024	0.019	0.0	0.0	100.0	1500.0	2000.0	1.0
0.72	0.028	0.026	0.091	0.081	150.0	12000.0	10500.0	1.5

Para la detección de los valores ruidosos, se utilizó la estadística de visualización de datos. Mediante programación, se generaron gráficas de caja y bigote, y mediante el conocimiento en corrosión, se determinó que ninguna variable fueran eliminadas.

Figura 10.

Gráficas de caja y bigote



2.3 Análisis univariado y correlación

Teniendo claro las características generales de cada dato individual, la idea en este apartado fue analizar la existen de alguna relación entre las variables predictoras y la variable a predecir(Pitting). Este análisis mostró una faceta de las posibles relaciones entre cada una de las variables y es la determinación de ciertos rangos de la variable x que están más inclinados al fenómeno de corrosión y cuánto es su porcentaje. Para ello se realizó un gráfico de tasa de conversión versus la variable a analizar. La tasa de conversión o proporción de corrosión La tasa de conversión proporcionó una idea general de qué condiciones o factores estuvieron asociados con una mayor probabilidad de corrosión por picaduras. Para ello se eligieron los

factores con mayor desviación estándar y que, según la literatura científica, son más comunes en incidir sobre la corrosión en picaduras.

El análisis de la matriz de correlación fue una herramienta estadística muy útil para el análisis de datos, ya que exploró las relaciones entre las variables, mediante un factor llamado coeficiente de Pearson que indica el nivel de significancia entre dos variables. Este análisis también se lo realizó evitar el sobre entrenamiento en el modelo de aprendizaje automático que se va a utilizó posteriormente en la ingeniería de variables y modelado.

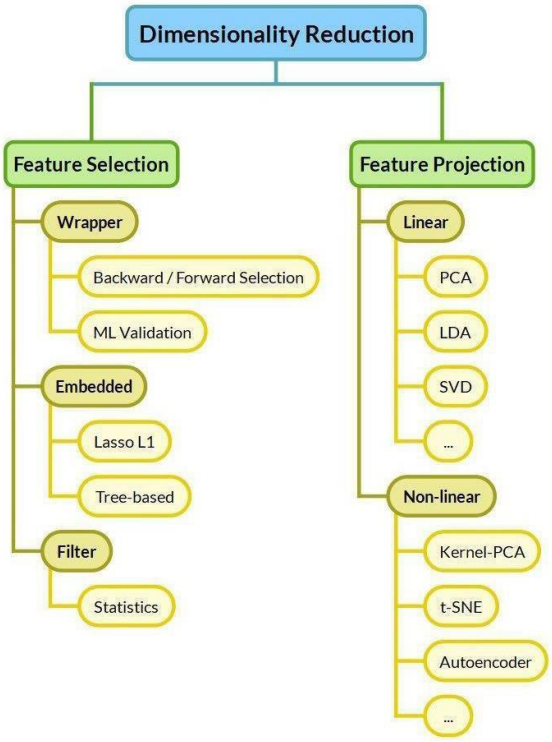
2.4 Ingeniería de características

Una vez explorado el conjunto de datos y analizado los valores inexistentes y extremos, y realizado el análisis de las variables, se prepararon los datos con la finalidad de construir el conjunto final de las características que se utilizaron para introducirlos en la etapa de modelado. Para ello se empleó la técnica de selección de características que consistió en reducir el tamaño del conjunto de características con la finalidad de: a) Mejorar el rendimiento de los algoritmos de aprendizaje automático; b) Aumentar la eficiencia computacional; c) Mejorar el análisis de características mediante la determinación de las características más importantes para establecer conclusiones.

Para este proyecto se eligió el método de envoltura o también denominado en inglés como “Whapped method”., que fue uno de los tantos métodos que se pudo utilizar como se observa en la figura. Este método permitió optimizar el rendimiento predictivo del modelo, a la par que resolvía uno de los objetivos de esta investigación y es el de saber cuál de las 26 características de corrosión tenían mayor importancia en la predicción de la corrosión en tuberías.

Figura 11.

Esquema del Método de Reducción de Dimensionalidades

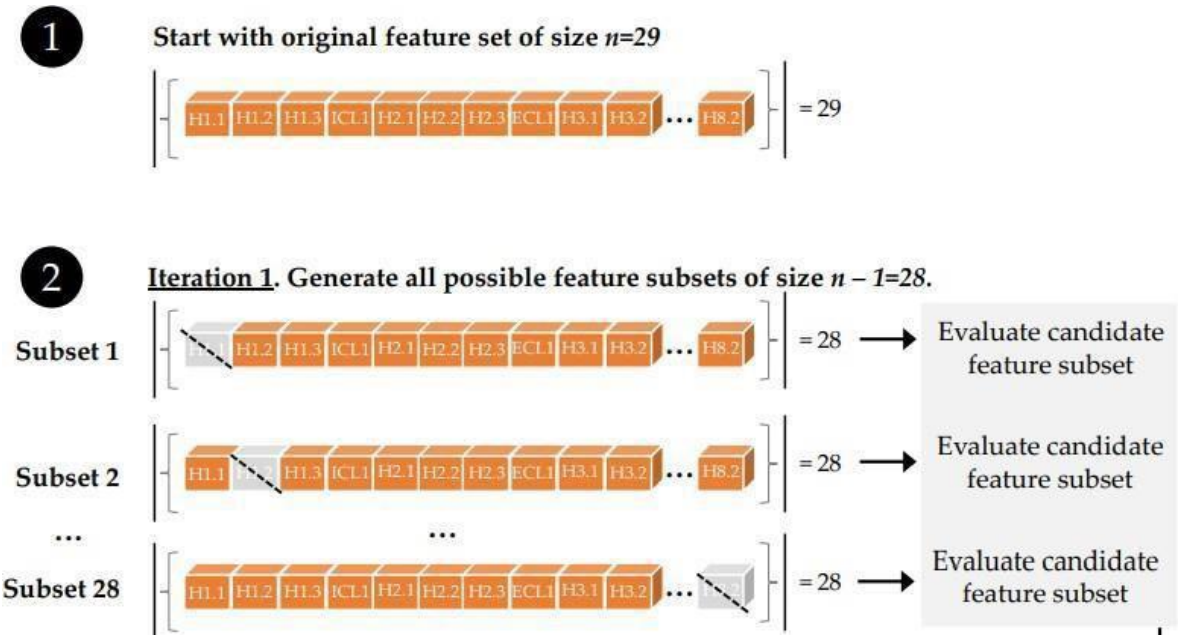


Nota. Reproducida de taxonomía de los algoritmos de reducción de Dimensionalidades, Ioannis Prapas,(www.learndatasci.com)

Dentro de este método existe una técnica llamada selección secuencial de características que consistía en encontrar el conjunto de características óptimo que garantice un rendimiento óptimo del modelo, sin necesidad de probar todas las combinaciones de características, sino simplemente un subconjunto de estas.

Figura 12.

Primera Iteración del Algoritmo Selección de Características



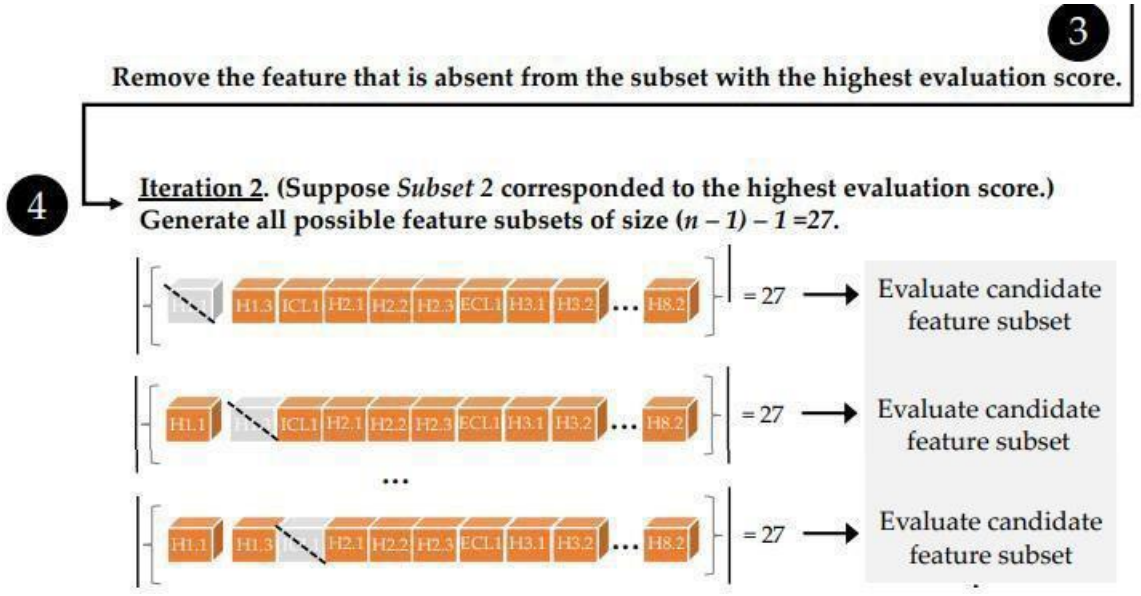
Nota. Información tomada de la página web www.sebastianraschka.com “sequential backward selection”, Sebastian Rashcka (2021)

El método comenzaba con el conjunto de características iniciales del proyecto, es decir: las 13 características de composición del acero, 8 características de solución y 4 características ambientales, la variable tiempo y la variable de tipo categórica si existe o no corrosión.

El siguiente paso consistía en la primera iteración del método: Se generaban los subconjuntos de características posibles de tamaño n-1, es decir, se el algoritmo consideraba el conjunto completo de características y eliminaba una característica a la vez. Se evaluaba mediante la validación cruzada y se registraba el rendimiento del subconjunto.

Figura 13.

Segunda Iteración del Algoritmo Selección de Características



Adaptado de “sequential backward selection”, Sebastian Rashcka (2021)

Posteriormente, el algoritmo evaluaba cuál de estos subconjuntos, presentaba un rendimiento más alto. El subconjunto con el rendimiento más alto, se seleccionaba para después, volver a repetir el proceso 1 y 2 hasta obtener el tamaño de características más deseado.

La finalidad de este método estuvo enlazada con uno de los objetivos del proyecto que consistía en identificar los factores de mayor importancia para asegurar la capacidad de generación del modelo.

2.4.1 CÓDIGO PARA LA IMPLEMENTACIÓN DEL MÉTODO SELECCIÓN SECUENCIAL DE CARACTERÍSTICAS

Importar librerías

Numpy es una librería que sirvió para convertir un dataframe en un arreglo. Requisito fundamental ya que los algoritmos de aprendizaje automático emplean arreglos.

Pandas es una librería de manipulación de datos. En este proyecto se lo utilizó para leer el archivo Excel que contienen los datos del conjunto de datos.

Figura 14.

Librerías Pandas y Numpy

```
import pandas as pd
import numpy as np
```

La librería Sklearn permite dividir el conjunto de datos de prueba y entrenamiento

Figura 15.

Módulos División en Entrenamiento y Prueba; Normalización de Valores

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
```

. Matplotlib permitió observar los resultados del rendimiento versus el número de características seleccionadas.

Figura 16.

Librería matplotlib

```
%matplotlib inline
import matplotlib.pyplot as plt
```

Mlxtend.feature_selection permitió importar la técnica SFS

Figura 17.

Modulo Mlxtend.feature_selection

```
from mlxtend.feature_selection import SequentialFeatureSelector as SFS
```

Métodos de cada librería

pd.read_csv es un módulo de pandas que permite la data en formato csv.

Figura 18.

Módulo de Pandas para Leer Data Csv

```
pd.read_csv
```

Se dividió el set de datos dependiendo de las características que fueron seleccionadas en la variable X, y la etiqueta en la variable Y:

Se llamó al módulo train_test_split con la finalidad de separar la variable X y Y en datos de prueba y entrenamiento. Los argumentos de este módulo, fueron las variables numéricas X, variable categórica Y, el porcentaje total de las observaciones que se dividieron en set de prueba y la cantidad de veces en las que las filas se combinan aleatoriamente.

Figura 19.

Variables de Entrenamiento y Prueba X e Y

```
X, y = dataset.iloc[:, 1:-1] ,dataset.iloc[:, -1]
X_train, X_test, y_train, y_test = train_test_split(
    X,
    y,
    test_size=0.3,
    random_state=42)

X_train.shape, X_test.shape
```

Para las observaciones numéricas fue necesario la normalización ya que resulta más sencillo para los modelos, trabajar con valores entre rangos de 0 y 1. Para eso se utilizó en módulo StandarScaler.fit_transform para transformar la data de entrenamiento y StandardScaler.transform para transformar la data de prueba.

Figura 20.

Normalización de las Variables Dependientes Tanto de Prueba Como de Entrenamiento

```
sc = StandardScaler()
X_train_std = sc.fit_transform(X_train)
X_test_std = sc.transform(X_test)
```

Se ajustaron los diferentes modelos de clasificación como por ejemplo el clasificador de los k vecinos más cercano en el conjunto de datos normalizados de entrenamiento. Para el módulo que permite importar el modelo de clasificación, se empleó arbitrariamente 5 “vecinos”.

Figura 21.

Definición y Entrenamiento del Modelo K vecinos

```
model = KNeighborsClassifier(n_neighbors=3)

model.fit(X_train_std, y_train)

print('Training accuracy:', np.mean(model.predict(X_train_std) == y_train)*100)
print('Test accuracy:', np.mean(model.predict(X_test_std) == y_test)*100)
```

```
Training accuracy: 84.28571428571429
Test accuracy: 83.33333333333334
```

Se importó el selector de características secuenciales (SFS) para poder seleccionar las mejores características. Como entrada recibe al modelo, el subconjunto de mejores características que se desea conocer, el tipo de SFS a utilizar, si se desea el SFS regular, métrica de puntuación a utilizar y la cantidad de veces que se utilizó la validación cruzada para la evaluación del subconjunto de características.

Por último, se entrenó el modelo con el módulo fit.

Figura 22.

Módulo SFS para determinar las variables más importantes.

```
sfs1 = SFS(model,
           k_features=5,
           forward=True,
           floating=False,
           verbose=2,
           scoring='accuracy',
           n_jobs=-1,
           cv=5)


sfs1 = sfs1.fit(X_train_std, y_train)
```

Para saber los índices de las mejores características del conjunto de datos se usó el módulo k_feature_idx_

Figura 23.

*Módulo k_feature_idx_*

```
sfs1.k_feature_idx_
```

Una vez seleccionado el mejor subconjunto, se evaluó el rendimiento del modelo ajustado a este subconjunto en particular.

2.5 Predicción de corrosión

Para predecir la existencia de corrosión se emplearon 2 modelos de clasificación: clasificador de k vecinos más cercanos y árbol de decisión. Entonces primero se definieron dos variables que contenían los módulos de dichos modelos.

Figura 24.

Variables que Definen los Algoritmos de Clasificación

```
# Call the Machine Learning Algorithms for Classfication
knn = KNeighborsClassifier()
decision_tree = DecisionTreeClassifier()
```

Se definieron los hiperparámetros de los dos algoritmos. Para ello, se generaron dos diccionarios que contienen diferentes valores de k vecinos y diferentes parámetros del árbol de decisión. El método arange permite generar valores del 1 al 30 para los k vecinos, y del 2 al 20 para la máxima amplitud en el árbol de decisión.

Figura 25.

Diccionarios que Contienen los Dos Hiperparámetros de los Algoritmos de Clasificación

```
knn_params = {'n_neighbors': np.arange(1, 30)}
dt_params = {'max_depth': np.arange(2,21,1)}
```

Se creó la función training_model con la finalidad de entrenar cualquier modelo sin necesidad de volver a programar. Dentro de la función se agregaron dos parámetros: model, donde se introduce el modelo y param_grid que hace referencia a los hiperparámetros usados por el modelo. GridSearchCV es un módulo que sirvió para identificar el mejor hiperparámetro.

Figura 26.

Función Training_model.

```
# Function to train each algorithm

def training_model(model, param_grid):
    model = GridSearchCV(model, param_grid=param_grid, scoring='accuracy')
    model.fit(X_train, y_train)
    return model
```

El atributo best_params_ identificó el mejor hiperparámetro para este modelo en concreto, mientras que el atributo best_estimator_fit reentrena el algoritmo usando el mejor parámetro. Finalmente, con el conjunto de prueba se evaluó el rendimiento del algoritmo, es decir, se verificó que eficiencia que tiene el algoritmo a la hora de estimar predicciones de datos no vistos previamente.

Figura 27.

Predicción Mediante los Datos de Prueba

```
knn_model = training_model(knn, knn_params)
knn_model.best_params_
```

```
{'n_neighbors': 1}
```

```
knn_model_final = knn_model.best_estimator_.fit(X_train_std, y_train)
```

```
y_pred_knn = knn_model_final.predict(X_test_std)
```

```
# Accuracy of model
print(knn_model_final.score(X_train_std, y_train))
print(knn_model_final.score(X_test_std, y_test))
```

```
1.0
0.8333333333333334
```

```
# Accuracy
accuracy_knn = accuracy_score(y_test, y_pred_knn)
print(accuracy_knn)
```

```
0.8333333333333334
```

CAPITULO 3

3. RESULTADOS Y DISCUSIÓN

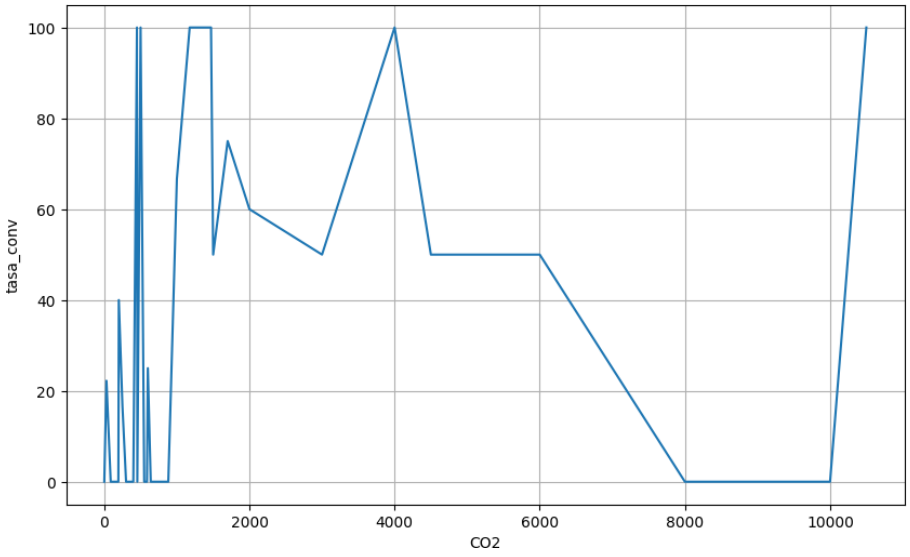
3.1 Análisis univariado de variables

El análisis univariado de variables muestra una faceta de las posibles relaciones entre una de las variables y la variable a predecir. Para ello se usó la proporción de casos en los que ocurre la corrosión en relación con el total de casos analizados (tasa de conversión). En este contexto se puede utilizar esta medida para analizar los factores/variables que afectan la corrosión por picaduras en tuberías de acero de petróleo. Dado que se trata de un factor binario, la tasa de conversión se convierte en una medida de prevalencia de corrosión en relación con el total de observaciones. En este trabajo, se analizaron los factores que podrían resultar más importantes sacar información, sobre el rango de valores que resultan corrosivos en este estudio: en este caso se analizaron el dióxido de carbono, el pH, temperatura y H2s. Por ejemplo: La forma en que la salinidad afecta a un material específico puede variar según la concentración de salinidad, la temperatura, la presencia de otros compuestos corrosivos, el flujo de fluidos y otros factores ambientales. Por lo tanto, para comprender completamente cómo la salinidad afecta a un material en particular, es necesario para establecer que rangos potenciales puede conllevar al fenómeno de corrosión

En la figura, se puede observar el comportamiento de la variable dióxido de carbono, sobre el porcentaje de casos en el que ocurre la corrosión. Se empezó a detectar ciertos patrones en los datos como, por ejemplo: de 8000 a 10.000Kpa, la tasa de conversión es del 0%, es decir, para este rango de valores, el 0% de las observaciones de los ductos, presentan corrosión.

Figura 28.

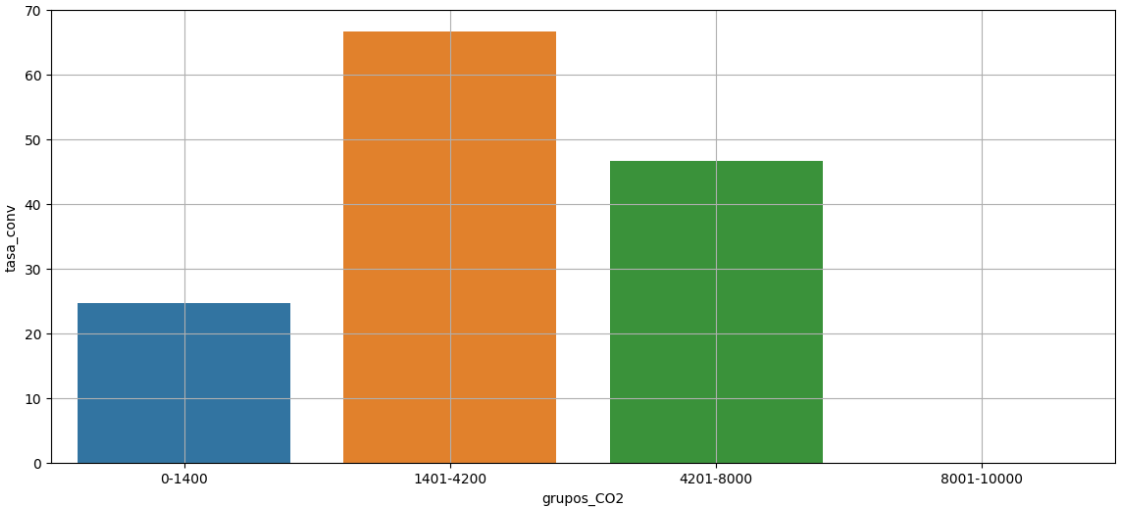
Gráfica de Línea Continua Proporción de Corrosión Vs Co2



Se establecieron 4 rangos aproximados para los valores de Co2 y se visualizaron de forma más detallada para determinar el grupo con la mayor tasa de conversión. En la figura, se establecieron los rangos de 0-1400kpa, 1400-4200kpa, 4200-8000kpa y 8000-10000kpa. Estos rangos dependieron de los patrones que adquiría la gráfica de línea continua, vista en el anterior gráfico. Finalmente se estableció que, para un rango de valores entre 1400-4200kpa, se tiene una mayor proporción de corrosión, es decir, mayor probabilidad de que los ductos se encuentren corroídos, con un 67% de probabilidad.

Figura 29

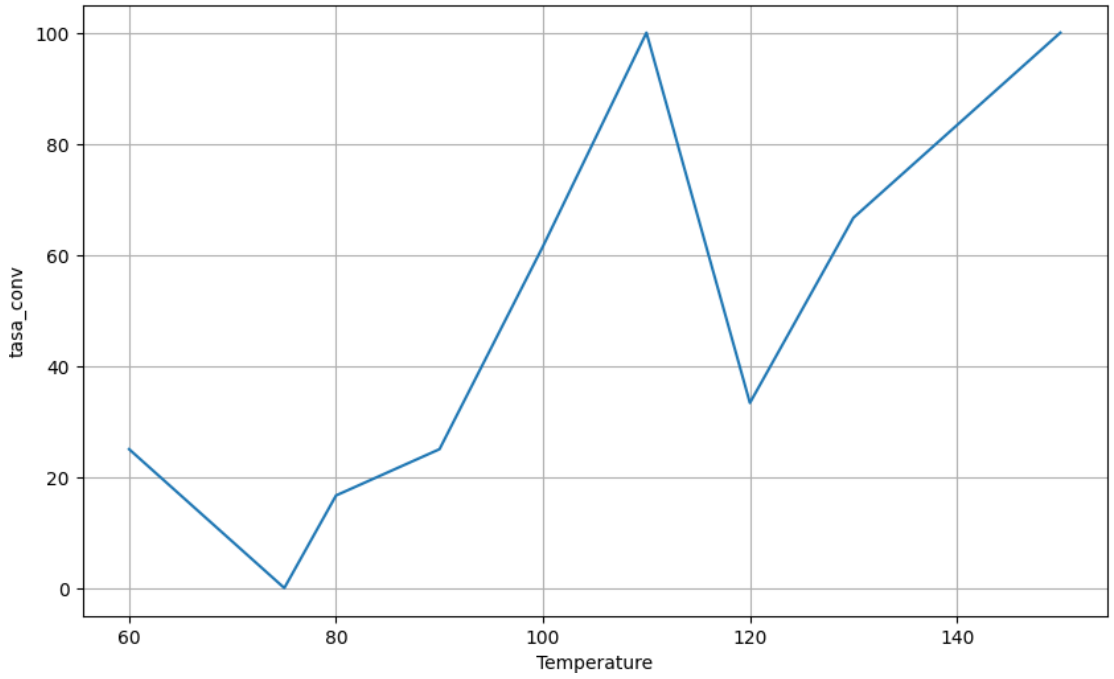
Gráfico de Barras Proporción de Corrosión vs Rangos



El mismo procedimiento se realizó para la temperatura. Se observaron patrones desde los 60-78, 78-100 ,100-120, y 120-150 grados centígrados, y se realizó una gráfica con estos rangos de valores versus la tasa de conversión o proporción de corrosión.

Figura 30.

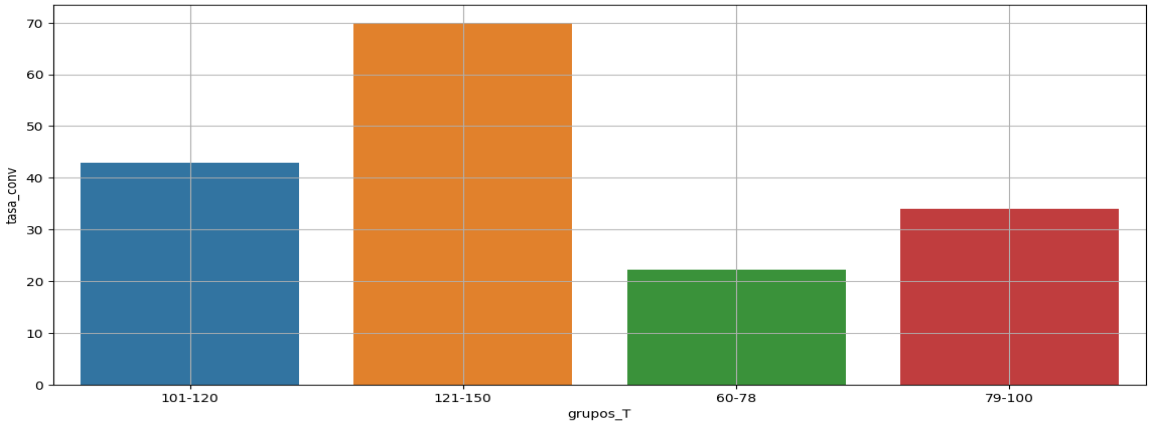
Gráfico de línea Proporción de Corrosión vs Valores Temperatura



En la figura 31, se tiene que hasta un 70% de las mediciones a los ductos, presentan corrosión con valores de temperatura circundantes entre los 120-150 grados centígrados.

Figura 31.

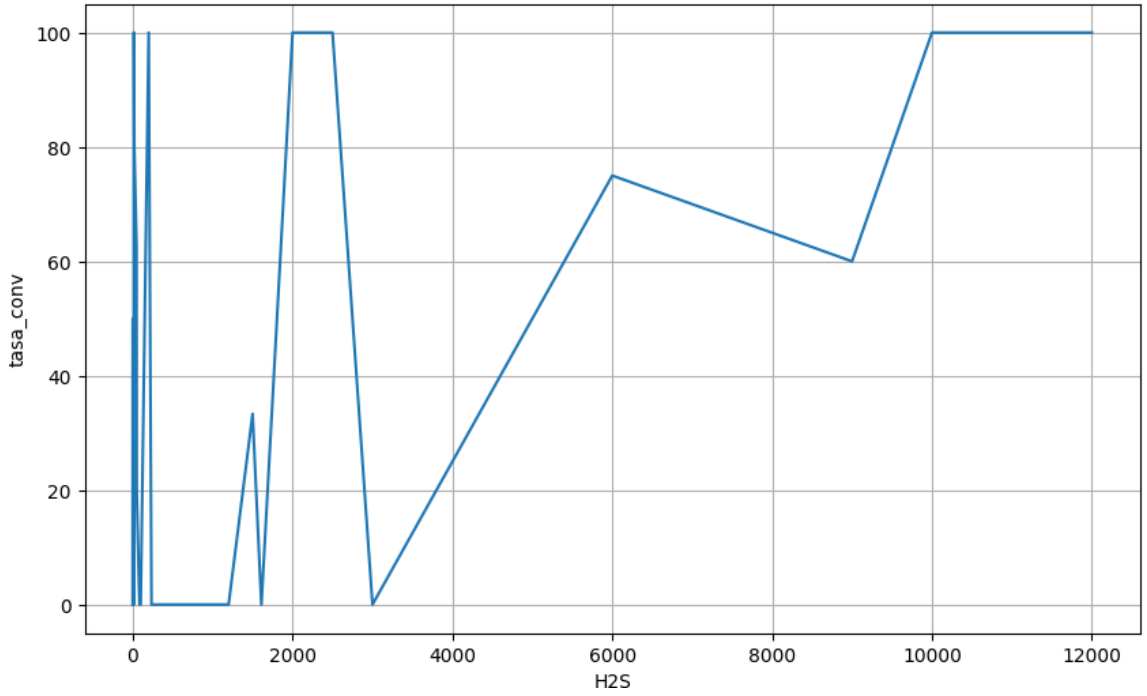
Gráfico de Barras Proporción de Corrosión vs Rangos Temperatura



En la figura 32, los valores concretos de 0 y 2000kPa, se puede apreciar una tasa de conversión en 100%, mientras que los valores oscilantes entre este rango, la mayoría presentan 0% de corrosión. De 2000kpa a aproximadamente 3000kpa existe una disminución de la tasa de conversión abrupta. A partir de 3000kpa hasta 6000kpa existen una relación lineal positiva entre el aumento del h2s y la tasa de conversión. De 6000-9000kpa existe una disminución lineal de la proporción de corrosión, de 9000-10000Kpa presenta nuevamente una relación positiva lineal. Finalmente, a partir de 10.000Kpa, siempre va a presentar una tasa de corrosión del 100%

Figura 32.

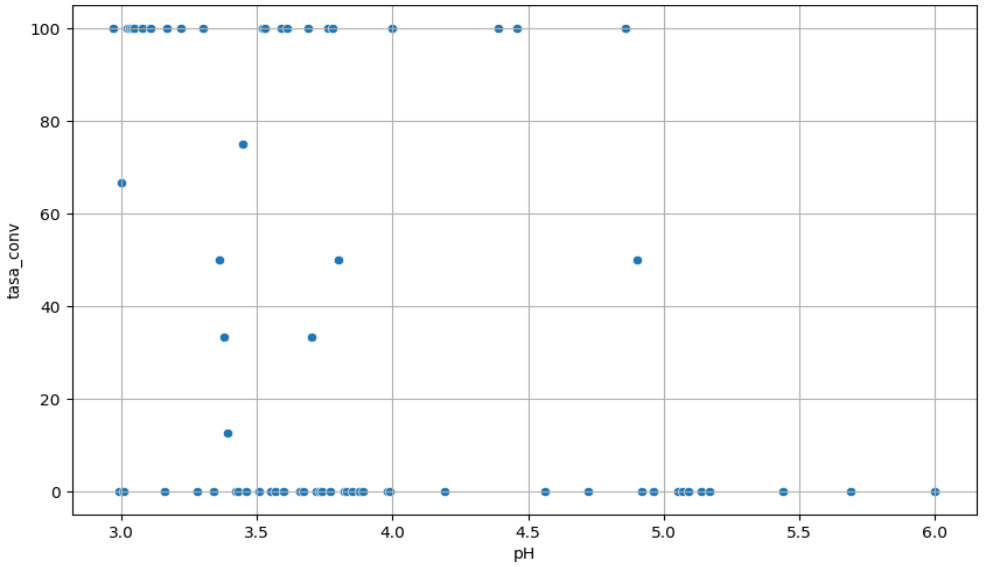
Gráfico de línea Proporción de Corrosión vs Valores H2s



Para la variable pH, se realizó un gráfico de dispersión en el que se puede apreciar para un rango de valores, la disparidad de resultados. Entonces se eligió un rango entre 3-3.5, 3.5-4 4-4.5,4.5-5

Figura 33.

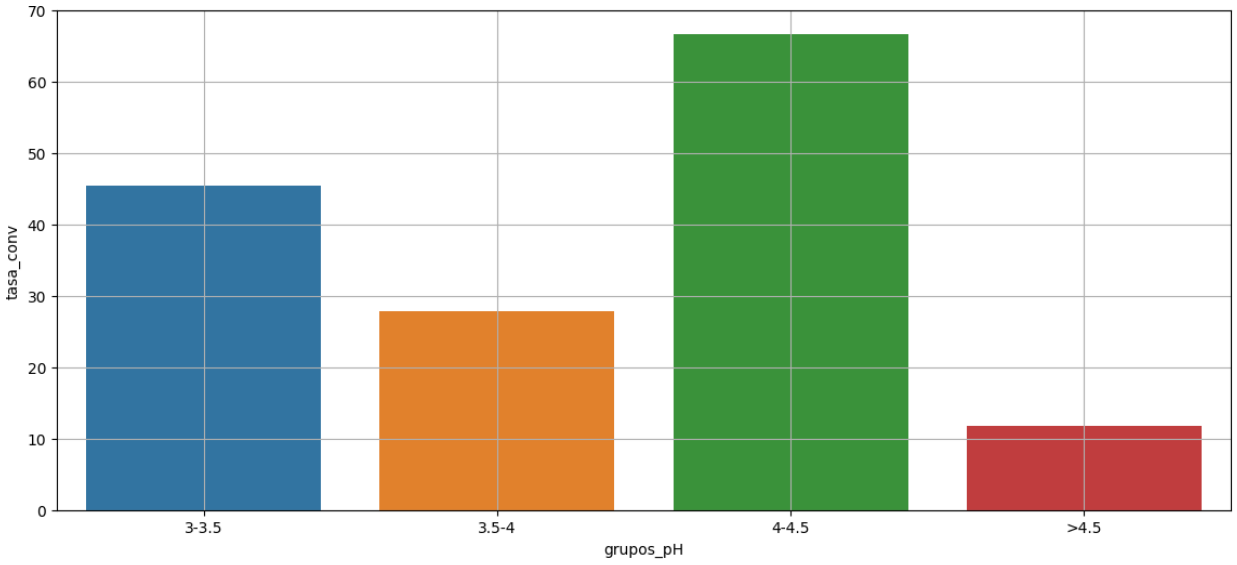
Gráfico de Dispersión Proporción de Corrosión vs Valores pH



En la figura se puede apreciar, que, para un rango de 4-4.5 de pH, se tiene una proporción mayor de corrosión alrededor del 67%.

Figura 34.

Gráfico de barras Proporción de Corrosión vs grupos pH



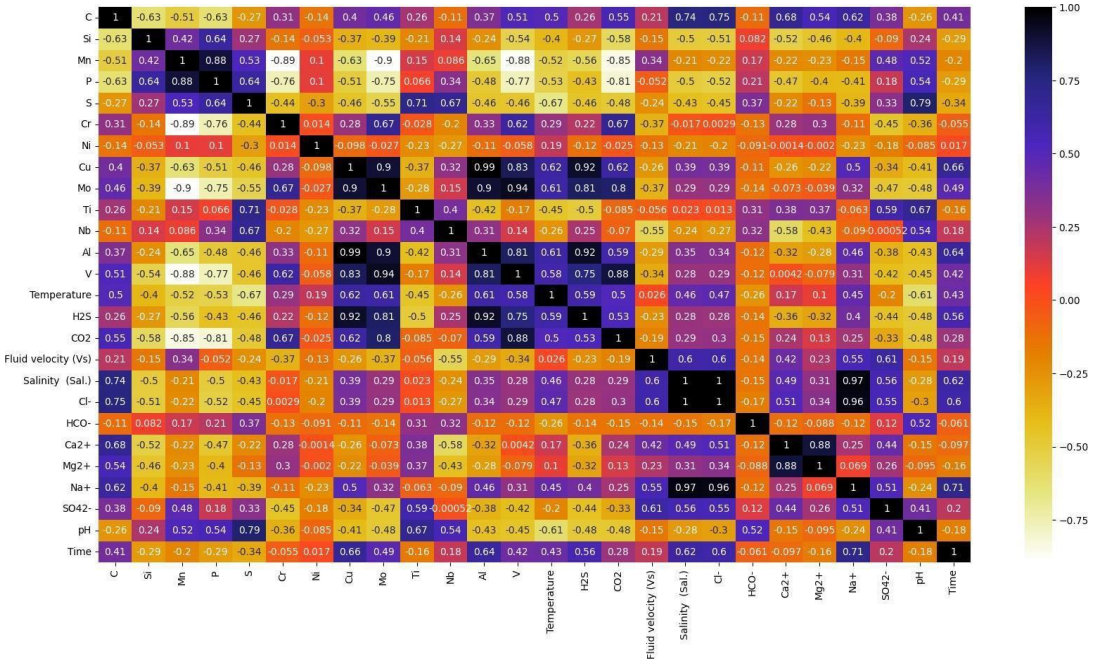
3.2 Análisis de correlación de Pearson

Después se realizó el análisis de correlación de Pearson. Uno de los componentes fundamentales en el análisis de datos es explorar las potenciales relaciones entre las variables que están siendo analizadas. Para ello, comúnmente se utiliza una medida estadística llamada correlación. La correlación nos indica la magnitud que pueda haber entre las variables.

En la matriz que se presenta se puede observar la matriz de correlación de Pearson con sus correspondientes magnitudes. Debido a la gran cantidad de valores a analizar, se creó una función que detecta las variables correlacionadas entre sí por más de un 90%. Estas variables fueron: {'Al', 'Cl-', 'H2S', 'Na+', 'V'}. Para evitar el sobreajuste, se eliminaron estas variables del dataset original y se procedió a calcular las variables más importantes del dataset usando ingeniería de características. A continuación, se presenta la matriz:

Figura 35.

Matriz de Correlación de las Variables Dependientes



En la matriz de correlación se puede observar 3 de las características de solución con valores de Pearson muy elevados: Salinidad, Ca y Cl, que varían de 0.96-1, por lo que se pudo afirmar con un alto porcentaje de confianza, que existe una correlación positiva muy alta entre estas 3 variables. La correlación entre la salinidad, e iones Na+ y Cl- tiene sentido debido a la naturaleza química de los elementos involucrados y cómo actúan sobre un ambiente corrosivo. La salinidad se refiere a la concentración de sales disueltas de las que se encuentra el Cloruro de sodio (NaCl) y generalmente es una de las principales sales disueltas en el agua de formación que se extrae junto al petróleo durante la producción.

Por otra parte, existió una alta correlación entre el sulfuro de hidrógeno con el aluminio y cobre. Esto fue un resultado indicativo ya que el H2s, Al y Cu tienen una relación específica en el contexto de corrosión en ductos de transporte de petróleo ya que algunos materiales son más susceptibles a la corrosión en presencia de H2s.

3.3 Ingeniería de características

Para cumplir con el objetivo de determinar las características más importantes que afectan a la corrosión por picaduras, se usó el método de selección secuencial de características.

Para esto se usó la librería de Scikit-Learn para este análisis, Se importó el selector de características secuenciales (SFS) para poder seleccionar las mejores características.

Finalmente, en el cuadro de abajo se puede ver las 4 variables más importantes en el dataset, que inciden en la variable de corrosión. Estas cuatro variables son el azufre(S), cobre(Cu), el dióxido de carbono(Co2) y el ácido sulfúrico (So4,) con una precisión de entrenamiento del 84.3% y de prueba del 83.3%. Las variables antes mencionadas son relevantes en el contexto de la corrosión en la industria petrolera.

Figura 36.

Entrenamiento del Modelo y Precisión Data de Entrenamiento y Prueba

```
model = KNeighborsClassifier(n_neighbors=3)

model.fit(X_train_std, y_train)

print('Training accuracy:', np.mean(model.predict(X_train_std) == y_train)*100)
print('Test accuracy:', np.mean(model.predict(X_test_std) == y_test)*100)
```

```
Training accuracy: 84.28571428571429
Test accuracy: 83.33333333333334
```

Figura 37.

Lista de Las Características más Importantes en la Predicción de Corrosión

```
dataset.columns[1:][list(sfs1.k_feature_idx_)].tolist()
```

```
['S', 'Cu', 'V', 'CO2', 'SO42-']
```

El azufre generalmente se forma en la fabricación del acero, como pequeñas partículas de sulfuro de hierro que se forman en el proceso de fundición en la estructura del metal. Esto puede debilitar el material y servir como un sitio inicial para la corrosión por picaduras.

El cobre actúa como cátodo y otro metal diferente actúa como ánodo más el contacto con un electrolito en este caso el agua de formación, influyen sobre el accionar de los agentes corrosivos.

El Co2 en presencia del agua de formación que se encuentra asociada el crudo forma el ácido carbónico que resulta ser un ácido débil. Este proceso es muy importante ya que influye en la capacidad de corrosión de los ductos. La concentración de este ácido depende de varios factores, entre ellos, la temperatura de la solución y el pH del agua. En el acero, el ácido carbónico interactúa con esta capa debilitando la protección contra corrosión

En términos de la industria petrolera, el CO2 suele estar presente en una solución corrosiva en forma de gas disuelto en el petróleo y es la principal causante de los problemas de corrosión interna por picaduras en ductos. HCO3- resulta otra característica de solución que afecta, aunque de menor medida el estado del acero. Este compuesto anódico se forma usualmente cuando el CO2 reacciona con el agua de formación que también contiene el petróleo y acidifica de forma local el ducto.

El sulfato resulta ser un ion presente en muchas aguas de formación asociadas a la extracción de crudo. Como en las otras variables, la corrosión de los ductos depende de la concentración de sulfato y las propiedades específicas del material y condiciones del entorno.

El agua acompañante del petróleo suele ir contaminado de ácido sulfhídrico (Hs2), este compuesto se descompone en H+ y HS-. Como los iones de sulfuro se obtienen en la cercanía de la superficie del metal, todas las concentraciones de Hs2 puede acelerar la acidez local del acero.

En presencia del ácido sulfhídrico la velocidad de corrosión se incrementa aún más con el aumento de la temperatura. La existencia de zonas expuestas al medio agresivo, crea un diferencial de potencial entre el Fe y el sulfuro de hierro que acaba desarrollándose en corrosión por picadura.

El efecto del cloruro (Cl-) sobre las paredes superficiales del oleoducto también tienen una gran influencia en las picaduras del acero, a mayor contenido de cloruro en el agua de formación, más grave resultaría la fisura.

El Ca+2 y Mg+2 tienen una capacidad de influir sobre la susceptibilidad en picaduras considerablemente grande. Estos componentes pueden cambiar la solubilidad del Co2 y H2s, haciendo que aumente. Cuando estos compuestos se disuelven en el agua, forman ácidos como (H2CO3) que puede corroer metales de acero.

En esta sección se presentan los resultados de los modelos seleccionados en este trabajo (árbol de decisión, SVM), con los que fue entrenado el conjunto de dato. La relación de división del conjunto de entrenamiento seleccionado fue 80%.

3.4 Predicción: Caso de estudio para la predicción de corrosión: ducto X80

Para poner en práctica la aplicabilidad del algoritmo, se utilizaron un ejemplo distinto al de las muestras de ductos analizadas en el conjunto de datos, que son las muestras del ducto de acero X80. Este ducto se ha desarrollado debido a la reciente demanda y transporte de gas natural que envuelve China. Gracias a su alta resistencia, alta tenacidad y buena soldabilidad, la composición química del ducto X80 se basa en un bajo contenido de carbono, alto contenido de manganeso, niobio y aleaciones apropiadas de cobre, níquel y molibdeno. A continuación, en la siguiente tabla se presenta la composición química de la muestra:

Tabla 4.

Composición química de la Muestra del Ducto X80

X80 pipeline Steel										
C	Si	Mn	P	S	Cr	Ni	Mo	Ti	Nb	V
0.061	0.28	1.86	0.011	0.0006	0.03	0.03	0.22	0.016	0.061	0.059

Tanto las características ambientales como las características de solución, fueron adaptadas de acuerdo a las condiciones representativas del entorno en el que se utiliza el material. Cabe recalcar que estos valores son ejemplos según condiciones específicas de un proyecto, ubicación geográfica, fuente de petróleo y gas. En la siguiente tabla se presentan las características de solución y ambientales propuestas:

Tabla 5.

Características de Solución de la muestra X80

Características solución							
Velocidad fluido	Salinidad	Cl-	HCO-	Ca2+	Mg2+	Na+	SO4
5 m/s	5000 mg/l	1000 mg/l	100 mg/l	200 mg/l	150 mg/l	3000 mg/l	100 mg/l

Tabla 6.

Características de Ambientales de la muestra X80

Características ambientales			
Temperatura,	'H2S',	'CO2',	pH
30	55 Kpa	180 Kpa	5.5

El resultado a estos datos fue que dicha muestra del ducto, presentó corrosión con un 83.3% de precisión para el modelo de K vecinos

CAPITULO 4

4. CONCLUSIONES Y RECOMENDACIONES

4.1 CONCLUSIONES

1. Las gráficas de proporción de corrosión (tasa de conversión) versus todas las variables analizadas, no presentaron linealidad lo que quiere decir que las características que involucran a la corrosión no siguen una relación directamente proporcional, lo que significa que a medida que se cambian las características de corrosión, la proporción de corrosión puede aumentar en algunos puntos y disminuir en otros. Sin embargo, al analizar los valores de las variables por rangos, se proporciona información valiosa sobre las condiciones en las que la corrosión es más pronunciada y permitir identificar rangos específicos de factores de corrosión asociados a una mayor tasa de corrosión.

2. Las variables 'Al', 'Cl-', 'H2S', 'Na+', 'V', están altamente correlacionadas con un 90% de correlación, lo que sugiere que los cambios en una o más de estas variables suelen estar acompañadas con cambios en las otras de manera predecible. Este cambio es de forma lineal, lo que significa que a medida que un valor de una variable aumenta, el valor de otra tiende a aumentar.

3. La implementación del método de selección secuencial de características fue un método beneficioso, ya que ayudó a identificar las variables más influyentes en el conjunto de datos en relación con la corrosión, mediante una visión fundamental de entender las variables críticas que contribuyen a este fenómeno en los ductos de transporte de petróleo.

4. El dióxido de carbono (Co2) se rige como un factor fundamental influyendo en la formación de ácido carbónico que influye en la acidez del medio y acelera los procesos corrosivos en los ductos.

5. La precisión del conjunto de datos de entrenamiento y prueba sugieren que el modelo de selección de características es capaz de generalizar y predecir la corrosión con un buen nivel de acierto. Muestra que el modelo no está sobre ajustado en los datos de entrenamiento lo que es una buena señal en su capacidad para manejar nuevas muestras.

6. Para la predicción de corrosión en muestras de ductos de petróleo mediante el modelo de k vecinos más cercanos, se llegó a un accuracy de 1 en el conjunto de entrenamiento que puede ser motivo de cautela en términos de sobreajuste, sin embargo, un accuracy de 0.83 en el conjunto de prueba es un indicador positivo que el modelo es capaz de generalizar y hacer predicciones de precisas de datos nuevos.

4.2 RECOMENDACIONES

1. Para mejorar el análisis de proporción de corrosión frente a las variables que las afectan se recomienda realizar un estudio detallado con la finalidad de comprender como estas variables interactúan entre sí y como sus efectos se potencian o contrarrestan mutuamente.

Realizar un análisis a profundidad para comprender si las correlaciones observadas implican relaciones causales directas o si existen factores subyacentes que impulsan a estas correlaciones.

2. Con respecto al método de selección secuencial de características y corrosión en ductos de transportes de petróleo se recomienda comparar los resultados de este método con otros métodos como el análisis de componentes principales para evaluar la eficiencia del método elegido en relación con el fenómeno de corrosión.

3. El evidente daño que causa el dióxido de carbono sobre las capas de acero de los ductos de petróleo hace necesario la transformación en la que la industria petrolera aborda el problema de la corrosión para ofrecer beneficios que sean relevantes en términos de eficiencia y sostenibilidad.

4. Se recomienda utilizar técnicas de regularización, como la validación cruzada para abordar el posible sobre ajuste del modelo y mejorar aún más su capacidad.

5. BIBLIOGRAFÍA

Vajo; Wei; Phelps ; Reiner ;Herrera; Cervantes ; Gidanian ; Bavarian ; Kappes . (2002). Application of extreme value analysis crecive corrosion. *Elsevier*,*45,* 497-509. doi:10.1016/S0010-938X(02)00129-4

Aja, M. B. (2014). *Análisis del Riesgo de Rotura en Servicio de Tuberías de gas natural.* Cantabria.[Tesis de grado Escuela Técnica Superior De Ingenieros Industriales y Telecomunicación] https://repositorio.unican.es/xmlui/bitstream/handle/10902/4417/364006.pdf?sequence=1&isAllowed=y

Álvarez, D. (2021). La metodología CRISP-DM*. Adictos al trabajo.* https://www.adictosaltrabajo.com/2021/01/14/metodologia-crisp-dm/

Barquín, M. (2014). *Análisis del riesgo de rotura en servicio de tuberías de gas natural.* [Tesis de grado Escuela Técnica Superior De Ingenieros Industriales y Telecomunicación] https://repositorio.unican.es/xmlui/bitstream/handle/10902/4417/364006.pdf?sequence=1&isAllowed=y

DXP IFS integrated flow solutions. (2019).*Features and benefits of pipeline transportation-Whoy pipelines are needed*.*Ifsolutions*. https://ifsolutions.com/features-benefits-of-pipeline-transportation-why-pipelines-needed/#:~:text=Pipelines%20can%20transport%20enormous%20amounts,rail%2C%20truck%2C%20or%20ships.

ECCA (2011). The basics of corrosion Technical Paper. *prepaintedmetal.eu*. https://prepaintedmetal.eu/repository/downloads/1.%20The%20Basics%20of%20Corrosion.pdf

Ehrnstén, U. (2020). Corrosion and stress corrosion cracking of austenitic stainless steels. *Comprehensive nuclear materials, 4,* 118-128.

Haya, P. (2021).La metodología CRISP-DM en ciencia de datos. *iic.uam.es* https://www.iic.uam.es/innovacion/metodologia-crisp-dm-ciencia-de-datos/#_ftn3

Huerta, E. O. (1997). *Corrosión y degradación de materiales(2a ed.*) Editorial Síntesis.

Islam, M. (2015). *Erosion, Corrosion and Erosion-Corrosion fo oil and gas pipeline steels.[Tesis de grado* Halifax, Nova Scotia] Redalyc. file:///C:/Users/Hp/Downloads/Islam-Md._Aminul-PhD-MATL-November-2015%20(4).pdf

Javaherdashti, R. (2008). *Microbiologically influenced corrosion.* Engineering Materials and Process, Western Australia,Springer. doi:10.1007/978-1-84800-074-2

Jun hu, Tian Yangyang, Haipeng Teng, Lijun Yu, Maosheng Zheng (2014). The probabilistic life time prediction model of oil pipeline corrosion crack. 1.

kdnuggets. (2014).What main methodology are you using for analytics data mining or data science projects?. https://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html

Nasiriany. (2019). Decision Tree Learning. *A comprehensive Guide to Machine Learning*,163-165. Universidad de California, Berkeley. https://snasiriany.me/files/ml-book.pdf

Sandoval, L. J. (2018). Algoritmos de aprendizaje automático para análisis y predicción de datos, 37. http://www.redicces.org.sv/jspui/bitstream/10972/3626/1/Art6_RT2018.pdf

Yin, Zhao, Bai, Feng, & Zhou. (2008). *Corrosion behavior of SM 80SS tube steel in stimulant solution containing H2S and CO2.* Editorial Board. doi:https://doi.org/10.1016/j.electacta.2007.12.039

Zhihao Qu, D. T. (2021). *Pitting Judgment Model Based on Machine Learning and Feature Optimization Methods.Frontiersin,8,1-8,* https://doi.org/10.3389/fmats.2021.733813

Zukhrufany, S. (2018). *The utilization of supervised machine learning in prediction corrosion to support preventing pipelines leakage in oil and gas industry.[Tesis de postgrado Universitetet i Stavanger]*. Redalyc. https://uis.brage.unit.no/uis-xmlui/bitstream/handle/11250/2565865/Zukhrufany_Stiffi.pdf?sequence=4&isAllowed=y

https://github.com/brycgonz/Tesis-Analisis-corrosion/blob/main/matriz_correlacion_tesis.ipynbe.com/drive/1oyYmSS8USOjrWPmgL2XjQGe5QoXMZq28?hl=es#scrollTo=9Tt0NNRlJNvh

https://github.com/brycgonz/Tesis-Analisis-corrosion/blob/main/tesis.ipynb

6. APÉNDICE

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from google.colab import drive
drive.mount('/content/drive')
ruta="/content/drive/MyDrive/tesis/dataset.XLSX"
data=pd.read_excel(ruta)
data.head()
data.dropna(inplace=True)
data.info()

df_num=data.loc[:,data.columns!="pit"].astype("float64")
df_num.info()

dataset=pd.concat([df_num,data["pit"]],axis=1,join="outer")
dataset.info()

print(f'Tamaño del set antes de eliminar las filas repetidas:
{dataset.shape}')
dataset.drop_duplicates(inplace=True)
print(f'Tamaño del set después de eliminar las filas repetidas:
{dataset.shape}')

cols_num =dataset.columns.tolist()
print(cols_num)
print(len(cols_num))
fig,ax=plt.subplots(nrows=4,ncols=7,figsize=(15,10))
plt.tight_layout()
```

```
fig.subplots_adjust(hspace=1)
for i,col in enumerate(cols_num):

  ax[i // 7, i % 7].set_title(col)

# Comencemos representando la variable a predecir de forma binaria:
# 'yes' = 1, 'no' = 0
diccionario = {'yes':1, 'No':0}
binario = dataset['pit'].map(diccionario)
dataset['y_bin'] = binario

# De esta forma resulta fácil calcular la tasa de conversión: el promedio
# de la columna 'y_bin'

fig, ax = plt.subplots(nrows=4, ncols=7, figsize=(15,10))
fig.subplots_adjust(hspace=20)
plt.tight_layout()
for i, col in enumerate(cols_num):
  bplt = sns.boxplot(x="y_bin", y=col, data=dataset, ax=ax[i // 7, i % 7])
  ax[i // 7, i % 7].set_xlabel('y_bin (1: yes, 0:no)')
  plt.tight_layout()
  ax[i // 7, i % 7].set_title(col)
  plt.tight_layout()

# Función para graficar proporción de corrosion

def graficar_tasas_conversion(var_num, var_predecir, type='line',
order=None):
    x, y = var_num, var_predecir
```

```
    # Generar agrupaciones (groupby), calcular proporción de corrosión
(mean),
    # multiplicarla por 100 (mul(100))
    grupo = dataset.groupby(x)[y].mean().mul(100).rename('Proporcion
Corrosion').reset_index()

    # Y generar gráfica
    if type=='line': # Útil para rangos continuos
        plt.figure(figsize=(10,6))
        sns.lineplot(x=var_num, y='Proporcion Corrosion', data=grupo)
        plt.grid()
    elif type=='bar': # Útil si los datos están divididos en rangos o son
categóricos
        plt.figure(figsize=(14,6))
        sns.barplot(x=var_num, y='Proporcion Corrosion', data=grupo,
order=order)
        plt.grid()
    elif type=='scatter': # Útil si los datos están divididos en rangos o
son categóricos
        plt.figure(figsize=(10,6))
        sns.scatterplot(x=var_num, y='Proporcion Corrosion', data=grupo)
        plt.grid()

# Nueva columna en el dataset: "grupos_CO2"
dataset.loc[:,'grupos_CO2'] = "0-1400"
dataset.loc[(dataset['CO2']>1400)&(dataset['CO2']<=4200), 'grupos_CO2'] =
"1401-4200"
dataset.loc[(dataset['CO2']>4200)&(dataset['CO2']<=8000), 'grupos_CO2'] =
"4201-8000"
dataset.loc[(dataset['CO2']>8000)&(dataset['CO2']<=10000), 'grupos_CO2'] =
"8001-10000"
```

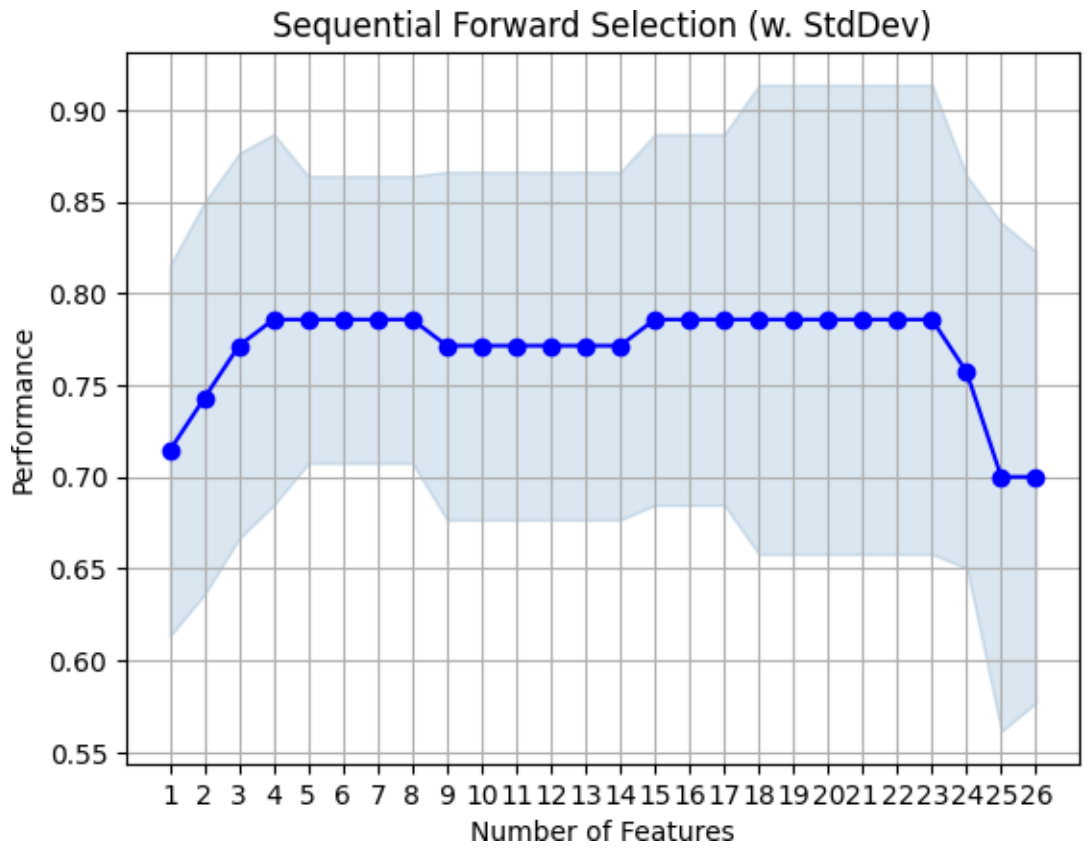
```
graficar_tasas_conversion('grupos_CO2','y_bin',type='bar')
graficar_tasas_conversion('Temperature','y_bin')

# Nueva columna en el dataset: "grupos_T"
dataset.loc[:,'grupos_T'] = "60-78"
dataset.loc[(dataset['Temperature']>78)&(dataset['Temperature']<=100),
'grupos_T'] = "79-100"
dataset.loc[(dataset['Temperature']>100)&(dataset['Temperature']<=120),
'grupos_T'] = "101-120"
dataset.loc[(dataset['Temperature']>120)&(dataset['Temperature']<=150),
'grupos_T'] = "121-150"
graficar_tasas_conversion('grupos_T','y_bin',type='bar')
#H2S
graficar_tasas_conversion('H2S','y_bin')

# Nueva columna en el dataset: "grupos_H2S"
dataset.loc[:,'grupos_H2S'] = "0-2000"
dataset.loc[(dataset['H2S']>2000)&(dataset['CO2']<=3000), 'grupos_H2S'] =
"2000-3000"
dataset.loc[(dataset['H2S']>3000)&(dataset['CO2']<=6000), 'grupos_H2S'] =
"3000-6000"
dataset.loc[(dataset['H2S']>6000)&(dataset['CO2']<=9000), 'grupos_H2S'] =
"6000-9000"
dataset.loc[(dataset['H2S']>9000)&(dataset['CO2']<=12000), 'grupos_H2S'] =
"9000-12000"
graficar_tasas_conversion('grupos_H2S','y_bin',type='bar')

#Phis
graficar_tasas_conversion('pH','y_bin',type="scatter")
# Nueva columna en el dataset: "grupos_CO2"
dataset.loc[:,'grupos_pH'] = "3-3.5"
```

```
dataset.loc[(dataset['pH']>3.5)&(dataset['pH']<=4), 'grupos_pH'] = "3.5-4"
dataset.loc[(dataset['pH']>4)&(dataset['pH']<=4.5), 'grupos_pH'] = "4-4.5"
dataset.loc[dataset['pH']>=4.5,'grupos_pH'] = ">4.5"
graficar_tasas_conversion('grupos_pH','y_bin',type='bar')
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from mlxtend.plotting import plot_sequential_feature_selection as plot_sfs
import numpy as np
from sklearn.metrics import accuracy_score
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split,GridSearchCV
from sklearn.preprocessing import StandardScaler
from sklearn.neighbors import KNeighborsClassifier
from mlxtend.feature_selection import SequentialFeatureSelector as SFS

X, y = dataset.iloc[:, 1:-1] ,dataset.iloc[:, -1]
X_train, X_test, y_train, y_test = train_test_split(
    X,
    y,
    test_size=0.3,
    random_state=42)

plt.figure(figsize=(20,10))
cor = X_train.corr()
sns.heatmap(cor, annot=True, cmap=plt.cm.CMRmap_r)
plt.show()

def correlation(dataset, threshold):
    col_corr = set()  # Set of all the names of correlated columns
```

```
    corr_matrix = dataset.corr()
    for i in range(len(corr_matrix.columns)):
        for j in range(i):
            if abs(corr_matrix.iloc[i, j]) > threshold: # we are
interested in absolute coeff value
                colname = corr_matrix.columns[i]  # getting the name of
column
                col_corr.add(colname)
    return col_corr
corr_features = correlation(X_train, 0.90)
len(set(corr_features))

corr_features
{'Al', 'Cl-', 'H2S', 'Na+', 'V'}

X_train.drop(corr_features,axis=1)
X_test.drop(corr_features,axis=1)

sc = StandardScaler()
X_train_std = sc.fit_transform(X_train)
X_test_std = sc.transform(X_test)

model = KNeighborsClassifier(n_neighbors=3)

model.fit(X_train_std, y_train)

print('Training accuracy:', np.mean(model.predict(X_train_std) ==
y_train)*100)
print('Test accuracy:', np.mean(model.predict(X_test_std) == y_test)*100)
```

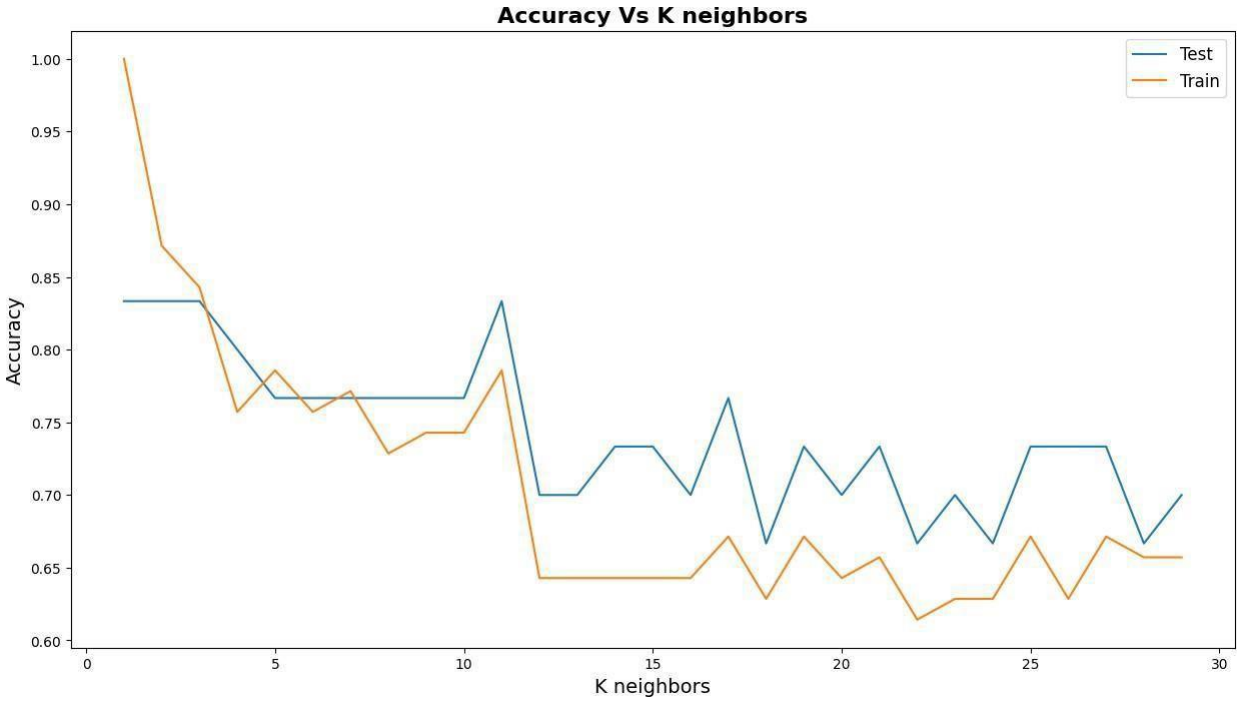
```
Training accuracy: 84.28571428571429
Test accuracy: 83.33333333333334


sfs1 = SFS(model,
           k_features=5,
           forward=True,
           floating=True,
           verbose=2,
           scoring='accuracy',
           n_jobs=-1,
           cv=5)

sfs1 = sfs1.fit(X_train_std, y_train)
[2023-08-26 00:41:40] Features: 5/5 -- score: 0.7857142857142858

dataset.columns[1:][list(sfs1.k_feature_idx_)].tolist()
['S', 'Cu', 'V', 'CO2', 'SO42-']

sfs1 = SFS(model,
           k_features="best", # or (1, 13) and then select by hand
           forward=True,
           floating=False,
           verbose=0,
           scoring='accuracy',
           cv=5)
sfs1 = sfs1.fit(X_train_std, y_train)
metric_dict = sfs1.get_metric_dict(confidence_interval=0.95)
fig1 = plot_sfs(metric_dict, kind='std_dev')
```

```
plt.title('Sequential Forward Selection (w. StdDev)')
plt.grid()
plt.show()
```




```
knn = KNeighborsClassifier()
decision_tree = DecisionTreeClassifier()


knn_params = {'n_neighbors': np.arange(1, 30)}
dt_params = {'max_depth': np.arange(2,21,1)}


def training_model(model, param_grid):
```

```
    model = GridSearchCV(model, param_grid=param_grid, scoring='accuracy')
    model.fit(X_train_std, y_train)
    return model

knn_model = training_model(knn, knn_params)
knn_model.best_params_

knn_model_final = knn_model.best_estimator_.fit(X_train_std, y_train)

y_pred_knn = knn_model_final.predict(X_test_std)
# Precisión del modelo
print(knn_model_final.score(X_train_std, y_train))
print(knn_model_final.score(X_test_std, y_test))

1.0
0.8333333333333334
```




```
dt_model = training_model(decision_tree, dt_params)
dt_model.best_params_

dt_model_final = dt_model.best_estimator_.fit(X_train_std, y_train)

y_pred_dt = dt_model_final.predict(X_test_std)

print(dt_model_final.score(X_train_std, y_train))
print(dt_model_final.score(X_test_std, y_test))

0.9857142857142858
0.7666666666666667
```

	Model	Accuracy
0	KNN	0.833333
1	Decision Tree	0.766667

```
def method_recomendado(model,C_value, Si_value, Mn_value, P_value,
S_value, Cr_value, Ni_value, Cu_value, Mo_value, Ti_value, Nb_value,
Al_value, V_value, Temperature_value, H2S_value, CO2_value,
Fluid_velocity_value, Salinity_value, Cl_value, HCO_value, Ca2_value,
Mg2_value, Na_value, SO42_value, pH_value, Time_value):

    predict = model.predict(np.array([C_value, Si_value, Mn_value,
P_value, S_value, Cr_value, Ni_value, Cu_value, Mo_value, Ti_value,
Nb_value, Al_value, V_value, Temperature_value, H2S_value, CO2_value,
Fluid_velocity_value, Salinity_value, Cl_value, HCO_value, Ca2_value,
Mg2_value, Na_value, SO42_value, pH_value, Time_value]).reshape(1, -1))
    return print(f'¿dicho caso presenta corrosión?: {predict[0]}')

prediction=method_recomendado(knn_model_final,0.061,0.28,1.86,0.011,0.0006
,0.03,0.03,0,0.22,0.016,0.061,0,0.059,30,55,180,5,5000,1000,100,200,150,30
00,100,5.5,0)
¿dicho caso presenta corrosión?: yes
```