



Universidad Nacional Mayor de San Marcos

Universidad del Perú. Decana de América

Facultad de Ingeniería Geológica, Minera, Metalúrgica y Geográfica

Escuela Profesional de Ingeniería de Minas

**Desarrollo de modelos predictivos de regresión en la
industria minera mediante el uso de algoritmo de
machine learning**

TESIS

Para optar el Título Profesional de Ingeniero de Minas

AUTOR

Oscar Adanver CASTILLO LIBERATO

ASESOR

Mg. David Ysaac MELGAR CABANA

Lima, Perú

2022



Reconocimiento - No Comercial - Compartir Igual - Sin restricciones adicionales

<https://creativecommons.org/licenses/by-nc-sa/4.0/>

Usted puede distribuir, remezclar, retocar, y crear a partir del documento original de modo no comercial, siempre y cuando se dé crédito al autor del documento y se licencien las nuevas creaciones bajo las mismas condiciones. No se permite aplicar términos legales o medidas tecnológicas que restrinjan legalmente a otros a hacer cualquier cosa que permita esta licencia.

Referencia bibliográfica

Castillo, O. (2022). *Desarrollo de modelos predictivos de regresión en la industria minera mediante el uso de algoritmo de machine learning*. [Tesis de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería Geológica, Minera, Metalúrgica y Geográfica, Escuela Profesional de Ingeniería de Minas]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios

Datos de autor	
Nombres y apellidos	Oscar Adanver Castillo Liberato
Tipo de documento de identidad	DNI
Número de documento de identidad	71058473
URL de ORCID	No aplica
Datos de asesor	
Nombres y apellidos	David Ysaac Melgar Cabana
Tipo de documento de identidad	DNI
Número de documento de identidad	41423051
URL de ORCID	https://orcid.org/0000-0001-6996-923x
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Jorge Enrique Soto Yen
Tipo de documento de identidad	DNI
Número de documento de identidad	03493146
Miembro del jurado 1	
Nombres y apellidos	Carlos Enrique Rodríguez Vigo
Tipo de documento de identidad	DNI
Número de documento de identidad	21555151
Miembro del jurado 2	
Nombres y apellidos	Mariano Pacheco Ortiz
Tipo de documento de identidad	DNI
Número de documento de identidad	06284228
Datos de investigación	
Línea de investigación	C.0.4.1. Desarrollo de modelos, simulación y optimización de procesos
Grupo de investigación	No aplica
Agencia de financiamiento	Sin Financiación
Ubicación geográfica de la investigación	País: Perú Departamento: LimaProvincia: Lima Latitud: - 12.0621065 Longitud: -77.0365256
Año o rango de años en que se realizó la investigación	2021-2022
URL de disciplinas OCDE	Minería, Procesamiento de minerales https://purl.org/pe-repo/ocde/ford#2.07.05



UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
(Universidad del Perú - Decana de América)
FACULTAD DE INGENIERÍA GEOLÓGICA, MINERA, METALÚRGICA Y GEOGRÁFICA
ESCUELA PROFESIONAL DE INGENIERÍA DE MINAS

Av. Colonial cdra. 53 – Ciudad Universitaria
Central Telefónica: 619-7000 anexos: 1110 - 1111
Lima 1 – Perú

ACTA DE SUSTENTACIÓN DE TESIS PARA OPTAR TÍTULO
PROFESIONAL DE INGENIERO DE MINAS

Vídeo conferencia Google Meet de la Escuela Profesional de Ingeniería de Minas de la Facultad de Ingeniería Geológica, Minera, Metalúrgica y Geográfica de la Universidad Nacional Mayor de San Marcos, el martes 17 de mayo del 2022, siendo las 17:30 horas, en presencia de los Señores Docentes designados como Miembros del Jurado.

Dr. JORGE ENRIQUE SOTO YEN

Presidente

Mg. CARLOS ENRIQUE RODRÍGUEZ VIGO

Miembro

Mg. MARIANO PACHECO ORTIZ

Miembro

Reunidos para el Acto Académico Público de la Sustentación de la TESIS de Don **OSCAR ADANVER CASTILLO LIBERATO**, Bachiller en Ingeniería de Minas, quien sustentó la Tesis Titulada: “**Desarrollo de modelos predictivos de regresión en la industria minera mediante el uso de algoritmo de machine learning**”, para la obtención del Título Profesional de Ingeniero de Minas.

Los miembros del Jurado Calificador, escuchada la sustentación respectiva, plantearon al graduando las observaciones pertinentes, que fueron absueltas a:

Aprobado, Bueno

El Jurado procedió a la calificación, cuyo resultado fue la nota de:

16 (Dieciséis)

Habiendo sido aprobada la Sustentación de la Tesis por el Jurado Calificador, el Miembro Presidente del Jurado, recomienda que la Facultad de Ingeniería Geológica, Minera, Metalúrgica y Geográfica, otorgue el **TÍTULO PROFESIONAL DE INGENIERO DE MINAS, a Don OSCAR ADANVER CASTILLO LIBERATO.**

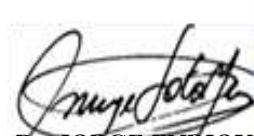
Siendo las 18:46 horas, se dio por concluido el acto académico, expidiéndose cinco (05) Actas Originales de la Sustentación de Tesis.

Ciudad Universitaria, 17 de mayo del 2022


Dr. JORGE ENRIQUE SOTO YEN
MIEMBRO PRESIDENTE


Mg. CARLOS ENRIQUE RODRÍGUEZ VIGO
MIEMBRO


Mg. MARIANO PACHECO ORTIZ
Miembro


Dr. JORGE ENRIQUE SOTO YEN
DIRECTOR
ESCUELA PROFESIONAL DE INGENIERÍA DE MINAS



RECOMENDACIONES

A large, empty rectangular box with an orange border, intended for writing recommendations.

Datos de la plataforma virtual institucional del acto de sustentación:

ID: <https://meet.google.com/fdx-hyrv-ktn>

Grabación archivada en:

https://drive.google.com/file/d/1dyMnGv0m_DI0tkdr4evUmpzRsRh_uppZ/view?usp=sharing

NOTA OBTENIDA:

PÚBLICO ASISTENTE: (Nombre, apellido y DNI)



UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

(Universidad del Perú - Decana de América)

FACULTAD DE INGENIERÍA GEOLÓGICA, MINERA, METALÚRGICA Y GEOGRÁFICA

ESCUELA PROFESIONAL DE INGENIERÍA DE MINAS

“Año del Fortalecimiento de la Soberanía Nacional”

INFORME DE EVALUACIÓN DE ORIGINALIDAD

DIRECTOR DE LA ESCUELA PROFESIONAL DE INGENIERÍA DE MINAS

Dr. Jorge Enrique Soto Yen

OPERADOR DEL PROGRAMA INFORMÁTICO DE SIMILITUDES

Tec. Alexander Iván Luyo Sánchez

DOCUMENTO EVALUADO:

Tesis para optar el Título Profesional de Ingeniero de Minas titulado: «DESARROLLO DE MODELOS PREDICTIVOS DE REGRESIÓN EN LA INDUSTRIA MINERA MEDIANTE EL USO DE ALGORITMOS DE MACHINE LEARNING»

AUTOR DEL DOCUMENTO:

Bach. OSCAR ADANVER CASTILLO LIBERATO

FECHA DE RECEPCIÓN DEL DOCUMENTO:

31/03/2022

FECHA DE APLICACIÓN DEL PROGRAMA INFORMÁTICO DE SIMILITUDES:

01/03/2022

SOFTWARE UTILIZADO

Turnitin

CONFIGURACIÓN DEL PROGRAMA DETECTOR DE SIMILITUDES

- Excluye textos entrecomillados
- Excluye fuentes para buscar similitud
- Excluye Bibliografía
- Excluye cadenas menores a 40 palabras

PORCENTAJE DE SIMILITUDES SEGÚN PROGRAMA DETECTOR DE SIMILITUDES

Ocho Por ciento (8 %)

FUENTES ORIGINALES DE LAS SIMILITUDES ENCONTRADAS

- | | |
|--|-----|
| ▪ www.cienciadedatos.net | <1% |
| ▪ aprendeia.com | <1% |
| ▪ code.i-harness.com | <1% |
| ▪ geekflare.com | <1% |
| ▪ Jesús Alberto Torres Guerra, Denis Mejía Cáceres, Patrick Moreyra Ramos, Jairo Oré Grados, Santos Oscoco Barrientos. "Geometalurgia y el futuro de la minería digital en el Perú", Revista | <1% |



UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

(Universidad del Perú - Decana de América)

FACULTAD DE INGENIERÍA GEOLÓGICA, MINERA, METALÚRGICA Y GEOGRÁFICA

ESCUELA PROFESIONAL DE INGENIERÍA DE MINAS

“Año del Fortalecimiento de la Soberanía Nacional”

del Instituto de investigación de la Facultad de minas, metalurgia y ciencias geográficas, 2021

• topbigdata.es	<1%
• repositorio.upsin.edu.mx	<1%
• idoc.pub	<1%
• es.wikipedia.org	<1%
• www.coursehero.com	<1%
• www.ibm.com	<1%
• tesis.ucsm.edu.pe	<1%
• docplayer.es	<1%
• 3lib.net	<1%
• repositorio.unsa.edu.pe	<1%
• Submitted to Pontificia Universidad Católica del Peru	<1%
• ichi.pro	<1%
• repositorio.una.ac.cr	<1%
• www.dspace.uce.edu.ec	<1%
• idus.us.es	<1%
• repositorioacademico.upc.edu.pe	<1%
• es.scribd.com	<1%
• bibdigital.epn.edu.ec	<1%
• pythondiario.com	<1%
• cybertesis.unmsm.edu.pe	<1%
• nanopdf.com	<1%
• Submitted to Pontificia Universidad Católica Madre y Maestra PUCMM	<1%
• empiezoinformatica.wordpress.com	<1%
• manualestutor.com	<1%
• bibliotecadigital.udea.edu.co	<1%

OBSERVACIONES

Ninguna

CALIFICACIÓN DE ORIGINALIDAD

Documento cumple criterios de originalidad, sin observaciones

04/04/2022



UNMSM

Firmado digitalmente por SOTO YEN
Jorge Enrique FAU 20148092282 soft
Motivo: Soy el autor del documento
Fecha: 05.04.2022 09:44:50 -05:00

DR. JORGE ENRIQUE SOTO YEN
DIRECTOR

ESCUELA PROFESIONAL DE INGENIERÍA DE MINAS

Dedicatoria

A mis padres Luis y Rafaela, y a mi hermano Erick

Agradecimientos

A mis hermanas, hermanos y a toda mi familia, gracias por su apoyo y comprensión.

A los ingenieros Brayan Huaranga Yantas y Eduardo Vásquez Alfaro, por haber ayudado en la consecución de la data para el desarrollo de esta investigación.

A los profesores Carlos Enrique Rodríguez Vigo, Mariano Pacheco Ortiz, Jorge Enrique Soto Yen y mi asesor David Ysaac Melgar Cabana, por sus recomendaciones y sugerencias para conseguir la mejor versión de esta investigación.

Resumen

Una de las características que las empresas mineras siempre buscan minimizar es el alto riesgo que implica operar un proyecto. Estos pueden estar influenciados por factores geológicos, ambientales, sociales, de mercado, operativos, políticos, etc. Para afrontar estos riesgos y aumentar sus beneficios, muchas empresas han iniciado la transformación digital implementando tecnologías basadas en Inteligencia Artificial. Los resultados obtenidos han mejorado significativamente la productividad, el consumo energético, reducido los niveles de contaminación y optimizado la seguridad de sus trabajadores.

La naturaleza propia de cada operación minera y el conocimiento limitado de estos temas en muchas operaciones actuales del Perú, nos incentivó a realizar modelos predictivos de regresión en casos de la industria minera mediante el uso de algoritmos de Machine Learning, un sub campo de la inteligencia artificial, utilizando un flujograma más didáctico e intuitivo.

El estudio propone soluciones en tres casos de la industria minera a través de modelos predictivos de regresión. Estos modelos permiten estimar con un alto grado de certeza: el precio de oro, el contenido de sílice en el concentrado de hierro y el consumo de combustible en camiones mineros de gran tonelaje. Este estudio es de enfoque cuantitativo y del tipo exploratorio, lo que permitió desarrollar estos modelos usando diferentes algoritmos de machine learning a partir de datos históricos.

Según las métricas de evaluación, los modelos que tuvieron un buen desempeño para predecir el precio de oro y el consumo de combustible se desarrollaron usando los algoritmos **SVR** y **Árbol de regresión**. Sin embargo, para predecir el contenido de sílice en el concentrado de hierro fue necesario el uso del algoritmo de ensamble **Random Forest**, principalmente por la complejidad del caso. Finalmente, es importante mencionar que para llegar a los resultados esperados hace falta un trabajo estructurado que implica conocimiento de ciencia de datos, conocimiento de minería, y un entendimiento claro del problema que se resuelve usando algoritmos de machine learning.

Índice

Capítulo I: Introducción.....	9
1.1.Situación problemática	10
1.2.Formulación de problema.....	11
1.2.1.General.....	11
1.2.2.Específico	11
1.3.Justificación.....	11
1.3.1.Importancia del tema.....	11
1.3.2.Viabilidad	12
1.3.3.Originalidad.....	12
1.3.4.Alcance y limitaciones	12
1.4.Objetivos.....	13
1.4.1.General.....	13
1.4.2.Específico	13
Capítulo II: Marco Teórico.....	14
2.1.Antecedentes del tema.....	14
2.2.Bases teóricas de Inteligencia Artificial y Machine Learning	18
2.2.1.Inteligencia Artificial	18
2.2.2.Machine Learning	19
2.2.3.Deep Learning	19
2.3.Técnicas y algoritmos de Machine Learning	20
2.3.1.Aprendizaje supervisado	21
2.3.2.Aprendizaje no supervisado	21
2.3.3.Aprendizaje reforzado	22
2.4.Minería de datos	24
2.4.1.Etapas de minería de datos	24
2.4.2.Técnica de minería de datos	26
2.5.Industria Minera	27
2.5.1.Proyecto minero	28

2.5.2.Etapas del proceso productivo de una mina	30
2.5.3.Extracción.....	31
2.5.4.Procesamiento	32
2.5.5.Fundición y Refinería.....	33
2.6.Inteligencia Artificial en minería	34
2.6.1.Como adoptar Inteligencia Artificial (IA) en minería.....	36
2.6.2.Casos de aplicación de Inteligencia Artificial en minería.	37
Capítulo III: Marco metodológico de la Investigación.....	41
3.1.Metodología de trabajo.....	41
3.2.Tipo de investigación	42
3.3.Diseño de proceso de investigación	42
3.4.Población y datos de objeto de estudio	43
3.5.Método y técnica de recolección de datos.....	43
Capítulo IV: Flujograma para generar modelos predictivos de regresión con el uso de algoritmos de machine learning.....	45
4.1.Comprensión del problema	45
4.2.Desarrollo del flujograma de trabajo.....	45
4.2.1.Categorización del problema.....	47
4.2.2.Determinación de métricas.....	49
4.2.3.Recolección de datos.....	51
4.2.4.Análisis de datos.....	51
4.2.5.Elección del algoritmo	53
4.2.6.Modelado.....	54
4.2.7.Evaluación del modelo	55
4.2.8.Predicción.....	58
4.3.Herramientas empleadas.....	59
4.3.1.Lenguaje de programación	59
4.3.2.Librerías.....	59
4.3.3.Programas y Software	62
Capítulo V: Desarrollo de modelos predictivos de los casos de aplicación.....	64
5.1.Modelo para la predicción del precio de oro.....	64

5.1.1.Compresión del problema	64
5.1.2.Aplicación del flujograma para el desarrollo del modelo	65
5.1.3.Análisis de resultados	92
5.2.Modelo predictivo para estimar sílice en el concentrado de hierro.....	95
5.2.1.Compresión del problema	95
5.2.2.Aplicación del flujograma para el desarrollo del modelo	96
5.2.3.Análisis de resultados	110
5.3.Modelo predictivo para estimar el consumo de combustible en camiones mineros de alto tonelaje.....	111
5.3.1.Compresión del problema	111
5.3.2.Aplicación del flujograma para el desarrollo del modelo	112
5.3.3.Análisis de resultados	121
Capítulo VI: Conclusiones.....	124
Capítulo VII: Recomendaciones.....	126
Bibliografía.....	128
Anexo.....	131

Lista de Figuras

Figura 1. Inteligencia Artificial y las Sub Áreas.	19
Figura 2. Funcionamiento de las redes neuronales	20
Figura 3. Aprendizaje Supervisado y no Supervisado.	22
Figura 4. Aprendizaje Reforzado.	23
Figura 5: Esquema CRISP-DM.....	25
Figura 6: Esquema para minería de datos de Microsoft.....	26
Figura 7: Cadena de valor de la minería	27
Figura 8: Etapas y componentes de un Proyecto Minero.....	29
Figura 9: Ciclo de vida de una Mina.	30
Figura 10: Procesos de una mina a tajo abierto.....	31
Figura 11: Procesos de una Mina Subterránea.	32
Figura 12: Procesos de Flotación.	33
Figura 13: Procesos de Lixiviación.....	33
Figura 14. Inteligencia Artificial en Minería.	35
Figura 15: Esquema de metodología de trabajo.	41
Figura 16. Flujograma para el desarrollo de modelos predictivos de regresión.	46
Figura 17. Esquema de división de datos.....	53
Figura 18. Elección del Algoritmo según datos de salida	54
Figura 19. Esquema para el proceso de modelado.	55
Figura 20. Diferentes modelos ajustados a un mismo dato	56
Figura 21. Comportamiento del error frente al Overfit	56
Figura 22. Proceso para realizar validación cruzada.	57
Figura 23. Diagrama para el ajuste de hiperparámetros.....	58
Figura 24. Variación del precio de Oro 2015-2021.	65
Figura 25. Correlación de Variables del Precio de Oro.	69

Figura 26. Histograma y gráfica de dispersión, Precio de Au Vs S&P 500.....	70
Figura 27. Diagrama de cajas de las variables principales.....	71
Figura 28. Esquema de división de datos para el Caso 1.	72
Figura 29. Ilustración Modelo Regresión Lineal	74
Figura 30. Ilustración del Modelo SVR	75
Figura 31. Ilustración del Modelo Árbol de Regresión.....	77
Figura 32: Partes de un diagrama de árbol de regresión.	78
Figura 33. Diagrama del Modelo Árbol de regresión	79
Figura 34. Modelo Multilineal para la predicción del precio de Oro.....	80
Figura 35. Esquema de validación cruzada.....	81
Figura 36. Modelo SVR para la predicción del precio de Oro.....	83
Figura 37. puntajes CV para diferentes ccp_alpha.....	87
Figura 38. Modelo inicial Árbol de Regresión para la predicción del precio Oro.	88
Figura 39. Diagrama del Árbol de Regresión del modelo inicial.....	89
Figura 40. Comparación de modelos para la predicción del precio de Oro.	91
Figura 41. Diagrama de flujo para la obtención del concentrado de Hierro	96
Figura 42. Variación del Concentrado de Sílice	98
Figura 43. Histograma del Concentrado Sílice	99
Figura 44. Ilustración del Modelo Random Forest	101
Figura 45. Ilustración del Modelo Gradient Boosting.....	102
Figura 46. Ilustración del Modelo Votin Regressor	103
Figura 47. Modelo SVR para la predicción de Concentrado de Sílice	104
Figura 48. Modelo Random Forest para la predicción de Concentrado de Sílice.....	106
Figura 49. Ajuste de los datos con Algoritmo Normal SVR.....	107
Figura 50. Ajuste de los datos con Algoritmo de Ensamble GB.....	108
Figura 51. Comparación de modelo para la predicción de Concentrado de Sílice	109
Figura 52. Camión Minero de gran tonelaje CAT 797E.....	112
Figura 53. Variación de consumo de combustible diario para del camión F100	115

Figura 54. Consumo promedio por hora de los camiones.....	116
Figura 55. Histograma del consumo diario de combustible de los camiones	117
Figura 56. Modelo SVR para la predicción del consumo de combustible.....	119
Figura 57. Modelo Random Forest para la predicción del consumo de combustible	119
Figura 58. Comparación de modelos para la predicción del consumo de combustible ...	120
Figura 59. Ajuste de los datos con el modelo GB	122
Figura 60. Ajuste de los datos con el modelo RF.....	123
Figura 61. Ajuste de los datos con el modelo SVR.....	123

Capítulo I: Introducción

Entre los desafíos que enfrentan las empresas mineras que desean implementar tecnologías basadas en Inteligencia Artificial están la cultura resistente y la falta de talento. Esto puede deberse principalmente a la falta de conocimiento sobre estos temas, la poca capacidad para tener datos estructurados o la poca predisposición e interés en las empresas mineras para invertir en esta nueva tecnología.

En los últimos años, muchas empresas mineras de alto capital han invertido en la investigación e implementación de tecnologías basadas en Inteligencia Artificial. Sin embargo, muchas operaciones actuales, sobre todo de la pequeña y mediana minería, tiene escaso conocimiento y han tenido dificultades técnicas y económicas para poder aprovechar estas ventajas competitivas. Esto demuestra que existe todavía una amplia oportunidad que pueden ser aprovechadas.

Este trabajo tiene como finalidad proporcionar de manera didáctica la forma en que se debe abordar las tecnologías basadas en machine learning (un sub campo de la inteligencia artificial) en la industria minera. Para este fin, desarrollaremos modelos predictivos de regresión usando algoritmos de machine learning aplicados a casos de la industria minera.

Los principales puntos que se desarrollarán en esta investigación son los siguientes.

1. Revisión bibliográfica.
2. Definición de concepto generales de Inteligencia Artificial (AI), Machine Learning (ML), minería de datos e industria minera.
3. Definición de población y datos para el desarrollo de los modelos.
4. Comprensión del problema para aplicar machine learning en minería.
5. Desarrollo del flujograma para generar modelos predictivos.
6. Descripción de algunas herramientas usadas.
7. Desarrollo de modelo predictivos de los casos.
8. Análisis de resultados de los ejemplos desarrollados.
9. Conclusiones.

1.1. Situación problemática

La naturaleza de la actividad minera, que actualmente está orientada a una producción a gran escala, hace necesario que las empresas busquen continuamente nuevas formas de resolver sus problemas para poder tener mayor eficiencia en todos sus procesos. Un pequeño cambio; ya sea la velocidad, rendimiento, consumo de energía o eficiencia, puede permitir la rentabilidad de una operación. Esta inquietud hizo que muchas empresas empiecen a adoptar tecnologías basadas en Inteligencia artificial, dado que tienen ventajas considerables respecto procesos tradicionales.

Los beneficios obtenidos por las empresas que ya están usando esta tecnología han tenido retornos alentadores. Sin embargo, su uso es relativamente nuevo, ya que aún no han evaluado claramente cuál será el impacto para facilitar este cambio de paradigma. Este estudio busca motivar la investigación para encontrar nuevas aplicaciones de herramientas tecnológicas en los diferentes procesos.

Con el uso de la inteligencia artificial se logra que las máquinas se comporten y “piensen” como los humanos. Además, los desafíos que hay en la minería, es que junto con la capacidad de generar volúmenes importantes de datos; de su uso, evaluación, y análisis depende el alcanzar progresivamente importantes niveles de eficiencia que favorecerá la aplicación de estas herramientas tecnológicas en la industria minera.

Uno de los pasos importantes para implementar Inteligencia Artificial, cuando se usa la técnica de aprendizaje supervisado, es la generación de modelos predictivos empleando los diferentes algoritmos de machine learning. Si bien es cierto que en muchas operaciones mineras actuales existe información del desarrollo y aplicación de modelos predictivos basados en machine learning, estos generalmente tienen información poco detallada o información muy general. Lo que puede causar poco entendimiento en los profesionales de la industria minera que tienen poco conocimiento en estos temas. Atendiendo a esta problemática, en esta investigación desarrollaremos modelos predictivos de regresión de tres casos de la industria minera a través de un esquema más didáctico e intuitivo, además, utilizaremos diferentes algoritmos de machine learning con el fin de conocer las bondades de cada uno de ellos al resolver problemas de minería.

1.2. Formulación de problema

1.2.1. General

¿De qué manera se desarrolla modelos predictivos de regresión en casos de la industria minera mediante el uso de algoritmos de machine learning?

1.2.2. Específico

¿De qué manera se desarrolla un flujograma para generar modelos predictivos de regresión en casos de la industria minera mediante el uso de algoritmos de machine learning?

¿De qué manera se desarrolla un modelo predictivo para estimar el precio del oro a partir de información histórica mediante el uso de algoritmos de machine learning?

¿De qué manera se desarrolla un modelo predictivo para estimar el contenido de sílice en el concentrado de mineral de hierro a partir de información histórica mediante el uso de algoritmos de machine learning?

¿De qué manera se desarrolla un modelo predictivo para estimar el consumo de combustible de camiones mineros de gran tonelaje a partir de información histórica mediante el uso de algoritmos de machine learning?

1.3. Justificación

1.3.1. Importancia del tema

La importancia del tema radica en aplicar algoritmos de machine learning en un campo bastante nuevo como es la industria minera. La aplicación de estas herramientas es sin duda muy relevantes para la supervivencia de las operaciones mineras a futuro. Además, en el corto plazo ayudará a las empresas a reducir sus costos significativamente como ya lo están haciendo en algunos casos excepcionales. En el ámbito académico despertará el interés de los estudiantes e interesados para que puedan seguir investigando, y puedan ver cómo aplicar estas herramientas en las diferentes áreas de un proyecto minero.

1.3.2. Viabilidad

La presente investigación se realizó utilizando información de diferentes fuentes confiables: opiniones y conocimiento de egresados e ingenieros de minas, consultoras, profesores, informes, trabajos sobre el tema. Además, la investigación se realizó mayormente utilizando herramientas informáticas y trabajo de gabinete, lo que contribuyó la viabilidad del estudio.

1.3.3. Originalidad

Si bien es cierto hay muchas empresas de capitales elevados que están empezando aplicar estas herramientas. Poco se ha difundido sobre la manera en que la usan; además, las empresas mineras medianas y pequeñas tienen poco o ningún conocimiento de estos temas.

1.3.4. Alcance y limitaciones

Mediante la presente investigación nos enfocaremos en tres procesos particulares de la industria minera, las cuáles contempla:

- Explotación – para el cual plantearemos una solución para evitar el consumo excesivo de combustible en camiones mineros en el sub proceso de transporte.
- Beneficio – para el cual plantearemos cómo obtener un concentrado de hierro con un contenido mínimo de sílice en el sub proceso de flotación.
- Comercialización – en este caso desarrollaremos un modelo para la estimación del precio de oro.

Las soluciones planteadas, sin duda, pueden mejorar los procesos mencionados a través de una implementación. Sin embargo, no se pueden definir con exactitud en cuanto puede ser la mejora de todo el proceso o el negocio minero en general, ya que estos contienen otras actividades unitarias que pueden tener incluso más importancia en el desempeño de cada proceso o del negocio minero en general.

Por otro lado, cuando se desarrolla aplicaciones basadas en machine learning podemos encontrar diferentes técnicas y algoritmos para su uso. Pero, también se debe tomar en consideración algunos aspectos fundamentales como: capacidad computacional, acceso a los datos, tiempo de procesamiento y objetivos. Además, es casi imposible tratar todos las técnicas y algoritmos de machine learning disponibles en un solo documento.

Inicialmente, se planteó desarrollar ejemplos de modelos predictivos usando la técnica de aprendizaje supervisado tanto para realizar tareas de regresión como para realizar tareas de clasificación, pero no pudimos encontrar datos para realizar problemas de clasificación, por lo que finalmente, solo desarrollamos modelos predictivos de regresión.

1.4. Objetivos

1.4.1. General

Desarrollar modelos predictivos de regresión en casos de la industria minera mediante el uso de algoritmos de Machine Learning.

1.4.2. Específico

- Desarrollar un flujograma para generar modelos predictivos de regresión en casos de la industria minera usando algoritmos de machine learning.
- Desarrollar un modelo predictivo para estimar el precio de oro a partir de información histórica mediante el uso de algoritmos de machine learning.
- Desarrollar un modelo predictivo para la estimar de contenido de sílice en el concentrado de mineral de hierro a partir de información histórica mediante el uso de algoritmos machine learning.
- Desarrollar un modelo predictivo para estimar el consumo de combustible de camiones mineros de gran tonelaje a partir de datos históricos mediante el uso de algoritmos de machine learning.

Capítulo II: Marco Teórico

2.1. Antecedentes del tema

a. Internacionales

Reveco (2019) realizó una investigación sobre el análisis predictivo de activos mineros para la obtención de intervalo de falla mediante algoritmos de machine learning en una mina de cobre y molibdeno a tajo abierto en Chile. El objetivo de este estudio fue desarrollar un modelo predictivo mediante algoritmos de machine learning para determinar el intervalo de falla de componentes principales de la gran minería. La metodología de trabajo usada se basa en la revisión bibliográfica, obtención de fechas en las que los motores hayan sido cambiadas, extracción de datos de análisis de aceite, etiquetado para cada modelo y preprocesamiento de datos, creación de modelos predictivos usando el programa AzureML, obtención de métricas de rendimiento, y análisis de resultados y conclusiones. Para la extracción de datos, primero obtuvo las fechas en las que los motores de los camiones fueron cambiadas por presentar alguna condición de falla, en el periodo comprendido entre enero del 2015 y diciembre de 2018, obteniendo un total de 44 fechas de cambio, además, con las fechas en la que un motor era cambiado de un camión, obtuvo la fecha de instalación. Teniendo la fecha de instalación y la fecha de cambio de los motores que presentaron alguna condición de falla a la fecha de cambio, obtuvo las muestras para realizar el análisis de aceite para finalmente, obtener una base de datos con las variables de las muestras.

En esta investigación, Reveco describe la importancia de hacer un análisis de la estructura y funcionamiento de motores Diesel de camiones mineros; extraer, procesar y etiquetar datos de muestras de análisis de aceite; crear modelos predictivos con algoritmos de machine Learning con estos datos extraídos; y a través de estos resultados hacer una retroalimentación de la estrategia, táctica y planificación de mantenimiento predictivo de la mina los bronce. Los resultados del desarrollo de esta investigación demostraron que a través de los datos de análisis de aceite se obtuvieron tres modelos predictivos con un desempeño satisfactorio, estos modelos lograron en conjunto generar una predicción global de comportamiento de motores Diesel de camiones mineros. Finalmente, concluye que 2 de los modelos, utilizando el método de Covarianza Robusta

(6%) para detectar anomalía y Decision Forest Regression para la predicción de tiempo de vida remanente, fueron incorporados exitosamente a la estrategia de área sintomática de motores Diesel de camiones mineros.

De igual manera Calisto (2018) realizó una investigación para un proyecto en Chile. Este estudio tiene como objetivo desarrollar un modelo predictivo en base a minería de datos y machine learning de un sistema de alerta temprana de ordenes de trabajo riesgosas, que permitiera articular de una menor forma la cadena de abastecimiento, planificación y ejecución, con la finalidad de mejorar la disponibilidad de materiales para mantenimiento y el nivel de servicio de Supply. La metodología de trabajo usada para el desarrollo de esta investigación se basa en el esquema de minería de datos CRISP-DM (Cross Industry Standard Process for Data Mining), el cual contempla el entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelamiento, evaluación y despliegue. Para el proceso de recolección de datos utilizó 4 bases de datos disponibles para finalmente consolidar la información en un Data Warehouse (se refiere a un sistema que combina, agrega y une informaciones provenientes de diferentes fuentes en un almacén de datos únicos y centralizado). Las bases de datos que utilizó fueron: WOPIFOT (mide el número de reservas planeadas) con un volumen de 18928 datos y 13 variables, MIFOT (ratio entre la cantidad de materiales entregados y materiales requeridos) con un volumen de 330553 datos y 5 variables, ONIQUA (software de inventario) con un total de 340080 datos y 7 variables, y Reporte RCA (análisis de causas raíces) con 2090 datos y 147 variables. Los resultados obtenidos describen que el modelo predictivo realizado con el algoritmo Gradient Boosting Tress, obtuvo el mejor resultado tanto en accuracy como performance general en cuanto a tiempo, interpretabilidad de los resultados y recursos computacionales. Finalmente, concluye que el modelo mencionado tiene una precisión de 82.2%, permitiendo tener un modelo predictivo confiable y además automatizar el análisis de las órdenes de trabajo, ayudando a los planificadores en enfocar su trabajo en órdenes de trabajos riesgosas y así optimizar su tiempo.

Rodríguez, Sánchez y Chica (2015) publicaron un artículo cuyo principal objetivo fue probar las capacidades de cuatro algoritmos de regresión de Machine Learning (Redes neuronales artificiales (ANN), Árbol de regresión (RT), Bosques aleatorios (RF) y las Máquinas de vectores de soporte (SMV)) para desarrollar un modelo predictivo que estime el potencial de depósito de mineral de oro a partir de información geológica, geoquímica, geofísica y derivada de EO-1. Esta

investigación realiza la integración de información geológica disponible en un distrito geológico, dentro de una geodatabase. Esta base de datos está constituida por 46 lugares de ocurrencia de oro, correspondiente a depósitos explotados y estructuras mineralizadas conocidas y datos físico-químicos. Los resultados de la aplicación de estos diferentes algoritmos al mapeo epitermal de la prospectividad de Au, indican que el algoritmo de RF superó a los otros algoritmos (ANNs, RTs, SVMs). El algoritmo de RF mostró una mayor estabilidad y robustez con diferentes parámetros de entrenamiento y mejores tasas de éxito y resultados de análisis. Por otro lado, concluyen que todos los algoritmos Machine learning son potenciales para utilizar cuando las evidencias de depósito de mineral son escasas. Además, identificaron que los parámetros del modelo de RF y RT se pueden interpretar para obtener información sobre los controles geológicos de la mineralización.

b. Nacionales

Penalillo (2021) realizó una tesis para estimar la sobre dilución en el método de explotación Sub Level Stoping de una operación minera. Este estudio tiene como objetivo la aplicación de algunas técnicas de aprendizaje automático para desarrollar modelos predictivos de regresión con el fin de mejorar los resultados de estimación de sobre dilución en tajos explotados por el método Sub Level Stoping. La metodología usada para el desarrollo de este trabajo se basa en una investigación con enfoque cuantitativo de tipo predictivo y tecnológico. En donde se describe también, la metodología de trabajo que utilizó para lograr estos objetivos como las siguientes: definir las variables operativas que intervienen en la investigación, recopilar la información definidas, aplicar diferentes técnicas de machine learning, validar y comparar los resultados con las métricas definidas. La recolección de datos se obtuvo de los reportes de operaciones, plan mensual y los informes de reconciliaciones, además, menciona que las herramientas usadas para el desarrollo de esta etapa se basan en el uso de Hojas de Excel y Software minero. Los resultados obtenidos con la aplicación de técnicas de machine learning permitieron la mejora de hasta 18% respecto a procesos tradicionales. Según menciona en el documento, estos permitieron la reducción de costo por el trabajo adicional por carguío y transporte de carga de dilución.

Torres, Mejía, Moreyra, Oré y Oscco (2021) realizaron un artículo científico donde abordan los avances tecnológicos utilizadas en la industria mineras. Así mismo, se presenta la situación actual de la Geometalurgia en las minas peruanas. La metodología que usaron para esta

investigación es no experimental del tipo descriptiva, ya que se describe la importancia del uso de la tecnología 4.0 en la Geometalurgia para mejorar los procesos operativos y la mejora de la rentabilidad de las operaciones mineras. Además, la realización de este estudio se basa en la recolección de casos de empresas mineras en el Perú, que tuvieron éxito con la implementación de nuevas tecnologías, y se detalla algunas otras en el mundo. En esta investigación hacen énfasis de la importancia de adoptar sistemas basadas en machine learning e inteligencia artificial para adoptar la automatización y digitalización en todos los procesos de la industria peruana. Finalmente, concluyen que, a través de la Geometalurgia, que abarca desde la explotación hasta las operaciones, estas innovaciones incidan significativamente en la productividad, seguridad y la capacidad de procesamiento de la información, para resolución de problemas con decisiones más acertadas que brinden valor económico al negocio minero.

Carrión (2018) presenta una aplicación de Inteligencia Artificial como herramienta para la solución de problemas en la planta concentradora mediante predicción de variables para una eficiente recuperación. La metodología de investigación es de enfoque cuantitativo, además, el trabajo para realizar esta investigación se basa en el desarrollo un modelo predictivo con el uso de redes neuronales, donde se introduce gran cantidad de datos reales anteriores de Planta Concentradora en este caso leyes de Estaño y Azufre del mineral Proveniente de mina y sus respectivas recuperaciones, con los cuales busca que la red Neuronal Aprenda para predecir la recuperación de la planta concentradora. El autor concluye que, de acuerdo a los resultados obtenidos, el modelo desarrollado con redes neuronales realiza predicciones con alto grado de exactitud para las recuperaciones futuras ayudando a tomar decisiones adecuadas en el proceso.

Si bien es cierto que las investigaciones descritas mencionan aplicaciones o desarrollo de tecnologías basadas en modelos predictivos de machine learning en casos de la minería, estas presentan una metodología muy general para el desarrollo de sus modelos, como es el caso de la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) para minería de datos. Otras investigaciones se basan en métodos descriptivos que, a través de la revisión bibliográfica o el estudio de casos de aplicación, llegan a la conclusión de alguna implementación exitosa de tecnologías basadas en modelos predictivos con machine learning en la industria minera. En esta investigación desarrollaremos modelos predictivos con un flujograma más didáctico e intuitivo

para un mejor entendimiento. Lo cual incluye la descripción de cada paso de flujograma para cada modelo desarrollado, además, la descripción de algunas herramientas usadas.

2.2. Bases teóricas de Inteligencia Artificial y Machine Learning

2.2.1. Inteligencia Artificial

Desde ya hace bastante tiempo hemos venido escuchando el término de inteligencia artificial y sus aplicaciones en los diferentes campos de la ciencia, pero ¿cómo surgió este campo de estudio y cómo lo aplicamos y lo definimos en la actualidad?

Turing, A. (1950) propició la discusión sobre la inteligencia artificial, donde pregunta: “¿puede pensar una máquina?”. McCarthy, J. (2004) décadas después, planteó una definición de inteligencia artificial donde la señala como: “Es la ciencia y la ingeniería para fabricar máquinas inteligentes, especialmente programas informáticos inteligentes, está relacionada con la tarea similar de usar computadoras para entender la inteligencia humana”.

Rouse, M. (2020) define la inteligencia artificial como: “es la simulación de proceso de inteligencia humana por parte de las máquinas, especialmente sistemas informáticos”. Esto incluyen procesos de aprendizaje, razonamiento y autocorrección. En palabras más simples, la inteligencia artificial es un software que es capaz de ejecutar capacidades humanas para resolver tareas simples hasta complejas. En la actualidad se está investigando para que Inteligencia Artificial puedan entender y reaccionar con su entorno a niveles más profundos.

Machine Learning vs Deep Learning

Tanto machine learning como Deep Learning son subcampos de la inteligencia artificial, a la vez que Deep Learning es un subcampo de Machine Learning. La utilización de ambos se ha ido aplicando en diferentes industrias, donde la principal diferencia es la forma en la que aprende cada algoritmo. En los siguientes puntos ampliaremos la definición con más detalle para cada una de ellas.

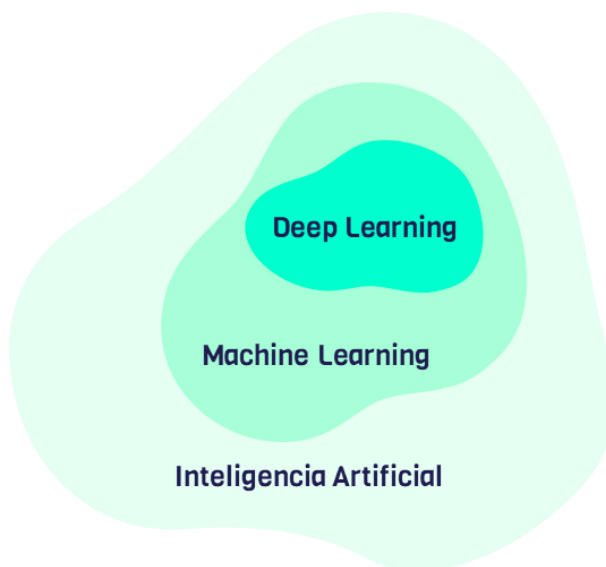


Figura 1. Inteligencia Artificial y las Sub Áreas.

Fuente: Elaboración propia

2.2.2. Machine Learning

El machine learning, es un subcampo de la inteligencia artificial, donde las personas intervienen para poder entrenarlos e identificar patrones basados en datos y hacer predicciones. También se puede definir como un conjunto de técnicas mediante las que un algoritmo que tiene que realizar una tarea es capaz de modificar su comportamiento. Esta técnica se basa en datos como resultados de acciones pasadas o en lo que le digan esos resultados.

2.2.3. Deep Learning

En Deep Learning, un subcampo de machine learning, los algoritmos se componen de capas de redes neuronales, a través de las cuales las máquinas son capaces de aprender predecir y tomar sus propias decisiones. La importancia de Deep learning radica en automatizar la extracción de características sin la intervención humana. Además, estos algoritmos soportan grandes volúmenes de datos.

El deep machine learning puede utilizar los conjuntos de datos etiquetados, también conocidos como aprendizaje supervisado, para informar a su algoritmo, pero no requiere necesariamente un conjunto de datos etiquetados. Puede ingerir datos no estructurados en su forma original (como por ejemplo texto o imágenes) y puede determinar automáticamente la jerarquía de características que distinguen diferentes categorías de datos. A diferencia del machine learning, no requiere intervención humana para procesar datos, lo que permite escalarlo de maneras más interesantes. (IBM.Cloud.Education, 2020)

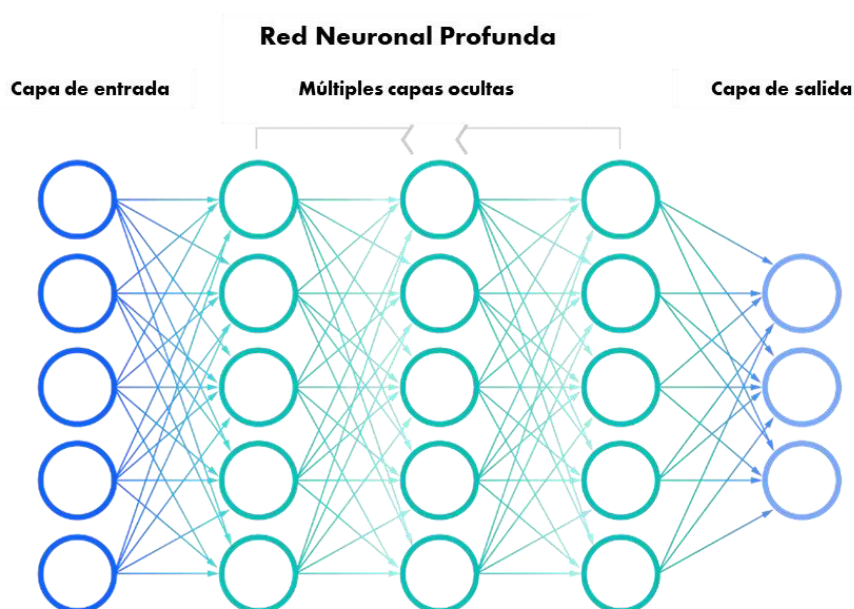


Figura 2. Funcionamiento de las redes neuronales

Fuente: Adaptado de Inteligencia Artificial, por IBM Cloud Education, 2020, <https://www.ibm.com/pe-es/cloud/learn/what-is-artificial-intelligence>

2.3. Técnicas y algoritmos de Machine Learning

Machine Learning ha crecido en popularidad debido a la gran cantidad de problemas provenientes de diferentes áreas del conocimiento para las que se han desarrollado soluciones basadas en las técnicas que componen esta sub rama de la inteligencia artificial.

En general, un problema de aprendizaje automático considera un conjunto de n muestras de datos y luego intenta predecir los datos desconocidos. Si cada muestra tiene las de un número y ,

por ejemplo, una entrada multidimensional, se dice que tiene varios atributos o características. En este apartado describiremos las ideas que hay detrás de las principales técnicas de Machine learning.

2.3.1. Aprendizaje supervisado

En el aprendizaje supervisado los datos vienen con etiquetas adicionales que queremos predecir. Estos pueden ser:

Regresión: si lo que queremos predecir consta de una o más variables continuas. Un ejemplo de un problema de regresión sería la predicción de la longitud de un tiburón en función de su edad y peso.

Clasificación: en este tipo de problema las muestras pertenecen a dos o más clases, y queremos aprender de los datos ya etiquetados con el fin de predecir la clase de datos no etiquetados. Otra forma de pensar en la clasificación es como una forma discreta de aprendizaje supervisado, donde uno tiene un número limitado de categorías, y para cada una de las n muestras proporcionadas tratar de etiquetarlas con la categoría o clase correcta.

2.3.2. Aprendizaje no supervisado

En esta técnica de aprendizaje, los datos de entrenamiento no tienen ningún objetivo correspondiente. El objetivo principal de este tipo de problemas puede ser descubrir grupos de ejemplos similares, llamadas agrupamiento; determinar la distribución de datos dentro del espacio de entrada, conocido como estimación de densidad; proyectar los datos desde una alta dimensión que reduzca el espacio a dos o tres dimensiones para fines de visualización.

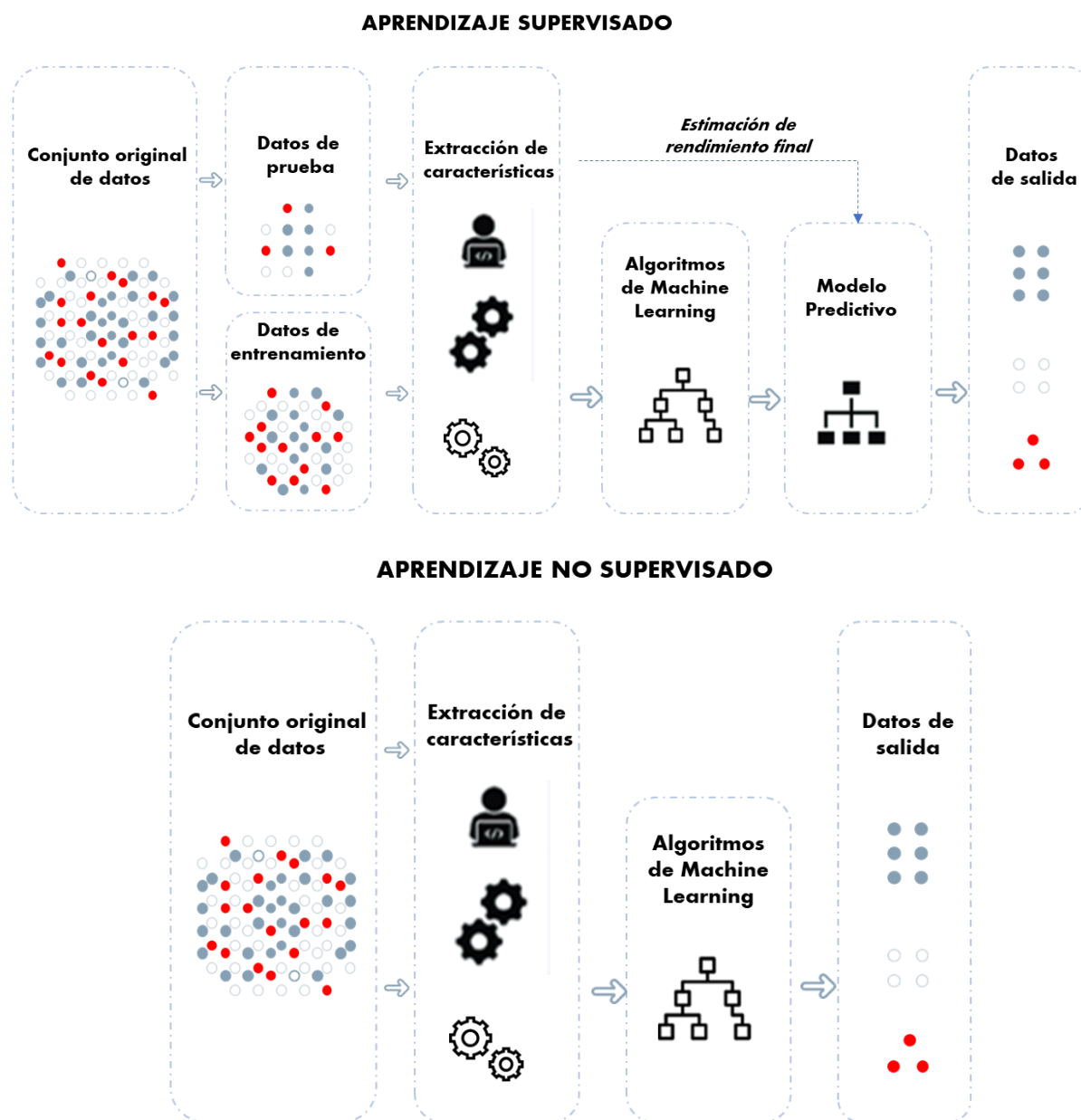


Figura 3. Aprendizaje Supervisado y no Supervisado.

Fuente: Elaboración propia

2.3.3. Aprendizaje reforzado

Las técnicas de aprendizaje por refuerzo se basan en modificar la respuesta del algoritmo. Para este fin usa el proceso de retroalimentación basado en un conjunto de recompensas y castigos.

Esta acción le permite al algoritmo identificar cuando alguna de las acciones que realice previamente obtuvo buenos resultados, o fue un comportamiento que deberá evitar en futuras ejecuciones. Esta técnica de aprendizaje intenta simular el proceso de aprendizaje humano, simulando la sensación de que el algoritmo aprende obteniendo información de cómo se modifica el mundo que lo rodea en respuesta de las acciones que produce.

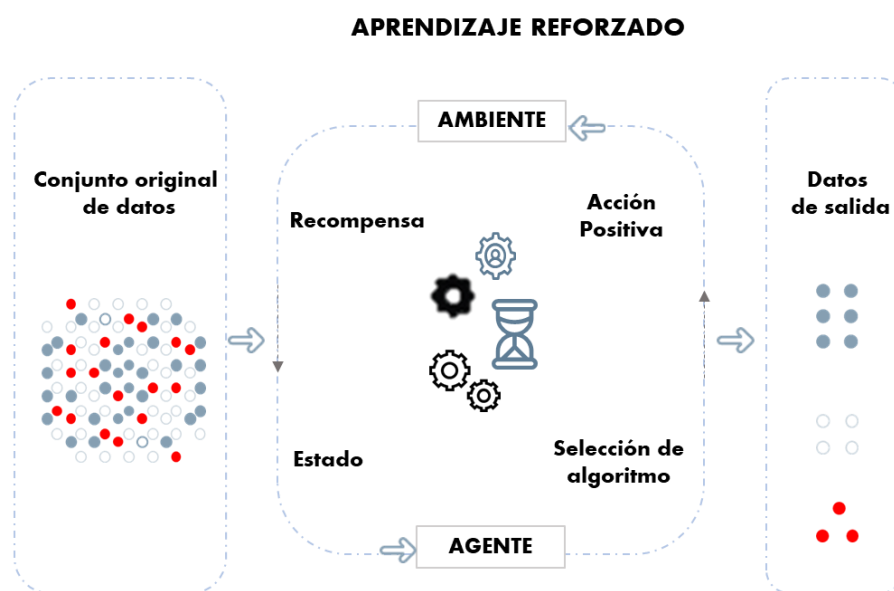


Figura 4. Aprendizaje Reforzado.

Fuente: Elaboración propia

En actualidad existen muchos algoritmos de machine learning que son usados para diferentes fines. A continuación, se menciona los de uso común, especialmente los algoritmos potenciales que se pueden usar en algunos procesos de minería.

- Regresión lineal
- Regresión polinomial
- Support Vector Regression (SVR)
- Árboles de clasificación y regresión (CART)
- Regresión logística
- Máquina de soporte vectorial (SMV)
- K – Means
- Random Forest

- Redes neuronales
- Gradient Boosting
- Voting Regressor.

2.4. Minería de datos

La minería de datos es un proceso con el cual las empresas buscan detectar información valiosa a partir de datos sin procesar. Este proceso que se realiza usando diferentes herramientas sofisticadas, busca deducir patrones y tendencias que existen en los datos, y que difícilmente se podría llevar a cabo mediante exploración tradicional. La finalidad de realizar minería de datos es desarrollar estrategias más efectivas de su negocio y sus procesos. Para llevar a cabo esta tarea se utiliza los métodos de inteligencia artificial, aprendizaje automático, estadística y sistema de base de datos.

Usualmente el término de minería de datos es confundido con el procesamiento de la información como pueden ser la recolección de datos, extracción, almacenamiento, análisis y estadísticas, sin embargo, estos procesos solo son pasos adicionales que ayudan a lograr el objetivo final. Se podría decir que los términos más adecuados para referirse a la minería de datos son el “descubrimiento” y “la detección de algo nuevo”.

2.4.1. Etapas de minería de datos

Según detalla en la guía de minería de datos CRISP – DM (2000) en el año1998, un grupo de actores claves en mercados de minería de datos se reunieron para desarrollar un protocolo estándar para desarrollar minería de datos, CRISP-DM (“Proceso estándar Cross Industry para la minería de datos”). Este protocolo ha sido aceptado por una gran cantidad de científicos de datos y desarrolladores hasta la actualidad. CRISP-DM emplea 6 etapas para extraer la información.

1. Comprensión del negocio
2. La comprensión de los datos
3. La preparación de los datos
4. Modelado
5. Evaluación
6. La implementación

La figura siguiente muestra las seis las seis fases de un proceso de minería de datos. La secuencia de datos no es rígida. Siempre es necesario avanzar y retroceder entre las diferentes fases. Depende del resultado de cada fase o que tarea determinado o requerimiento en particular se está realizando. Además, el círculo hace referencia al proceso cíclico que comúnmente se suele realizar en un trabajo de minería de datos, hasta lograr el objetivo deseado y seguir mejorando la implementación.

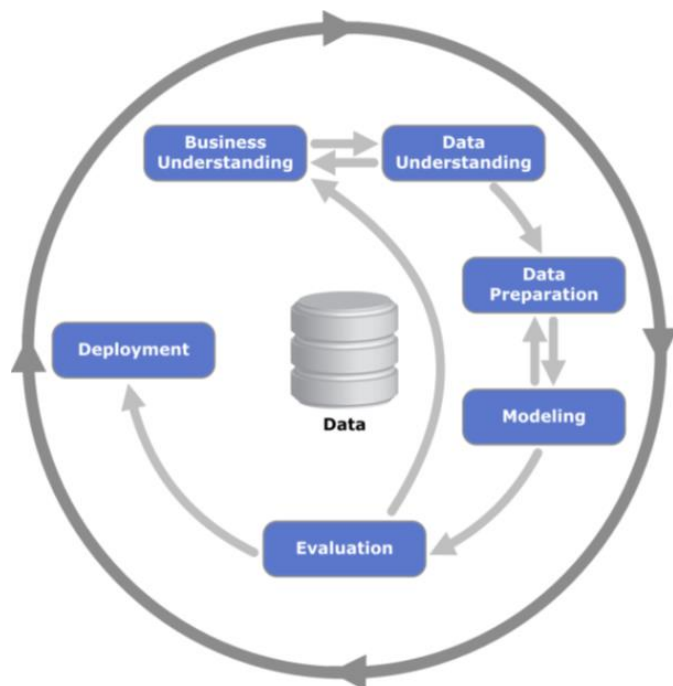


Figura 5: Esquema CRISP-DM

Fuente: Adaptado de *Cross Industry Standard Process for Data Mining*, por Wikipedia Colaboradores, 2021.

Otro esquema de minería de datos que presenta Microsoft consta de los siguientes 6 pasos:

1. Definir el problema
2. Preparación de datos
3. Exploración de datos
4. Creación de modelos
5. Exploración y validación de modelos
6. Implementación y actualización de modelos.

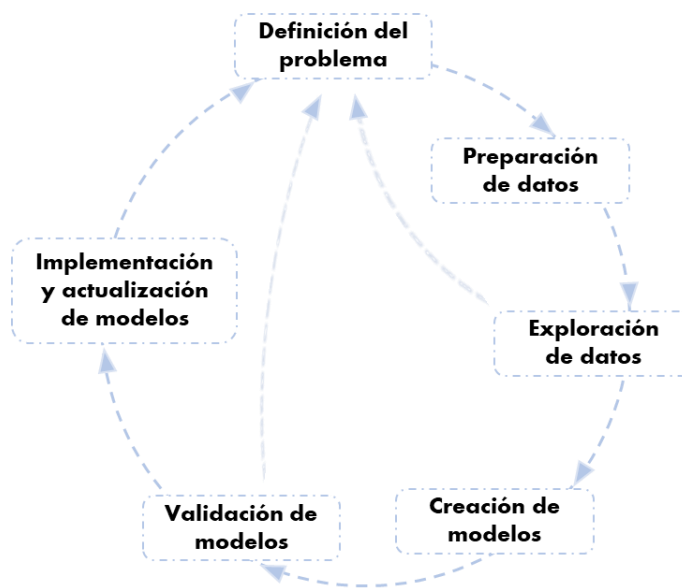


Figura 6: Esquema para minería de datos de Microsoft

Fuente: Adaptado de (Microsoft, 2021)

Si bien es cierto que estos esquemas y terminologías han sido adoptados por una gran comunidad de desarrolladores, estos son modificados y adaptados de acuerdo a una necesidad en particular o de acuerdo a los objetivos trazados cuando se realiza minería de datos. Además, estos esquemas fueron desarrollados por científicos de datos y expertos en analíticas de datos, por lo cual es normal realizar esquemas más simples y de fácil entendimiento para cualquier profesional.

2.4.2. Técnica de minería de datos

Las técnicas de minería de datos sirven para construir modelos de conocimiento, mediante las cuales se pueden identificar patrones de comportamiento de las variables. También es posible usar diferentes técnicas a la vez para generar modelos, pero se debe tomar en consideración que generalmente se realiza un pre procesamiento para cada técnica.

En la actualidad, existe una enorme cantidad de técnicas para realizar minería de datos, algunos de uso más comunes son:

- Asociación
- Clasificación
- Agrupamiento

- Regresión
- Patrones secuenciales
- Redes neuronales

2.5. Industria Minera

La industria minera es uno de los sectores productivos más importantes en el mundo. Esta industria, perteneciente al sector extractivo, es esencial para nuestras vidas y está presente en ella diariamente. Los productos que se obtiene a través de la minería son usados para diferentes tipos de construcción, redes eléctricas y comunicaciones, así como en la fábrica artículos, vehículos, productos electrónicos, medicinas, dispositivos medicinales y muchos otros.

Las principales actividades de la industria minera comprenden el cateo, prospección, exploración, explotación, labor general, beneficio, comercialización, y transporte minero. Estas actividades se realizan generalmente siguiendo el esquema que se muestra en la figura siguiente.

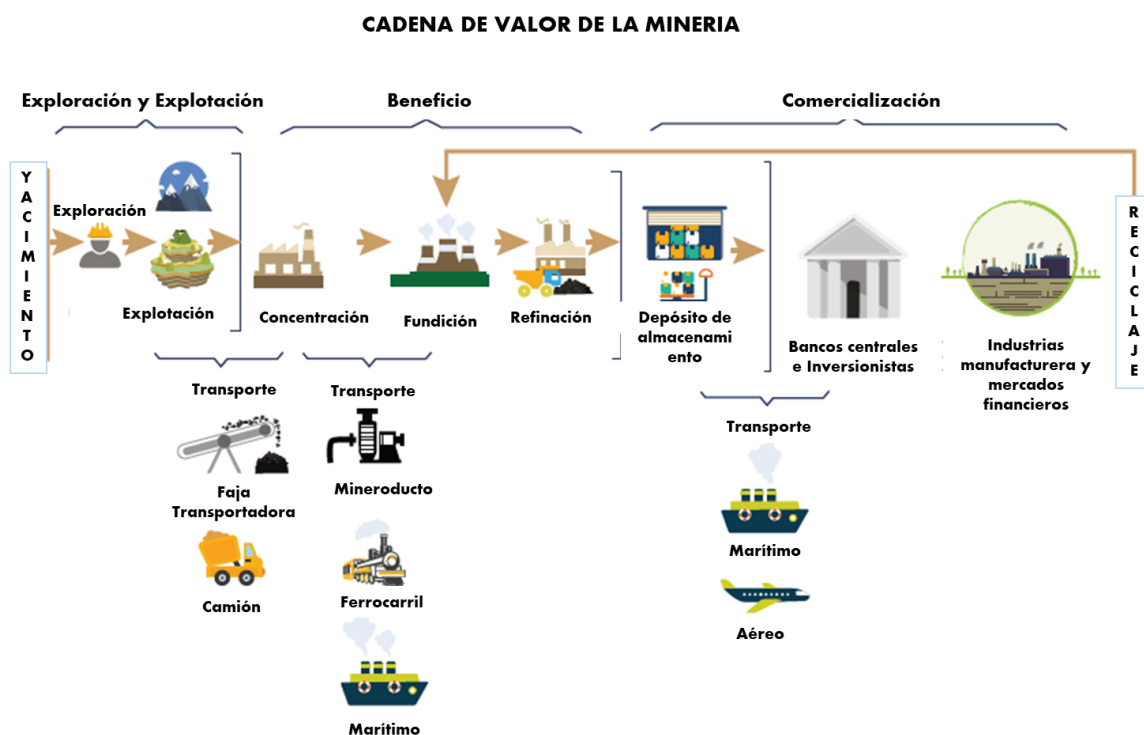


Figura 7: Cadena de valor de la minería

Fuente: adaptado de *La Industria de la minería en el Perú* (p.40), por Osinergmin, 2017

2.5.1. Proyecto minero

Un proyecto minero es una planificación, que consiste en un conjunto de actividades que permiten incrementar el valor de la empresa minera. La idea de un proyecto minero puede pasar desde buscar una solución a un problema (operacional, tecnológico, ambiental) hasta aprovechar una oportunidad de negocio (ampliación de capacidad, abrir una nueva operación). Estas ideas tienen que ser transformadas en estudios que permitan demostrar su viabilidad y la alternativa más conveniente. En esta parte describiremos las etapas y características de un proyecto nuevo de inversión.

Etapas de un proyecto minero

1. Pre inversión

- Exploración
- Preparación de estudios
- Aprobación de estudios y permisos

2. Inversión

- Desarrollo
- Construcción

3. Operación

- Explotación
- Expansión

4. Cierre

- Cierre
- Post Cierre



Figura 8: Etapas y componentes de un Proyecto Minero

Fuente: tomado de *Formulación y Evaluación de Proyectos Mineros* (p.36), por F. Gala, 2018

Características de un proyecto minero

1. Largo periodo de gestación
2. Intensiva en capital
3. Alto riesgo
 - Riesgos geológicos
 - Riesgos ambientales
 - Riesgos sociales
 - Riesgos del proyecto mismo
 - Riesgos de mercado
 - Riesgos políticos
4. Recursos no renovables
5. Estudios multidisciplinarios
6. El estudio de factibilidad debe hacerse en etapas

2.5.2. Etapas del proceso productivo de una mina

Una mina es un conjunto de labores necesarias que permiten la extracción del material valioso de mineral de un yacimiento. En la actualidad, las minas se pueden dividir siguiendo diferentes criterios. Según Wikipedia (2021) “las minas pueden ser divididas siguiendo varios criterios. El más amplio tiene en cuenta si las labores se desarrollan por encima o por debajo de las superficies, dividiéndose, respectivamente, en minas a cielo abierto y en minas subterráneas”.

Ciclo de vida de una mina



Figura 9: Ciclo de vida de una Mina.

Nota: Elaboración propia con información tomada de *Etapas del proceso productivo de una mina*, por M. Cnfield, 2012, Grupo Antofagasta Minerals

Los principales procesos productivos de una mina comprenden:

1. Extracción
 - Extracción de mineral desde la mina hasta la planta de procesamiento
2. Procesamiento
 - Reducción de tamaño por método físico para liberar las partículas metálicas
 - Aumento de la concentración de los metales por métodos físicos – químicos
3. Fundición
 - Separación de los metales contenidos en los concentrados
4. Refinación
 - Purificación de los metales producto de la fundición para su transformación industrial.

2.5.3. Extracción

Las actividades de esta etapa tienen como principal objetivo extraer la roca desde la mina para ser enviada a la siguiente etapa. Este proceso de extracción del mineral valioso de un yacimiento se puede llevar a cabo generalmente mediante dos formas:

- Extracción a tajo abierto
- Extracción subterránea

Extracción de mina a tajo abierto

El objetivo de este proceso es extraer el material roto desde la mina para ser enviada a la etapa siguiente de procesamiento (mineral) o a botaderos (baja ley y estéril).

Las principales actividades o subprocesos involucrados son:

- Perforación
- Voladura
- Carguío
- Transporte



Figura 10: Procesos de una mina a tajo abierto.

Nota: Elaboración propia con información tomada de *Etapas del proceso productivo de una mina*, por M. Cnfield, 2012, Grupo Antofagasta Minerals

Extracción de mina subterránea

El objetivo es extraer el material roto desde la mina para ser enviada a la planta de procesamiento de mineral. La operación de una mina subterránea se puede llevar a cabo mediante

diferentes métodos de explotación y sus variantes. La elección de estos métodos van depender de aspectos técnicos, económicos, geo mecánicos o escala de la explotación.

Las principales actividades o subprocesos involucrados son:

- Perforación
- Voladura
- Carguío
- Transporte



Figura 11: Procesos de una Mina Subterránea.

Nota: Elaboración propia con información tomada de *Etapas del proceso productivo de una mina*, por M. Cnfield, 2012, Grupo Antofagasta Minerals

2.5.4. Procesamiento

Esta etapa consiste en someter al mineral extraído de la mina a varios subprocesos con la finalidad aumentar su concentración (contenido metálico) y prepararlo para el proceso de fundición y refinación o en algunos casos para la venta final. Los métodos que se eligen para llevar acabo estos procesos metalúrgicos van a depender generalmente de las características del yacimiento.

Los principales métodos usados son:

- Flotación – en sulfuros (minerales profundos)
- Lixiviación – en óxidos (minerales en superficie)

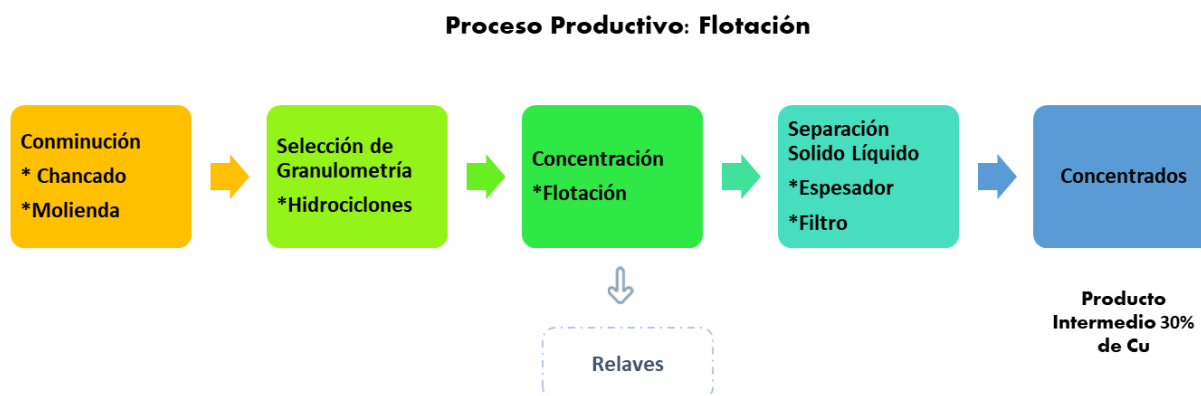


Figura 12: Procesos de Flotación.

Nota: Elaboración propia con información tomada de *Etapas del proceso productivo de una mina*, por M. Cnfield, 2012, Grupo Antofagasta Minerals



Figura 13: Procesos de Lixiviación.

Nota: Elaboración propia con información tomada de *Etapas del proceso productivo de una mina*, por M. Cnfield, 2012, Grupo Antofagasta Minerals

2.5.5. Fundición y Refinería

Procesos de fundición – este proceso consiste en la separación del metal valioso que contiene el concentrado por un proceso pirometalúrgico (un proceso que consiste en la fundición del concentrado con altas temperaturas).

Este proceso permite obtener metales con pequeñas cantidades de impureza. Por ejemplo, el Cobre Blíster – 99.2 % de Cu. Además, a través de este proceso podemos obtener otros subproductos importantes que pueden ser utilizados en otros procesos.

Proceso de refinación – Este proceso te permite obtener metales con alto grado de pureza y listos para ser usados en la industria. Por ejemplo, cátodos de cobre – 99.99% de Cu.

- Refinación a fuego (pirometalurgia)
- Refinación por electro – obtención (EW)

2.6. Inteligencia Artificial en minería

La industria minera es una actividad que requiere búsqueda permanente de eficiencia para sus diferentes procesos. A pesar de los esfuerzos de las empresas mineras por mejorar las operaciones cada día, todavía vienen enfrentando a desafíos cada vez mayores y difíciles de controlar con las tecnologías heredadas. Uno de estos problemas es mejorar constantemente la productividad. Junto a ello, la minería ha venido enfrentado otros problemas como la volatilidad de los precios de los metales, cambio en la naturaleza del trabajo, madurez de las minas, barreras de innovación, situaciones políticas y conflictos sociales.

Para poder enfrentar estos problemas a futuro, muchas de las empresas están buscando tecnologías avanzadas relacionadas a la Inteligencia artificial. Estas herramientas integradas a la inteligencia humana harán que el flujo de trabajo mejore significativamente.

Actualmente, muchas operaciones mineras están optando por la digitalización en sus procesos con el fin de monitorear sus actividades en tiempo real, lo cual le permite optimizar el uso de sus recursos. Así, se ha mejorado el control del consumo excesivo de energía y combustible, que no solo tiene implicancias económicas sino también ambientales. Un claro ejemplo del uso de tecnología para optimizar el consumo energético es la mina Kankberg de Boliden, que, gracias a la implementación de una solución digital Ventilation Optimizer redujo el consumo energético de los ventiladores en un 54% (Castro, 2021).

Otra de las innovaciones que ayuda a mejorar las actividades unitarias en las operaciones mineras es el uso de vehículos autónomos. Estas permiten mejorar los tiempos de producción, además, reduce la exposición de los operadores a riesgos que pueden tener consecuencias desfavorables. Los vehículos autónomos están fabricados para laborar en condiciones en las que difícilmente podrían estar los humanos. Este tipo de equipos se está implementando en la planta Pilbara de la empresa Río Tinto, con su aplicación han logrado aumentar en 1000 horas el trabajo

por camión con una reducción del 15% de costo; demás, ayudó a generar un aumento de flujo efectivo de 500 millones de dólares (Sepúlveda, 2019). En el Perú la empresa Angloamerican anunció que viene planificando la implementación de camiones y perforadores autónomos en su mina Quellaveco.

Deloitte & Northern Centre for Advanced Technology (2020) a través de una serie de entrevista a empresas mineras que están desarrollando aplicaciones basadas en inteligencia artificial y machine learning, llegaron a la conclusión que la adopción de cualquier tecnología digital está plagada de desafíos, desde periodos de transición que carecen de resultados tangibles, hasta innumerables desafíos relacionados con la calidad de los datos, la cultura organizacional y las capacidades limitadas de algunos aspectos. Sin embargo, también menciona que la adopción de estas tecnologías permite a las empresas tomar decisiones más precisas y rápidas, mejorar la salud y la seguridad, aumenta la eficiencia y garantiza menos errores humanos. Un panorama muy bueno para que las empresas mineras adopten tecnologías basadas en inteligencia artificial y machine learning.

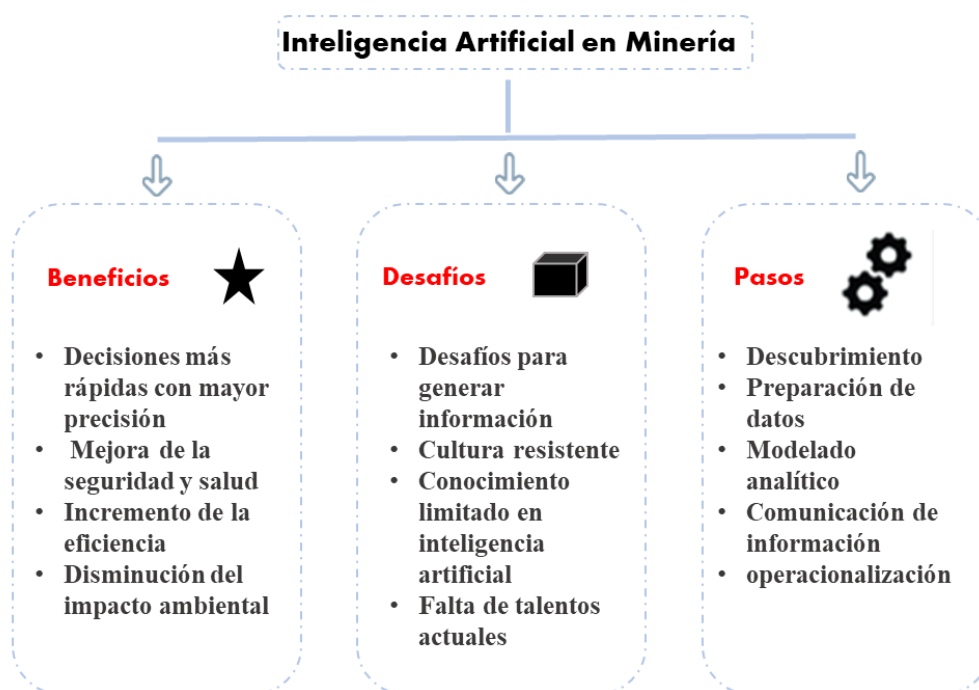


Figura 14. Inteligencia Artificial en Minería.

Nota: Elaboración propia, el esquema fue elaborado tomando datos de *Future of Mining with AI*, por Deloitte & NORCAT, 2020

2.6.1. Como adoptar Inteligencia Artificial (IA) en minería

Empresas mineras que ya integran fuentes de datos dispares y utilizan la tecnología para generar conocimientos y capacidades de alto nivel, comprenden bien que la tecnología por sí sola no es suficiente. Para cumplir los objetivos de la implementación de IA en minería se requiere el equilibrio adecuado de habilidades comerciales y técnicas, comenzando por conocer la diferencia entre simplemente adquirir tecnología y desarrollar un conjunto de funciones y capacidades analíticas adecuadas (Deloitte & NORCAT, 2020). En última instancia, se trata de implementar aplicaciones de inteligencia artificial sensatas y del tamaño adecuado que permitan a las empresas derivar soluciones puntuales y a largo plazo.

Omnia AI (2020), la unidad de Inteligencia Artificial (IA) de la consultora Deloitte, describe estos 5 pasos para ayudar a las empresas mineras que invierten en tecnología de IA optimizar sus esfuerzos y resultados:

1. **Descubrimiento** - Los líderes de la industria minera confirman que la planificación y el descubrimiento cuidadoso desde el principio de la implementación de una tecnología pueden reducir significativamente el riesgo. Las empresas comienzan alineando los objetivos y aspiraciones de los líderes empresariales con la forma en que se ve el futuro de la minería para la organización. A continuación, se hacen las preguntas correctas para identificar a los usuarios clave y los casos de uso; comprender las tendencias y oportunidades de la industria; identificar los recursos y determinar dónde la IA puede ayudar mejor.
2. **Preparación de datos** – Uno de los puntos fundamentales para implementar IA es centrarse en datos con un alto potencial de impacto. Para este fin se debe implementar una metodología bien estructurada, repetible, revisable y transferible. Con un énfasis en el diseño de planes de contingencia para tener en cuenta lo imprevisto, así como la contabilidad de las suposiciones y la identificación de cuándo detener el análisis para reevaluar e iterar. En resumen, esta fase implica transformar los datos de origen sin procesar en una forma que se pueda utilizar para desbloquear conocimiento y generar decisiones de negocios.

3. **Modelado analítico** - Al representar la fase de desarrollo de un compromiso tecnológico, el objetivo ahora es utilizar las mejores prácticas de la industria para analizar los datos estructurados a fin de determinar el modelo de datos necesarios para abordar el desafío, desde el aprendizaje automático hasta los modelos de aprendizaje profundo basados en la complejidad del caso comercial. Durante esta fase se mejora la gestión de la complejidad operativa, como la contabilidad, aspectos como la profundidad de la mina y otros factores relevantes para un análisis preciso. El modelado basado en el contexto también ayuda a abordar los riesgos imprevistos, incluida la falta de infraestructura técnica disponible para la implementación o la baja producción durante la implementación inicial de la tecnología relacionada con la IA.
4. **Comunicación de información** - Se genera una visualización de datos intuitiva y traduce constantemente los datos numéricos en un impacto operativo, mediante el cual todas las partes interesadas comprenden cómo la tecnología contribuye a los ingresos y al desarrollo de la capacidad organizativa. El proceso también permite a los líderes actuar sobre la base de los conocimientos y explorar más a fondo los modelos analíticos.
5. **Operacionalización** - El establecimiento de pautas estructuradas en torno a la naturaleza repetible e iterativa de las soluciones visualizadas, brinda a los líderes mineros la flexibilidad en la gestión diaria de sus capacidades analíticas para identificar qué unidades se beneficiarán de la solución, y evaluar más a fondo los enfoques operativos. Las soluciones operativas deben estar posicionadas para resolver el desafío, documentando los procesos a lo largo del camino para garantizar que las iniciativas se puedan llevar desde la idea hasta la implementación.

2.6.2. Casos de aplicación de Inteligencia Artificial en minería.

A continuación, se describe algunas aplicaciones de inteligencia artificial aplicados en algunas operaciones mineras.

Goldspot Discoveries Inc.

Utiliza la inteligencia artificial para poder detectar yacimientos complejos utilizando datos geológicos, topográficos y mineralógicos. Su tecnología llamada “LithoLens” permite examinar de

forma automática y consistente imágenes de puntos de perforación, fotografías de afloramiento y sensores de pozo con el fin de crear registros geológicos actualizados y precisos. Para ello, usa algoritmos de aprendizaje profundo que son procesados en la nube, generando un conjunto de datos cada vez más grandes y valiosos (Goldspot, 2021).

IBM Watson

IBM en alianza con Goldcorp Inc. están trabajando para llevar la tecnología de Watson a la industria minera. Watson es una tecnología de IA que tiene capacidad de comprender, aprender, razonar e interactuar. Esta tecnología ayuda analizar grandes cantidades de datos provenientes de perforaciones e informes de estudios geológicos. Entre las principales ventajas de utilizar esta tecnología está: determinar áreas específicas para la exploración, alcanzar los objetivos de exploración más rápidos y calcular modelos geológicos con más certeza. Esto mejorará la capacidad de los geólogos para extraer nueva información a partir de datos existentes, además, de ofrecer conocimientos regionalizados que ayudarán en el proceso de exploración (Walker, 2019).

Rio Tinto.

Usa la tecnología de Machine Learning para implementar camiones autónomos en algunas de sus operaciones. Estos camiones autónomos, proporcionados por la empresa japonesa Komatsu, encuentran su camino usando GPS de precisión y buscan obstáculos usando sensores de radar y láser. Con el fin de anticipar un futuro más eficiente para todas sus minas, Rio Tinto también está en proceso de automatizar todas sus operaciones de Australia, en ello incluyen: cargadores autónomos, perforadoras autónomas y transporte de concentrado autónomo. En el futuro la empresa tiene planeado implementar un tren autónomo para el transporte de sus minerales (Walker, 2019).

Tombra.

Es un equipo que utiliza machine learning para clasificar los materiales que serán extraídos. El equipo realiza la clasificación a través del color, transmisión rayos x o sensores de infrarrojo. La diversa gama de tecnología de reconocimiento avanzado que tiene los sensores de Tombra, ofrece solución de clasificación para cualquier tipo de material de mineral. Las soluciones de clasificación que tiene Tombra combinan el manejo de materiales, el reconocimiento y las

tecnologías de expulsión de aire presurizado en un sistema completamente optimizado que separa de manera confiable los minerales valiosos de la roca estéril (Walker, 2019).

Petra and GE

Estas dos empresas han utilizado la tecnología de machine learning para poder predecir las fallas que podrían tener los equipos, y las sobrecargas que podrían tener los molinos en una planta de diamantes.

Deep Pit Technology.

A través de su aplicación web MineExpert implementó en la mina Constancia un modelo de predictibilidad del Throughput usando algoritmos de machine learning. Según Eder Lagos (2021) la tecnología de Throughput ayuda con el rastreo del mineral desde el frente de minado hasta llegar a la chancadora; así como también les permite predecir con efectividad mayor al 90% los eventos que ocurrirán en el molino.

RockMass Technologies.

Esta empresa ha implementado sensores de última generación para capturar datos en tiempo real e identificar posibles planos de falla en las superficies rocosas. Para ello, utiliza un hardware portátil para analizar superficies rocosas y proporcionar los datos al usuario en minutos. Esto se hace con un software que analiza los datos 18 veces más rápido que los métodos manuales actuales. El resultado es un método consistente y eficiente para realizar una evaluación precisa y rápida de los riesgos potenciales (Deloitte & NORCAT, 2020).

Thorough Tec.

La empresa aprovecha la capacidad de IA para implementar sensores portátiles con el fin de monitorear continuamente el comportamiento de los trabajadores. Este proceso ayuda generar datos que luego serán analizados automáticamente para detectar tendencias de comportamiento problemáticas, y recomendar capacitación correctiva altamente enfocado en un entorno basado en simuladores. Esto permite a los gerentes de mina no solo responder de manera proactiva a brechas en las capacitaciones de los operadores, sino también optimizar las intervenciones de capacitación para adaptarse mejor a las necesidades individuales de los trabajadores y, en última instancia, elevar los estándares de desempeño del operador en toda la fuerza laboral (Deloitte & NORCAT, 2020).

Ionic Engineering.

Usa el aprendizaje automático para mejorar enormemente el reconocimiento de imágenes para la identificación de leyes de cobre; con ello reduce la tasa de erros en comparación con los empleados que realizan el trabajo. Al entrenar la máquina utilizando redes neuronales para que identifique las características que el operador está buscando, se reduce enormemente el tiempo de operación y la calidad del trabajo; aumentando así el potencial para reducir costos (Deloitte & NORCAT, 2020).

Shyft Inc.

En las minas subterráneas la ventilación es la que genera mayor consumo de energía. Para resolver este problema, Shyft Inc. está empleando machine learning con datos históricos de seis años para pronosticar picos de energía. El ajuste automático de los sistemas de ventilación mediante el control de procesos integrado, ayuda a reducir significativamente los costos de energía (Deloitte & NORCAT, 2020).

Sin duda alguna el éxito de las empresas mineras dependerá cada vez más de los cambios en la forma de hacer negocios. Una de ellas es el uso de la tecnología que garantice la utilización de los recursos de manera óptima, además, que ayude a la adaptación de trabajadores a las habilidades cambiantes del negocio minero. Empresas como Rio Tinto han comprobado el incremento en la capacidad de realizar tareas, toma de decisiones más rápidas, interacciones atractivas, además de generar resultados comerciales más sólidos.

Capítulo III: Marco metodológico de la Investigación

3.1. Metodología de trabajo

Para lograr los objetivos planteados es muy importante definir el esquema que se debe seguir para obtener resultados esperados. A continuación, se presenta de manera general un esquema de la metodología usada para el desarrollo de esta investigación.

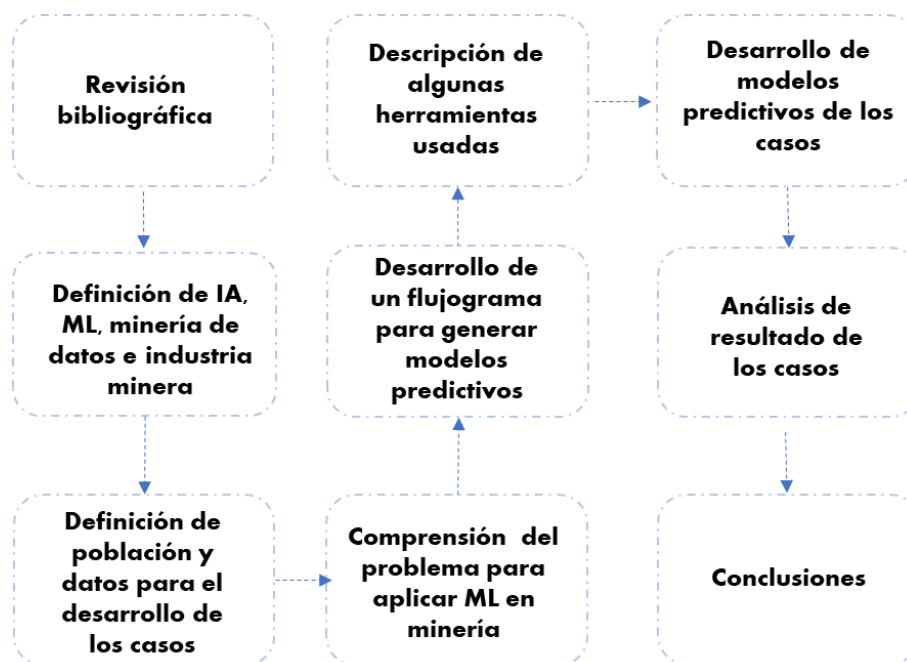


Figura 15: Esquema de metodología de trabajo.

Fuente: Elaboración propia

La descripción de este esquema consta de los siguientes puntos:

- Revisión bibliográfica (Descrita en el cap II, 2.1)
- Definición de Conceptos (Descrita en el cap II, 2.2)
- Definición de población y datos para el desarrollo de los modelos.
- Comprensión del problema para aplicar machine learning en minería.
- Desarrollo del flujograma para generar modelos predictivos y descripción de sus etapas – Con base en revisión de trabajos similares y el esquema general para

realizar minería de datos, desarrollaremos y describiremos todas las etapas del flujograma más intuitivo y de fácil entendimiento.

- Descripción de algunas herramientas usadas.
- Desarrollo de modelo predictivos de los casos – en esta etapa aplicaremos el flujograma desarrollado con el fin proponer soluciones basadas en modelos predictivos de machine learning, estos son:
 - Modelo predictivo para estimar el precio de oro
 - Modelo predictivo para estimar el contenido de sílice en el concentrado de hierro.
 - Modelo predictivo para estimar el consumo de combustible de camiones mineros de gran tonelaje.
- Análisis de resultados de los ejemplos desarrollados.
- Conclusiones.

3.2. Tipo de investigación

Bajo la exposición de Hernández Sampieri & Baptista Lucio, la investigación será de enfoque cuantitativo, del tipo exploratorio y en menor medida correlacional. Esto quiere decir que recopilaremos datos de diferentes fuentes confiables para poder generar modelos predictivos utilizando el flujograma desarrollado. Además, el alcance exploratorio implica que se aplicará diferentes algoritmos para generar modelos predictivos. Esto permitirá hacer un análisis comparativo de los algoritmos que mejor se adecuen a nuestros objetivos, con ello, dejaremos un campo abierto para que otras investigaciones puedan analizar la efectividad de los diferentes modelos de predicción existente.

3.3. Diseño de proceso de investigación

Nuestra investigación será no experimental y de diseño transversal, ya que no vamos a intervenir o manipular deliberadamente las variables de estudio. El diseño transversal hace referencia a que recopilaremos los datos para nuestra investigación en un momento único.

3.4. Población y datos de objeto de estudio

Caso 1: para generar un modelo predictivo de precio de oro se buscó datos históricos en un intervalo de 01 de enero del 2015 al 22 de julio del 2021 con una frecuencia diaria.

Caso 2: para generar un modelo predictivo que indique el % de sílice en un concentrado de hierro, se buscó una información histórica del proceso de flotación de una planta de procesamiento de mineral de hierro.

Caso 3: para generar un modelo predictivo para la estimación de consumo de combustible, se buscó una información histórica de consumo diario de combustible de camiones mineros de gran tonelaje modelo Cat 797F.

3.5. Método y técnica de recolección de datos

Para la presente investigación se usó el método secundario de recolección de datos. Este método consiste en recolectar los datos recurriendo a fuentes secundarias. Estas fuentes contienen información para el desarrollo de la presente investigación. Además, su fácil acceso y uso permitió el ahorro de tiempo y recurso.

Para poder tener datos confiables, auténticas y creíbles se utilizaron las siguientes fuentes de información:

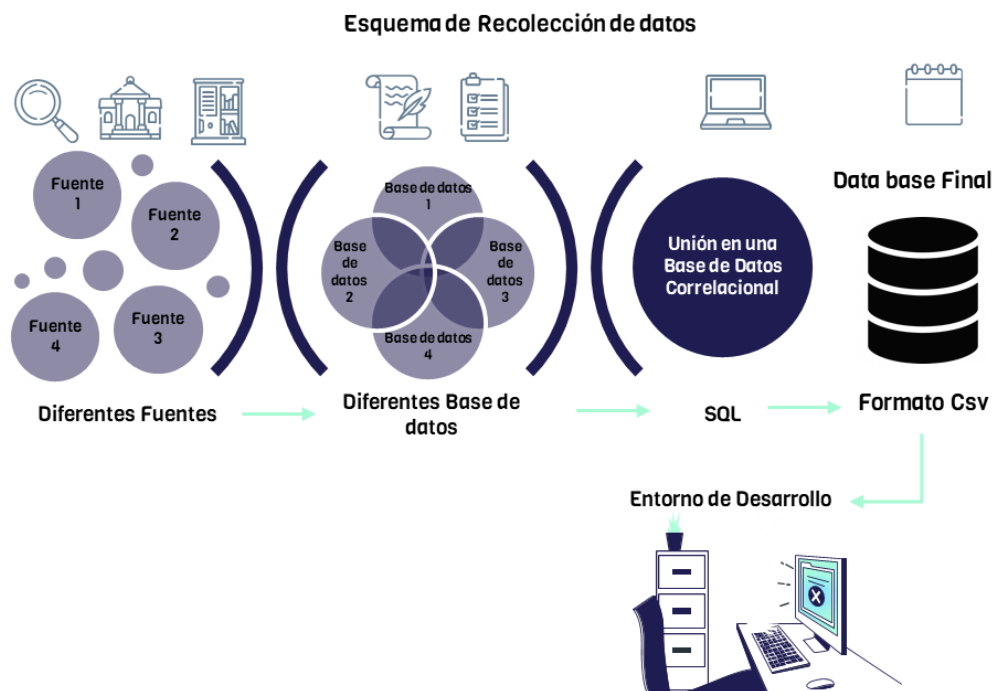
Fuentes oficiales:

- Ministerio de Energía y minas (Precio del Oro)
- Oficina nacional de estadísticas de china (PBI de China)
- Banco mundial (PBI de los estados unidos)

Fuentes privadas:

- Investing.com (Índice del precio del dólar)
- fred.stlouisfed.org (Tasa de la reserva federal e inflación de los Estados Unidos)
- Gold.org (Consumo mundial de oro, Producción de oro, Costo de producción)

- Investing.com (Índice S&P, Precio de petróleo)
- Kaggle.com (Base de datos concentrado de mineral de hierro)
- Ing. B Huaranga Yantas (Base de datos de historial del consumo de combustible)
- Ing. E Vásquez Alfaro (Base de datos de accidentes en minería).



Fuente: Elaboración propia

Capítulo IV: Flujograma para generar modelos predictivos de regresión con el uso de algoritmos de machine learning

4.1. Comprensión del problema

La minería es una industria que nos brinda un amplio campo para la implementación de nuevas tecnologías. Sin embargo, la sola aparición de una tendencia nueva o una tecnología novedosa no justifica su aplicación. Para evitar caer en falsas expectativas, es importante identificar el problema que se quiere resolver y de qué manera el uso de una tecnología diferente impactará en la mejora de resultados frente a los procesos tradicionales.

La gran mayoría de los expertos en IA y los altos ejecutivos de las principales operaciones mineras en el Perú, coinciden que la tecnología que se quiere implementar tiene que cumplir principalmente estos tres objetivos:

- Ayuden a mejorar la productividad
- Ayude a mejorar la toma de decisiones en tiempo más corto
- Que tengan un retorno de la inversión que se hace en la implementación de nuevas tecnologías.

Para lograr estos objetivos, es muy importante tener al menos una idea del camino que seguiremos e ir buscando más información en base a las necesidades particulares. Si bien es cierto que existen diferentes términos o cada objetivo tiene su propio flujograma para desarrollar aplicaciones basadas en machine learning, todos comienza con la comprensión del problema que se quiere resolver y/o al negocio que se quiere aplicar.

4.2. Desarrollo del flujograma de trabajo

Con base en revisiones bibliográficas y trabajos similares, hemos creado un flujograma para el desarrollo de modelos predictivos con la técnica de aprendizaje supervisado. A continuación, explicaremos los detalles del flujograma (Ver **figura 16**) para luego aplicarlas en los casos de ejemplos de esta investigación.

Si bien es cierto que en el flujograma que se muestra en la **Figura 16** hay un orden que se debe respetar inicialmente al desarrollar modelos con algoritmos de machine learning, es fundamental resaltar que, una vez obtenida un resultado preliminar en el desarrollo del modelo, es normal la interacción de los diferentes pasos para ir mejorando y obtener un resultado deseado.

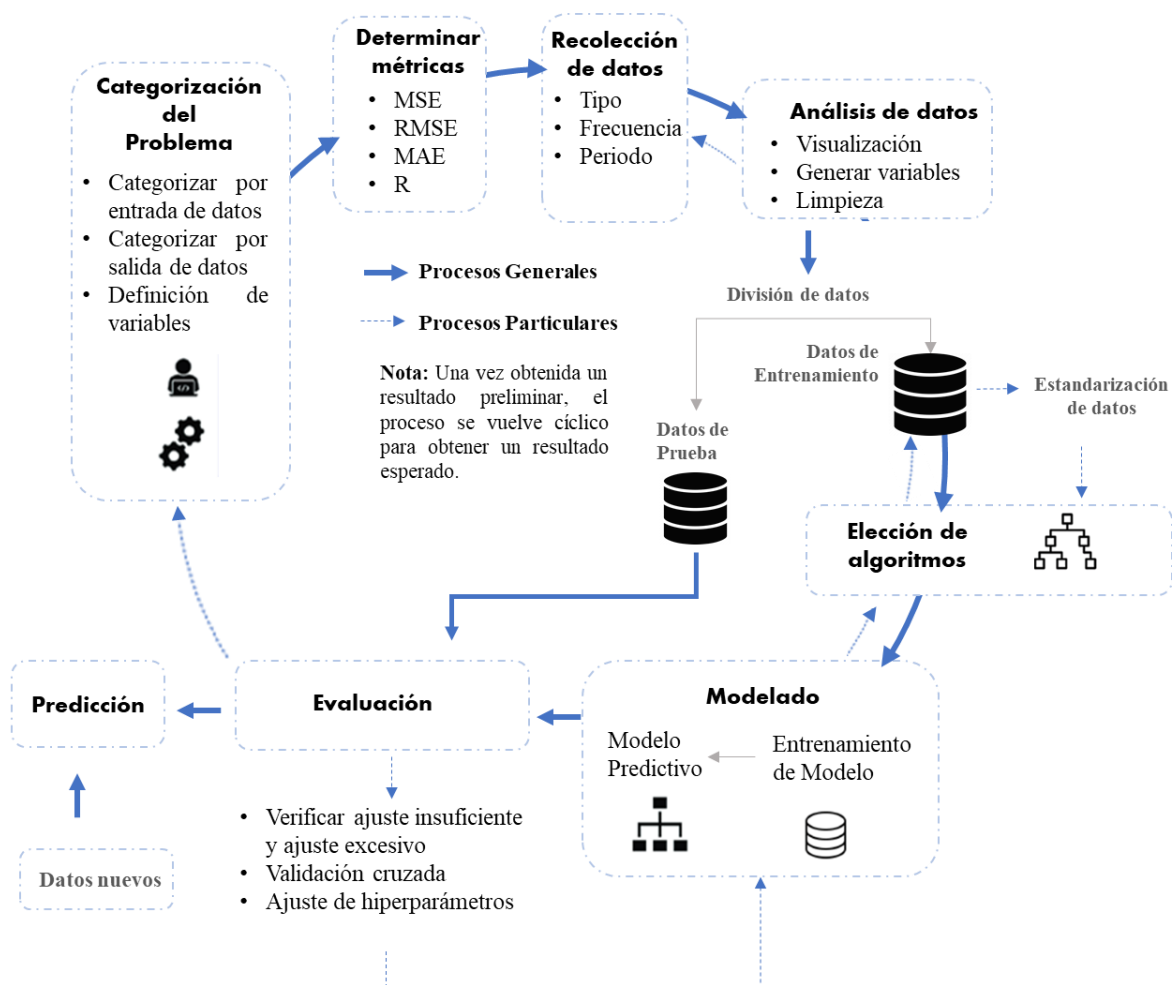


Figura 16. Flujograma para el desarrollo de modelos predictivos de regresión.

Fuente: Elaboración propia

4.2.1. Categorización del problema

Dentro de la industria minera existen muchos problemas potenciales en las cuales se pueden aplicar las diferentes técnicas de machine learning, con el fin de realizar diferentes tareas. Este paso del flujograma consiste en identificar qué técnica de machine learning emplearemos para resolver un problema determinado. La importancia de este paso radica en que nos ayudará a definir las variables necesarias que se usará en los siguientes pasos.

Para la categorización del problema nos basaremos en un proceso que consiste en dos pasos.

Categorizar por entrada:

- Si ha etiquetado datos, es un problema de aprendizaje supervisado
- Si tiene datos no etiquetados y desea encontrar una estructura, es un problema de aprendizaje sin supervisión.
- Si desea optimizar una función objetivo interactuando con un entorno, es un problema de aprendizaje por esfuerzo.

Categorizar por salida:

- Si la salida del modelo es un número, es un problema de regresión.
- Si la salida del modelo es una clase, es un problema de clasificación.
- Si la salida del modelo es un conjunto de grupos de entrada, es un problema de agrupamiento.
- ¿Quieres detectar alguna anomalía? Eso es detección de anomalías.

Tabla 1. Métodos de Machine Learning

Técnicas ML	Tipo de Data	Análisis	Objetivo	Enfoque
Aprendizaje Supervisado	Datos de entrada y salida son conocidos y pre categorizados	Offline	El resultado se predice utilizando la data de entrada categorizada	Regression Clasificación
Aprendizaje no Supervisado	Solo los datos de entrada son conocidos	En tiempo real	El resultado se predice en base a los patrones en los datos de entrada	Clustering Representation Learning
Aprendizaje por reforzamiento	Data no predefinida	En tiempo real y con interacción de un agente	El resultado se predice en base a patrones y al enfoque de prueba y error	Optimización de estrategias Behavior learning

Fuente: Adaptado de *Desarrollo de un modelo para la predicción del precio del cobre empleando herramientas de Machine Learning* (p.41), por F. Gamarra, 2020.

Para un mejor entendimiento mencionaremos algunos ejemplos de casos que actualmente han sido implementados en algunas operaciones.

Empresa: Shyft Inc

Técnica de Machine Learning: Aprendizaje supervisado

Tarea o enfoque: Regresión

Solución: En las minas subterráneas la ventilación es la que genera mayor consumo de energía. Para resolver este problema, Shyft Inc está empleando machine learning con datos históricos de seis años para pronosticar picos de energía. El ajuste automático mejorado de los sistemas de ventilación mediante el control de procesos integrado, ayuda a reducir significativamente los costos de energía.

Empresa: Tombra

Técnica de Machine Learning: Aprendizaje supervisado

Tarea o enfoque: Clasificación

Solución: Es un equipo que utiliza machine learning para clasificar los materiales que serán extraídos. El equipo realiza la clasificación a través del color, transmisión rayos x o sensores de infrarrojo. La diversa gama de tecnología de reconocimiento avanzado que tiene los sensores de Tombra ofrece solución de clasificación para cualquier tipo de material de mineral. Las soluciones de clasificación que tiene Tombra combinan el manejo de materiales, el reconocimiento y las tecnologías de expulsión de aire presurizado en un sistema completamente optimizado que separa de manera confiable los minerales valiosos de la roca estéril.

Empresa: Ionic Engineering

Técnica de Machine Learning: Aprendizaje no supervisado

Tarea o enfoque: Agrupamiento

Solución: Usa el aprendizaje automático para mejorar enormemente el reconocimiento de imágenes para la identificación de leyes de cobre, con ello reduce la tasa de erros en comparación con los empleados que realizan el trabajo. Al entrenar la máquina utilizando redes neuronales, para que identifique las características que el operador está buscando, se reduce enormemente el tiempo de operación y la calidad del trabajo; aumentando así el potencial para reducir costos.

4.2.2. Determinación de métricas

La definición de las métricas de medición para evaluar la precisión del modelo es muy importante para desarrollar modelos predictivos con algoritmos de machine learning, ya que este definirá la capacidad predictiva del modelo. Esto se llevará a cabo tomando en consideración los objetivos empresariales o de acuerdo al problema a resolver.

Para el desarrollo de los ejemplos utilizaremos las métricas que se pueden encontrar en la librería Scikit – Learn. La elección de las métricas se basó en la identificación y categorización de los problemas escogidos para presentar en los ejemplos que desarrollaremos en el “Capítulo V”. A continuación, describiremos cada una de ellas.

a) Error Cuadrático Medio (MSE)

El error cuadrático medio (MSE) es una métrica que se calcula como la media o el promedio de las diferencias al cuadrado entre los valores objetivo predichos y esperados en un conjunto de

datos. El MSE tiene el efecto de castigar a los errores más grandes. Es decir, cuando mayor se la diferencia entre el valor predicho y esperado, mayor será el valor de MSE.

Está representado por la siguiente fórmula.

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

Donde: $y_i = \text{valore real}$
 $\hat{y}_i = \text{valore predicho}$
 $N = \text{número de muestras}$
 $i = \text{iésimo - valor}$

b) Error Cuadrático Medio (RMSE)

El RMSE simplemente es la raíz cuadrada de MSE, lo que lo diferencia es que esta métrica está en las mismas unidades del valor objetivo. Es muy común usar la pérdida de MSE para entrenar un modelo predictivo de regresión y usar RMSE para evaluar y reportar su desempeño.

Está representada por la siguiente fórmula:

$$RMSE = \sqrt{MSE}$$

c) Error Absoluto Medio (MAE)

El error absoluto medio (MAE) se calcula como el promedio de los valores de error absoluto. A diferencia del MSE o RMSE que castigan los errores más grandes que los errores más pequeños, en cambio, en el MAE las puntuaciones aumentan linealmente con los aumentos del error. Es una métrica muy popular debido a que las unidades de puntuación de error coinciden con las unidades del valor objetivo que se predice.

Está representada por la siguiente fórmula:

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

Donde: $y_i = \text{valore real}$
 $\hat{y}_i = \text{valore predicho}$
 $N = \text{número de muestras}$
 $i = \text{iésimo - valor}$

d) Puntaje R^2 o el coeficiente de determinación

Esta métrica de error es más intuitiva que los anteriores mencionados debido a que están libre de escala. En este caso no importa si los valores son muy grandes o pequeños, esta métrica siempre estará entre $-\infty$ y 1. Un valor cercano a 1 significa que el error es demasiado pequeño, por el contrario, un error cercano a cero o negativo significa un error demasiado grande.

Está representada por la siguiente fórmula:

$$R^2 = 1 - \frac{\text{MSE}(\text{model})}{\text{MSE}(\text{baseline})}$$

4.2.3. Recolección de datos

La recopilación de datos es muy determinante cuando se genera modelos predictivos usando algoritmos de machine learning, estos pueden definir el algoritmo que se va usar o influir directamente en el desempeño final del modelo. Para ello, definiremos el tipo de dato, el periodo, la frecuencia, la fuente, número de datos, etc. Estos puntos serán descritos con mayor detalle en cada ejemplo de aplicación (Ver cap V, 5.1.2).

4.2.4. Análisis de datos

El primer paso es comprender los datos, dado que estos desempeñan un papel importante en la elección del algoritmo a utilizar. Los algoritmos no siempre funcionan bien con cualquier tipo de datos. Algunos pueden funcionar bien con una pequeña cantidad de datos, mientras que otros necesitan una gran cantidad de datos para funcionar bien. Algunos algoritmos son muy sensibles a los valores atípicos y otros los manejan bien. El análisis de datos consta de los siguientes puntos:

a. Visualización de datos

Entre las estadísticas, resúmenes y mejor visualización tenemos.

- Percentiles que nos ayudan a identificar el rango para la mayoría de los datos.
- Los promedios y las medianas que describen la tendencia central.
- Las correlaciones que indican las relaciones sólidas.
- Los diagramas de cajas pueden identificar valores atípicos.
- Las gráficas de densidad y los histogramas muestran la dispersión de los datos.

b. Limpieza de datos

Para realizar la limpieza de datos es muy importante verificar si existen datos incompletos, incorrectos, inexactos, no pertinentes, e irrelevantes; la presencia de estos datos puede influenciar en la efectividad del modelo. Uno de puntos más discutidos en la presencia de datos atípicos, si bien es cierto que en la actualidad existen muchos métodos o técnicas para reemplazar estos valores; como la mean, median, winsorizing o la misma eliminación. Estas técnicas van a depender de la lógica o los objetivos. Además, estas técnicas son más útiles cuando los datos tienen una distribución normal y que son perturbados por algunos valores extremos.

En muchos casos los datos atípicos en las variables tienen una justificación. Es decir, no se deben a un error en la construcción en la base de datos. Eliminarlos sería un error más que una solución. La eliminación o sustitución puede causar sesgo, disminución del tamaño en la muestra y puede afectar la correlación o distribución de la varianza. Según Guozhong et al. (1996) algo de ruido en los datos ayuda a desarrollar un mejor conjunto de entrenamiento para los algoritmos desarrollados, puesto que ayuda a reducir el ajuste excesivo del modelo, lo que a su vez ayuda a que el modelo se adapte mejor a conjuntos de datos nuevos e invisibles.

Es importante mencionar que en muchas ocasiones se necesitan incluir y generar otros datos a partir de datos originales, este proceso se hará siempre en cuando sea necesario o de acuerdo al algoritmo que se está usando. Finalmente nos aseguraremos que los datos a trabajar puedan cumplir con los siguientes requisitos: Exactitud, entereza, validez, consistencia, uniformidad, densidad y unicidad.

c. División de datos

Un buen modelo es aquel que puede hacer buenas predicciones con datos que no haya visto con anterioridad. Para probar si el modelo puede predecir correctamente estos datos nuevos utilizaremos el Splits, lo cual consiste en dividir los datos del conjunto en entrenamiento y prueba. Una buena evaluación en el conjunto de pruebas es un indicador útil del buen rendimiento del modelo en general, además es necesario tener las siguientes consideraciones.

- El conjunto de prueba debe ser lo suficientemente grande (30% del total de los datos)
- Usar diferentes conjuntos de prueba (para que el modelo no aprenda indirectamente)

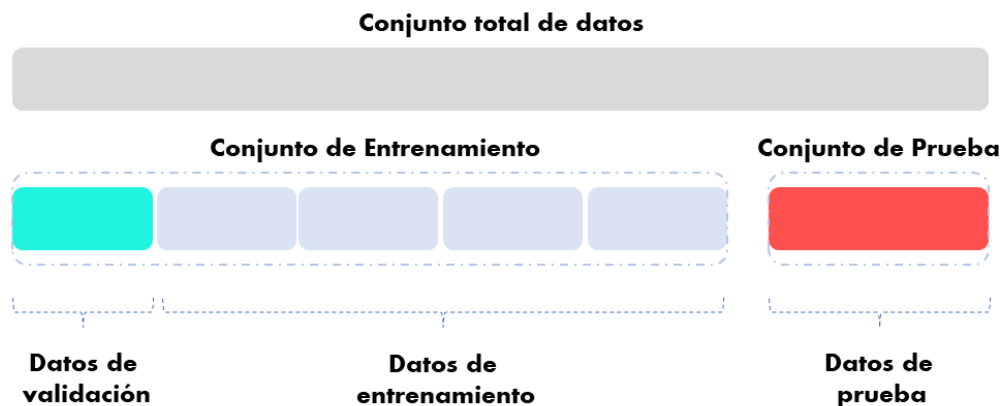


Figura 17. Esquema de división de datos.

Fuente: Elaboración propia

Conjunto de entrenamiento: Datos de entrenamiento + Datos de validación.

Conjunto de prueba: Un subconjunto para probar el modelo.

d. Estandarización de datos

La estandarización de los datos es un requisito importante en muchos estimadores de machine learning. Si las características individuales no se parecen más o menos a datos estándares distribuidos normalmente o tienen diferentes escalas, es posible que no funcione correctamente los modelos predictivos. En el caso actual para usar estimadores de machine learning en la generación de modelos predictivos, escalaremos las características para que se encuentren entre un valor mínimo y máximo, específicamente entre 0 y 1. Esto se puede lograr utilizando MinMaxScaler de la librería Scikit-Learn. Es importante mencionar que no todos los algoritmos necesitan esta etapa

4.2.5. Elección del algoritmo

Luego de haber realizado los pasos anteriores y teniendo en cuenta un panorama más claro de la situación, es momento de elegir los algoritmos que sean aplicables al caso presentado. Antes de hacerlo, es importante tener en consideración los siguientes puntos.

- Si el modelo cumple con nuestros objetivos.
- Qué tan explicable es el modelo.
- Que tan rápido es el modelo
- Cuanto tiempo se tarda en construir un modelo

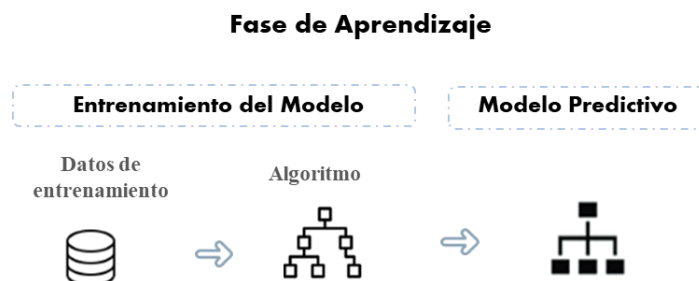


Figura 19. Esquema para el proceso de modelado.

Fuente: Elaboración propia

4.2.7. Evaluación del modelo

Se analiza la performance del modelo haciendo una comparación entre los resultados obtenidos con la predicción y la data set de prueba. Con este paso verificamos la capacidad predictiva del modelo. En caso sea necesario se retorna al paso anterior para la optimización de los hiperparámetros o interactuar con algún otro paso del flujograma.

Para llegar a los resultados finales se realizaron consideraciones especiales de acuerdo a la necesidad de cada técnica de machine learning. A continuación, se mencionará de manera general algunos de ellos que se tomaron en cuenta para llegar a los resultados obtenidos. En el desarrollo de los ejemplos se describirá con mayor detalle cada una de ellas.

a. Ajuste insuficiente frente a ajuste excesivo

Cuando el modelo intenta ajustar un polinomio a través de los datos, este tiene suficiente flexibilidad para dar cuenta casi perfectamente con todos ellos, sin embargo, también se puede ajustar a perturbaciones aleatorias de los datos. Se dice que este modelo tiene un sobre ajuste o también conocido como Overfit (Ver **figura 20**, derecha), es decir, el modelo aprende de los datos y de los errores aleatorios que lo contiene. Por otro lado, tenemos cuando el modelo intenta ajustar una línea que pase por todos los datos, sin embargo, los datos son intrínsecamente más complicados que una línea por lo que nunca podrá adaptarse la predicción a nuestros datos por la falta de flexibilidad. Se dice que el modelo tiene un bajo ajuste también conocido como Underfit (Ver **figura 20**, izquierda), es decir, el modelo tiene un alto sesgo. Para evitarlo, es una práctica común, cuando se realiza un experimento de aprendizaje automático (supervisado), mantener parte de los

datos disponibles como un conjunto de prueba, esto se logra realizando el Split o separación de datos.

Finalmente, lo que se busca es un buen modelo que tenga un error bajo tanto en los datos de entrenamiento como en los datos de prueba como muestra la **figura 21**.

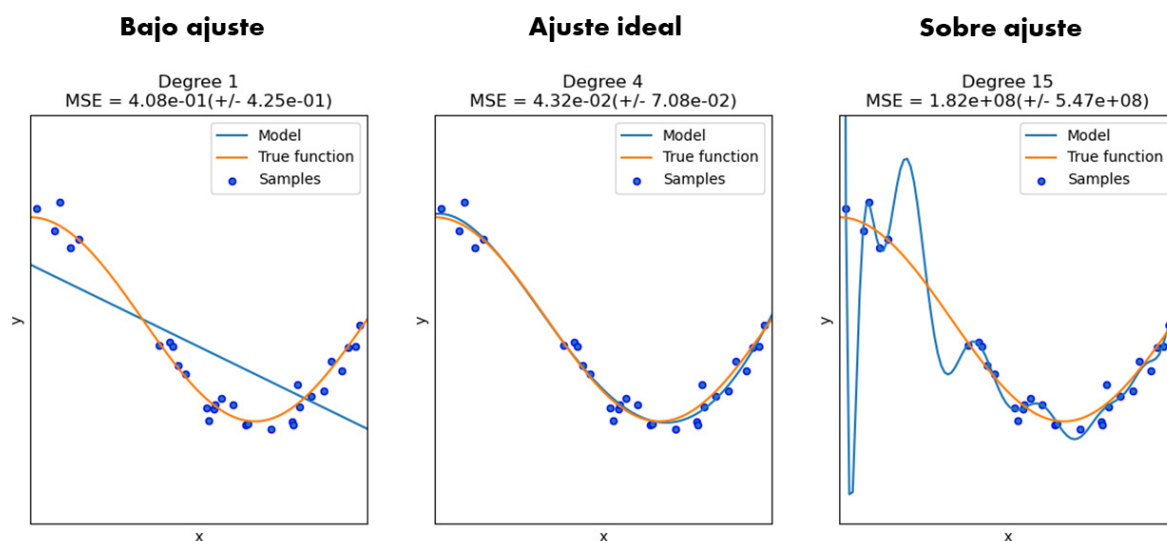


Figura 20. Diferentes modelos ajustados a un mismo dato

Fuente: Adaptado de (scikit-learn, 2021)

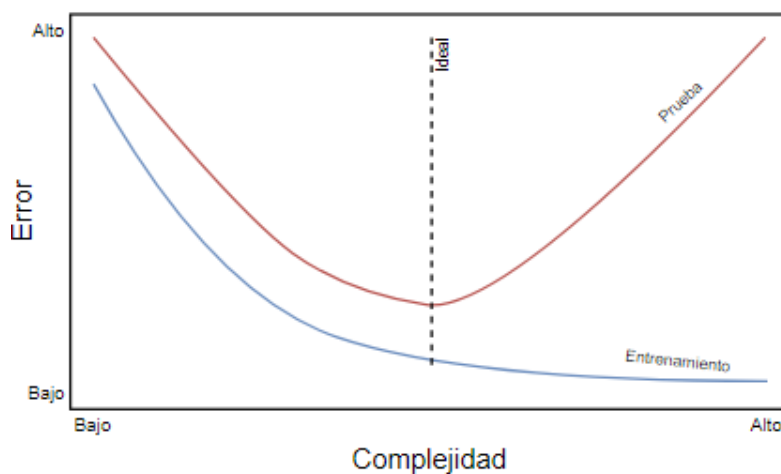


Figura 21. Comportamiento del error frente al Overfit

Fuente: Adaptado de (scikit-learn, 2021)

b. Validación cruzada (CV)

Esta técnica en lugar de generar un corte (Split), genera múltiples cortes, pero en tiempos diferidos. Una forma sencilla de verlo es imaginar que son múltiples cortes de entrenamiento y prueba. Esta técnica utiliza un parámetro K, el cual indica cuantas veces se repetirá la división. Cada vez que se repite el corte se construye un modelo de predicción y se evalúa. CV se encarga de recopilar estos resultados y presentarnos un promedio de las métricas obtenidas. Este método tiene la ventaja de usar todo nuestro conjunto de datos en el proceso de validación. Por lo tanto, la métrica de error es más cercana al error que nos mostraría una vez puesto en producción.

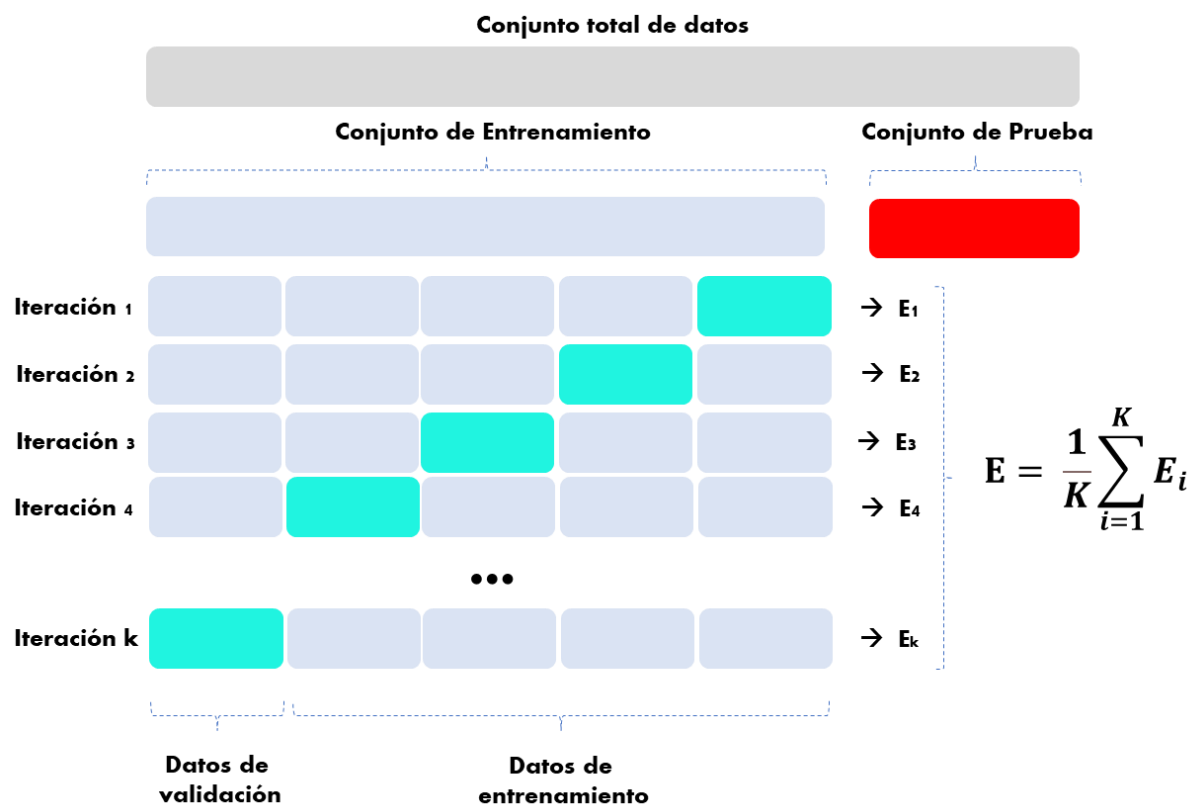


Figura 22. Proceso para realizar validación cruzada.

Fuente: Elaboración propia

c. Ajustes de hiperparámetros

Los hiperparámetros son parámetros que no se aprenden directamente en los estimadores. Estos más bien pasan como argumentos al constructor de las clases de estimador. Por ejemplo, en el método SVR incluyen C, Kernel y gamma; o en los métodos de ensamble como el Random Forest se necesita calcular el número óptimo de estimadores. En los ejemplos desarrollados (Ver cap. V) se explicará con más detalle el ajuste de hiperparámetros para cada algoritmo utilizado.

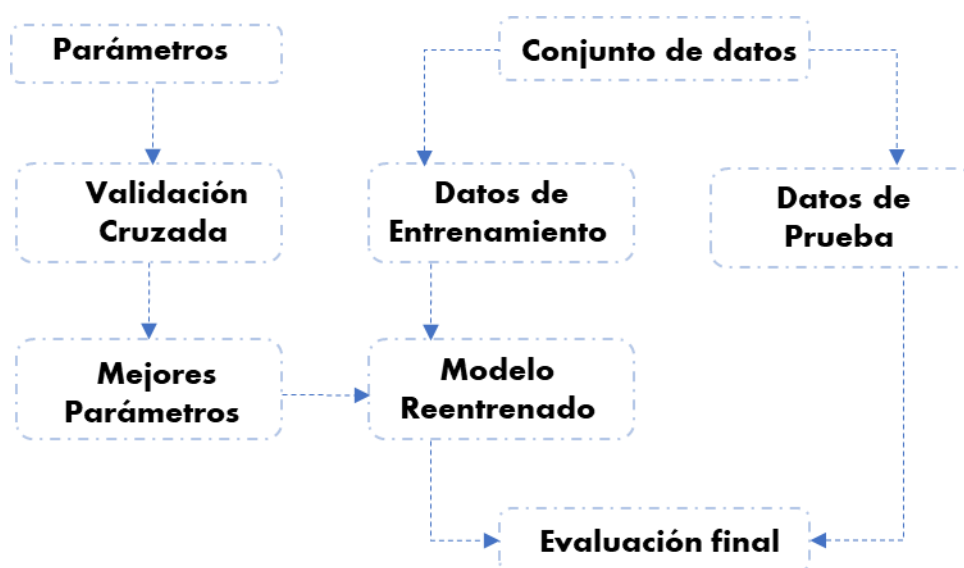


Figura 23. Diagrama para el ajuste de hiperparámetros con la técnica de búsqueda de cuadrícula.

Fuente: Adaptado de (scikit-learn, 2021)

4.2.8. Predicción

Luego de haber probado y evaluado el desempeño de diferentes modelos con diferentes algoritmos, se emplea el modelo más conveniente para predecir datos nuevos con la finalidad de probar una implementación o seguir mejorando los modelos.

4.3. Herramientas empleadas

En esta sección describiremos las principales herramientas utilizadas para el desarrollo de los ejemplos. Estas herramientas no son definitivas para toda aplicación. En la actualidad, existe una enorme cantidad de opciones disponibles que se pueden usar de acuerdo a las necesidades particulares.

4.3.1. Lenguaje de programación

a. Python

Python es un lenguaje de programación interpretado, orientado a objetos y de alto nivel con semántica dinámica. La sintaxis simple y fácil de aprender de Python enfatiza la legibilidad, lo que reduce los costos de mantenimiento del programa. Python admite módulos y paquetes, lo que fomenta la modularidad del programa y la reutilización del código.

Con el creciente número de aplicaciones de inteligencia artificial en diferentes industrias, se ha fomentado el uso Python debido a su fácil uso. A continuación, mencionaremos algunos puntos que hacen de Python uno de los más populares para el desarrollo de inteligencia artificial según Jean Francois Puget (2020) del departamento de aprendizaje automático de IBM:

- Una gran cantidad de librerías
- Una barrera de entrada baja
- Flexibilidad
- Independencia de la plataforma
- Legibilidad
- Buenas opciones de visualización
- Apoyo comunitario
- Creciente popularidad

4.3.2. Librerías

Una de las principales razones por las que el lenguaje de programación Python es el más utilizado en la inteligencia artificial es la gran cantidad y variedad de librerías para su uso. Una librería o biblioteca es un módulo o grupo de módulos publicados por diferentes fuentes que

incluyen un fragmento de código pre escrito que permite a los usuarios alcanzar alguna funcionalidad o realizar diferentes acciones. Estas librerías proporcionan elementos de diferentes niveles para que los desarrolladores no tengan que codificarlos desde el principio.

Cuando hacemos machine learning necesitamos procesar datos continuos, y para ello las librerías de Python nos permiten acceder, manejar y transformar los datos. Las librerías que mencionaremos son los que empleamos para el desarrollo de los ejemplos.

a. Scikit-Learn

Scikit-learn es una librería de Machine Learning de software gratuito para el lenguaje programación Python. En ella podemos encontrar varios algoritmos que realiza tareas de clasificación, regresión y agrupamiento; dentro de las cuales incluyen máquinas de vectores de soporte, bosques aleatorios, aumento de gradiente, k-means, etc. Scikit-learn está diseñada para inter operar con otras bibliotecas numéricas y científicas de Python, Numpy y SciPy.

b. Pandas

Esta librería se usa para análisis y estructuras de datos de alto nivel. Permite fusionar y filtrar datos, así como recopilarlos de otras fuentes externas como Excel, por ejemplo. Entre las funcionalidades más importantes tenemos:

- Visualización de DataFrame rápida y eficiente
- Tiene herramientas para leer y escribir datos en diferentes formatos como CSV, Microsoft Excel, base de datos SQL y el formato rápido HDF5
- Alineaciones inteligentes de datos y manejo integrado de datos faltantes
- Remodelación de datos
- Las columnas se pueden insertar y eliminar de las estructuras de datos
- Fusión y unión de conjuntos de datos
- Altamente optimizado su rendimiento
- Python con pandas se usa en una amplia variedad de dominios académicos y comerciales, que incluyen finanzas, neurociencia, economía, estadística, publicidad, análisis web y más

c. Numpy

Es una librería de Python que se utiliza para trabajar con matrices, además, tiene funciones para trabajar en el dominio de álgebra lineal y transformada de Fourier. La importancia de trabajar con *Numpy* radica en que esta nos proporciona un objeto de matriz que sea hasta 50 veces más rápido que las listas tradicionales de Python.

d. Matplotlib

Matplotlib es una librería completa que sirve para crear visualizaciones estáticas, animadas e interactivas en Python. Entre las principales funcionalidades que tiene podemos mencionar las siguientes.

- Crea gráficos de calidad
- Realiza figuras interactivas que pueden hacer zoom, desplazarse y actualizarse
- Se puede personalizar los estilos y diseños
- Se pueden exportar a diferentes formatos de archivo
- Se puede usar paquetes de terceros creados en Matplotlib

e. Seaborn

Seaborn es una biblioteca de visualización de datos de Python basada en Matplotlib. Proporciona una interfaz de alto nivel para dibujar gráficos estadísticos atractivos e informativos. A diferencia de Matplotlib Seaborn ayuda a generar gráficos más profesionales.

f. Plotly

La librería de gráficos Python de *Plotly* crea gráficos interactivos con calidad de publicación. Ejemplos de cómo hacer diagramas de líneas, diagramas de dispersión, gráficos de áreas, gráficos de barras, barras de error, diagramas de caja, histogramas, mapas de calor, subtramas, múltiples ejes, gráficos polares y gráficos de burbujas.

4.3.3. Programas y Software

a. Jupyter Notebook

Jupyter Notebook (anteriormente IPython Notebooks) es un entorno informático interactivo basado en la web para crear documentos de Jupyter notebook. El término "notebook" puede hacer referencia coloquialmente a muchas entidades diferentes, principalmente la aplicación web Jupyter, el servidor web Jupyter Python o el formato de documento Jupyter según el contexto. Un documento de Jupyter Notebook es un documento JSON, que sigue un esquema versionado y que contiene una lista ordenada de celdas de entrada/salida que pueden contener código, texto (usando Markdown), matemáticas, gráficos y texto enriquecidos, generalmente terminado con la extensión ".ipynb".

b. Anaconda (Distribución de python)

Es una distribución libre y abierta de los lenguajes Python y R, utilizada en ciencia de datos y aprendizaje automático (machine learning). Esto incluye procesamiento de grandes volúmenes de información, análisis predictivo y cómputos científicos. Está orientado a simplificar el despliegue y administración de los paquetes de software.

Las diferentes versiones de los paquetes se administran mediante el sistema de gestión de paquetes conda, el cual lo hace bastante sencillo de instalar, correr, y actualizar software de ciencia de datos y aprendizaje automático como puede ser Scikit-team, TensorFlow y SciPy.

c. Graphviz

Graphviz es un programa de visualización gráfica de fuente abierta. La visualización de gráficos es una forma de representar información estructural como diagramas de gráficos y redes abstractos. Tiene importantes aplicaciones en redes, bioinformática, ingeniería de software, diseño de bases de datos y web, aprendizaje automático y en interfaces visuales para otros dominios técnicos.

d. Microsoft Excel

Excel es un programa computacional incluido en el paquete Microsoft Office, y sirve para la creación, manejo y modificación de hojas de cálculo. Se puede utilizar en varios dispositivos y sistemas operativos.

e. SQL server

Microsoft SQL Server es un sistema de gestión de base de datos relacional (RDBMS) producido por Microsoft.

SQL es un lenguaje de computación para trabajar con conjuntos de datos y las relaciones entre ellos. Los programas de bases de datos relacionales, como Microsoft Office Access, usan SQL para trabajar con datos. A diferencia de muchos lenguajes de computación, SQL no es difícil de leer y entender, incluso para un usuario inexperto. Al igual que muchos lenguajes de computación, SQL es un estándar internacional reconocido por organismos de estándares como ISO y ANSI. SQL no se usa solo para manipular datos, sino también para crear y modificar el diseño de objetos de base de datos, como tablas.

Capítulo V: Desarrollo de modelos predictivos de los casos de aplicación

5.1. Modelo para la predicción del precio de oro

5.1.1. Compresión del problema

Las fluctuaciones en el precio de oro es sin duda uno de los riesgos para las empresas mineras productoras de oro. Ante una caída en los precios, muchos están obligados al cierre de sus operaciones debido a que no pueden controlar sus altos costos. El oro, por otro lado, casi siempre ha sido un refugio seguro para los inversionistas en alguna crisis económica, por lo que su precio sube. Sucede todo lo contrario cuando las economías vuelven a la normalidad, las inversiones apuntan a otros negocios con mayor costo de oportunidad. Bajo este escenario, las empresas están obligadas a realizar un análisis de variación futura de los precios para poder planificar sus proyectos.

Realizar un análisis del precio de oro para los próximos años, no solo es muy importante para las empresas, sino también para el estado. Con una tendencia alcista de los precios, el estado puede ayudar acelerar la puesta en marcha de proyectos, incentivar más inversión, incentivar más exploración, proyectar mayores ingresos por impuestos e incluso alentaría a la minería artesanal. Dado su importancia, Predecir el precio de oro puede ayudar a que las empresas, estado, traders, proveedores y todos los interesados proyectar sus actividades con indicadores positivos.

En este ejemplo, desarrollaremos un modelo predictivo con algoritmos de machine learning para estimación del precio de oro a partir de algunas variables que influyen en su fluctuación. Ante la necesidad de saber qué variables determinan en mayor medida la fluctuación de su precio, se identificará a través de los diferentes métodos la importancia y la influencia de estas variables que determinan su precio final.

Objetivo: Crear un modelo para estimar el precio del oro e identificar cuáles son las variables más determinantes para su variación.



Figura 24. Variación del precio de Oro 2015-2021.

Fuente: Elaboración propia

5.1.2. Aplicación del flujograma para el desarrollo del modelo

En esta sección se describe los pasos del flujograma que se aplicará en el desarrollo del modelo predictivo para estimar el precio del oro. También muestra algunas líneas de código ejecutados en los pasos del flujograma. En el capítulo 4.2 (Flujograma de trabajo) se describe los pasos que tomaremos en consideración al aplicar algoritmos de machine learning para generar modelos predictivos. Por lo tanto, mencionaremos algunos términos de manera general, puesto que la definición de estas se describió en el capítulo correspondiente (Ver cap. IV, 4.2).

5.1.2.1. Categorización del problema

Según lo expuesto en la Categorización del problema (Ver cap. IV, 4.2.1), para predecir el precio de oro aplicaremos el método de aprendizaje supervisado, debido a que se está trabajando con datos continuos y conocidos tanto para los datos de entrada como para los datos de salida. Por otro lado, el problema tendrá como dato de salida un valor numérico, por lo cual tendremos un problema de regresión. Con base en esta categorización, definimos las variables que utilizaremos en el desarrollo del modelo.

Cuando se hace un análisis de los factores que influyen en la fluctuación del precio de oro, no existe una sola causa que sea la responsable de su precio, sino por el contrario, nos encontramos con múltiples elementos y eventos que influyen en forma conjunto en su variación. Dentro de la cuales podemos tener la demanda, la producción, factores globales, sector financiero, etc. A través de un análisis paciente, entrevistas y revisión bibliográfica, se tomaron en consideración las variables más importantes. Para tal proceso se tomaron en consideración la accesibilidad de datos, tiempo y fuentes confiables. A continuación, se mencionan las variables predictoras que finalmente se tomaron en consideración.

Variable a predecir

- Precio de Oro

Variables predictoras

- Índice del precio del dólar
- Tasa de interés de la reserva federal de los Estados Unidos
- Producto bruto interno de los Estados Unidos
- Producto bruto interno de China
- Producción mundial del precio de Oro
- Costo de producción del Oro
- Inflación de los estados unidos
- Índice S&P 500
- Consumo mundial de oro
- Precio del petróleo.

5.1.2.2. Métricas de medición

Según los objetivos planteados y la categorización del problema, las métricas a tomar en consideración para este caso será el RMSE, MAE y el R^2 . Para saber la descripción de cada una de estas métricas revisar el apartado correspondiente (Ver cap. IV, 4.2.2).

5.1.2.3. Recolección de datos

Para el presente caso, los datos (Ver **Tabla 2**) se recogieron con una frecuencia diaria entre las fechas 01 de enero del 2015 al 22 de julio del 2021, con lo que finalmente se trabajará con 1620 datos para cada variable.

La obtención de todos los datos con una frecuencia diaria fue difícil de encontrar, debido a que algunos reportes solo brindan información mensual o trimestral. Para este caso en particular se tuvo que homogeneizar los datos con menor frecuencia, es decir, mantener los valores constantes en los periodos ajustados a otras variables que si tenían información con una frecuencia diaria. Finalmente, después de un cuidadoso trabajo en la selección de datos de forma separada para cada variable ubicadas en diferentes fuentes, se pudo cargar en una base de datos relacional (SQL) con la finalidad de facilitar en su posterior manejo.

5.1.2.4. Análisis de datos

Según las características de las variables y del problema, estamos ante un problema de pronóstico de series de tiempo. Para ello tomaremos como variable dependiente el precio de oro y el resto de variables como variables independientes, por lo cual tendremos la aplicación de algoritmos de machine learning con múltiples datos de entrada. Es importante mencionar que las variables múltiples deben respetar el orden de tiempo a la que pertenecen. El orden de los datos con la cual se trabajó se muestra en la **Tabla 3**.

Tabla 2. Variables para la predicción del precio de Oro.

Variable		Unidades	Frecuencia	Fuente de información
Nombre	Abreviatura			
Índice de precio de dólar	usdx	Uni	diaria	Investing
Tasa de interés reserva federal	tasa_FED	%	mensual	Fred.org
PBI china	pbi_china	%	trimestral	National Bureau of Statistics of
PBI estados unidos	pbi_usa	%	trimestral	Fred.org
Producción mundial de oro	produccion	tn	trimestral	Gold.org
Costo de producción de oro	costo_prod	U\$/tn	mensual	Gold.org
Inflación de estados unidos	infla_usa	%	mensual	Fred.org
Índice S&P 500	S&P_500	Uni	diaria	Investing
Consumo mundial de oro	consumo	tn	trimestral	Gold.org
Precio del petróleo	prec_petr	U\$/barril	diaria	Investing
Precio de oro	y	U\$/onza	diaria	Investing

Fuente: Elaboración propia

Tabla 3. Vista general de los datos

1	df.head(10)										
	usdx	tasa_FED	pbi_usa	pbi_china	produccion	costo_prod	infla_usa	S&P_500	consumo	prec_petr	y
Date											
01/01/2015	91.38	0.11	2.6	7.3	785.2	940	234.75	2058.95	1098.95	51.69	1186.8
02/01/2015	91.38	0.11	2.6	7.3	785.2	940	234.75	2046.25	1098.95	52.86	1337.7
05/01/2015	91.62	0.11	2.6	7.3	785.2	940	234.75	2016.00	1098.95	50.02	1353.5
06/01/2015	91.74	0.11	2.6	7.3	785.2	940	234.75	1994.50	1098.95	48.84	1362.2
07/01/2015	92.12	0.11	2.6	7.3	785.2	940	234.75	2019.50	1098.95	51.21	1353.6
08/01/2015	92.60	0.11	2.6	7.3	785.2	940	234.75	2055.00	1098.95	52.78	1351.6
09/01/2015	92.15	0.11	2.6	7.3	785.2	940	234.75	2035.25	1098.95	54.06	1357.4
12/01/2015	92.18	0.11	2.6	7.3	785.2	940	234.75	2022.50	1098.95	53.61	1373.2
13/01/2015	92.50	0.11	2.6	7.3	785.2	940	234.75	2016.00	1098.95	53.53	1374.5
14/01/2015	92.35	0.11	2.6	7.3	785.2	940	234.75	2007.50	1098.95	52.14	1364.5

Fuente: Elaboración propia

Visualización de datos

Una herramienta muy interesante para la visualización de datos es la matriz de Pearson. Esta nos ayuda a identificar la correlación, ya sea directa o indirecta, de las variables predictoras con la variable a predecir. Por otro lado, identificar altas correlaciones entre predictores ayuda a analizar las variables en cuestión para evitar problemas de colinealidad al generar el modelo.

En la **Figura 25** se observa que el *costo de producción*, *el consumo*, *la inflación en los Estados Unidos* y el *índice S&P 500* son los que mayormente influyen en la fluctuación del precio del Oro. Las otras variables que influyen en menor medida es el *índice del dólar* y la *tasa de interés de la reserva federal* de los Estados Unidos. Esta herramienta también ayuda identificar a las predictores más relevantes de una manera fácil y rápida, y que en muchos casos sirve para realizar el primer filtro. Por otro lado, las variables que podrían generar el problema de colinealidad son la *inflación de los estados unidos* y el *índice S&P_500* debido a su alta correlación entre ambas variables.

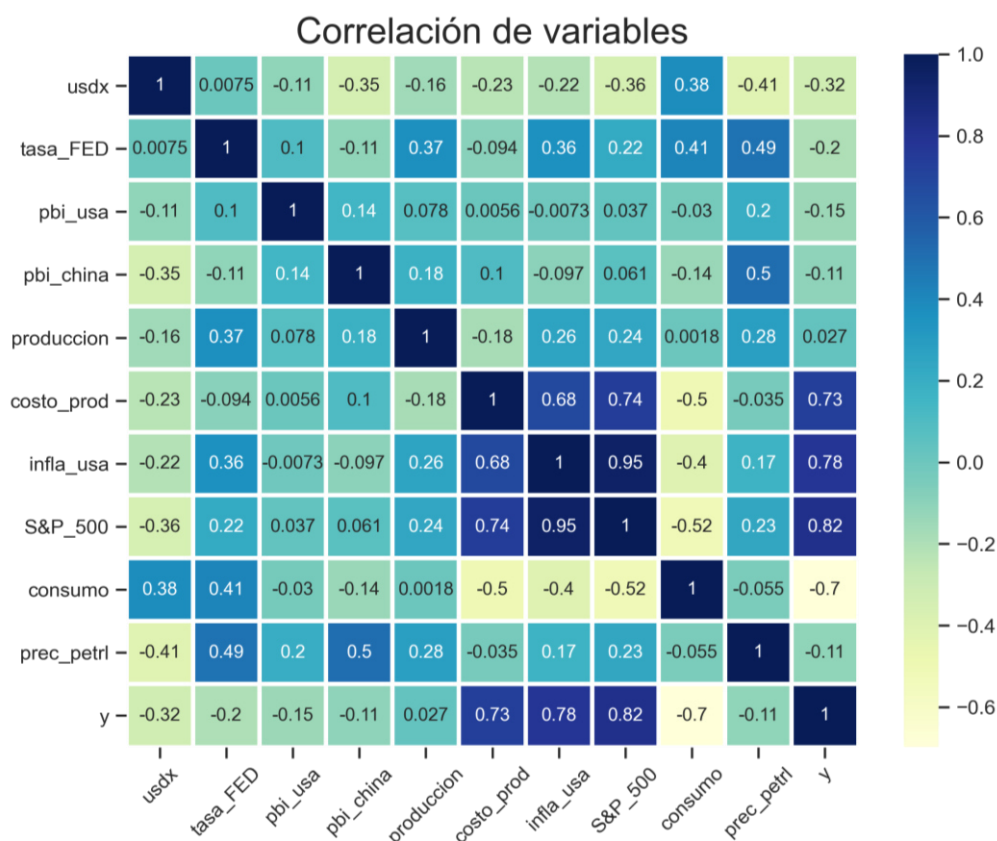


Figura 25. Correlación de Variables del Precio de Oro.

Fuente: Elaboración propia

En la **Figura 26** se muestra gráficamente la relación entre el índice *S&P 500* y la variación del *precio de oro*(y), además, vemos en el centro el gráfico la dispersión de ambas variables, así como los histogramas de cada variable en la parte superior y lateral derecho. En esta gráfica la distribución de datos de las variables no posee una distribución normal. Por el contrario, tienen una alta desviación estándar por lo que es necesario realizar el preprocesamiento de los datos antes de generar el modelo. Otra utilidad de la visualización del histograma es que nos permite ir identificando que algoritmos utilizar para el tipo de dato que tenemos. Algunos algoritmos tienen un desempeño muy bueno en datos de alta desviación estándar y con presencia de valores atípicos, mientras que algunos algoritmos son muy sensibles ante la presencia de datos atípicos.

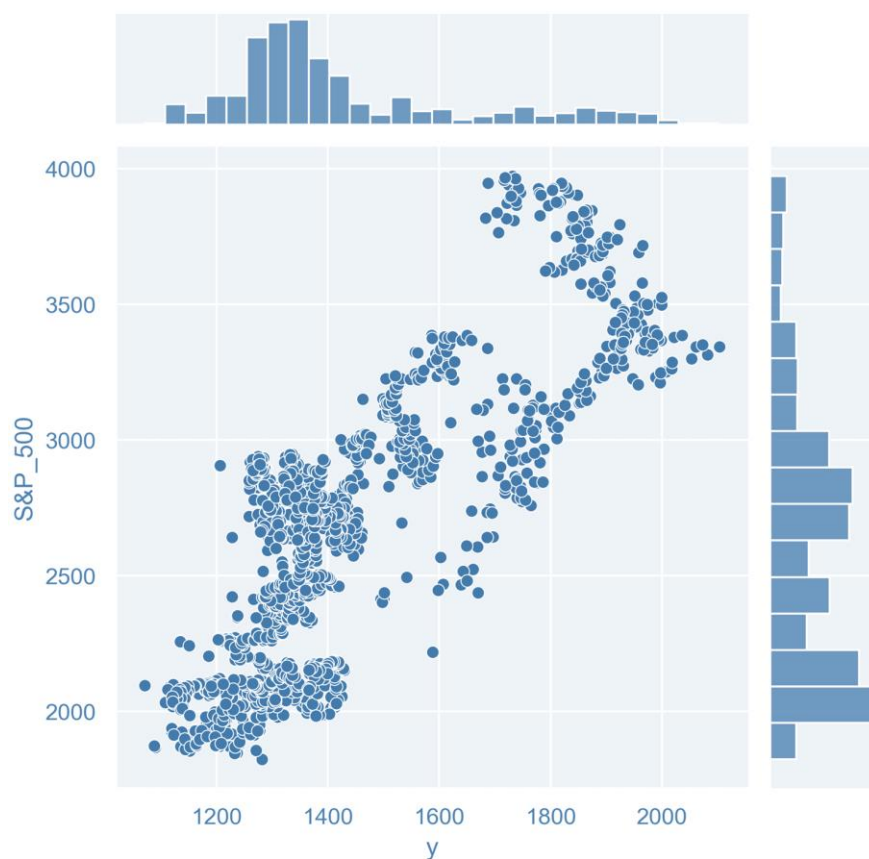


Figura 26. Histograma y gráfica de dispersión, Precio de Au Vs S&P 500.

Fuente: Elaboración propia

Limpieza de datos

Una de las herramientas que se suele usar con bastante frecuencia para la limpieza de datos es el diagrama de caja, este es un método para representar gráficamente grupos de datos numéricos a través de sus cuartiles. El cuadro se extiende desde los valores del cuartil Q1 a Q3 de los datos, con una línea en la mediana (Q2). Los bigotes se extienden desde los bordes de la caja para mostrar el rango de los datos. De forma predeterminada, no se extienden más de $1.5 \times \text{IQR}$ ($\text{IQR} = Q3 - Q1$) desde los bordes de la caja, terminando en el punto de datos más lejano dentro de ese intervalo. Todos los valores que están fuera de los bigotes son considerados datos atípicos.

```
1 import numpy as np
2 np.log(1+df).boxplot(column=['tasa_FED', 'pbi_usa', 'pbi_china', 'prec_petr1'],
3                       figsize=(20,12),
4                       fontsize=20)
5
```

La **Figura 27** muestran un diagrama de cajas y bigotes a partir de las variables. Para una mejor visualización se hizo una representación logarítmica de los datos.

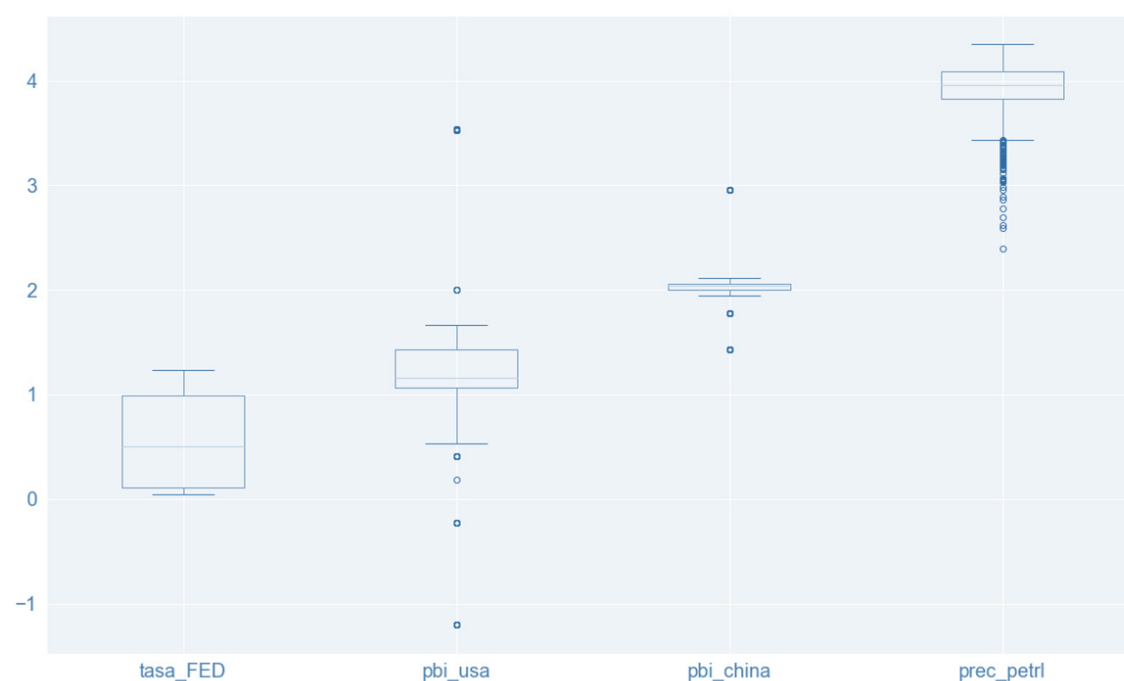


Figura 27. Diagrama de cajas de las variables principales

Fuente: Elaboración propia

Según este criterio se observa que algunas variables poseen datos atípicos. Tanto el PBI de Estados Unidos como el PBI de China sufrieron una enorme caída durante el primer año de la pandemia (2020), pero en el mismo periodo tuvieron un enorme crecimiento con respecto al año 2020. Las otras variables que también tienen datos atípicos es el precio del petróleo y el precio de oro, este tuvo un pico máximo durante la pandemia. Sin embargo, estos valores atípicos de nuestras variables tienen una justificación, es decir, no se deben a un error de construir la base de datos, eliminar estos datos atípicos sería un error más que una solución.

La eliminación o sustitución puede causar sesgo, disminución del tamaño de la muestra o puede afectar la correlación o distribución de la varianza. Como mencionamos anteriormente (Ver cap. IV, 4.2.4, b.), tener algo de datos ruidosos pueden ayudar obtener mejores resultados de predicción con datos nuevos.

División de datos

Los datos obtenidos para el desarrollo del modelo se dividirán en tres conjuntos importantes. El Primero es el conjunto de datos de entrenamiento, segundo es el conjunto de datos de validación cruzada y tercero los datos de prueba. Los datos de entrenamiento es el conjunto más grande y su función principal es entrenar el modelo, los datos de conjunto de validación cruzada se utilizan para ajustar algunos modelos, y los datos de prueba sirven para ver el rendimiento de los modelos desarrollados. La división empezará asignando el 75% al conjunto de entrenamiento y el 25% al conjunto de prueba. A su vez del 75% se divide en 60% para entrenamiento, mientras que el 15% para validación cruzada y finalmente el 25% se mantiene como conjunto de prueba.

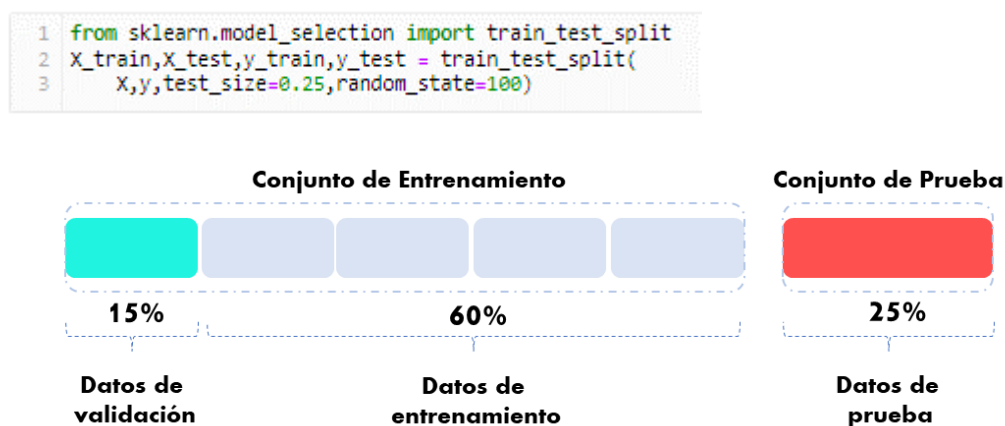


Figura 28. Esquema de división de datos para el Caso 1.

Estandarización de datos

En el caso actual para usar estimadores de machine learning en la predicción del precio del oro, escalaremos las características para que se encuentren entre un valor mínimo y máximo, específicamente entre 0 y 1, esto se puede lograr usando MinMaxScaler de la librería Scikit-Learn. En el caso del algoritmo árbol de decisión no será necesario el preprocesamiento de datos.

5.1.2.5. Elección de algoritmos

El alcance de esta investigación indica que es de tipo exploratorio, con base en este tipo de investigación probaremos algunos algoritmos que hay bajo la categorización del problema. Para ello, recurriremos a tres algoritmos: regresión lineal, árbol de regresión y SVR.

a) Regresión lineal

Este modelo nos permite encontrar la relación que existe entre una variable dependiente y una independiente, así como predecir el valor esperado de una variable aleatoria (y) cuando(x) toma una variable. La función que representa a este modelo simple es la siguiente:

$$y = w_0 + w_1x$$

En el caso de la predicción del precio de oro (y) estamos tomando muchas variables independientes, entonces estamos ante una regresión lineal múltiple. La notación matemática para representarlo es la siguiente:

$$y(w, x) = w_0 + w_1x_1 + w_2x_2 + w_3x_3 \dots \dots \dots w_px_p$$

A lo largo del módulo, designamos el vector $w = (w_1, w_2 \dots w_p)$ como coeficientes y w_0 como el parámetro de error, la (y) nuestra variable dependiente o a predecir, además, x_i como las variables independientes.

La librería Scikit Learn ofrece LinearRegression el cual ajusta a un modelo lineal (Ver **figura 16**) con coeficientes $w = (w_1, w_2 \dots w_p)$ para minimizar la suma residual de cuadrados entre los objetivos observados con el conjunto de datos y los objetivos predichos por la aproximación lineal. Matemáticamente, resuelve un problema de la forma:

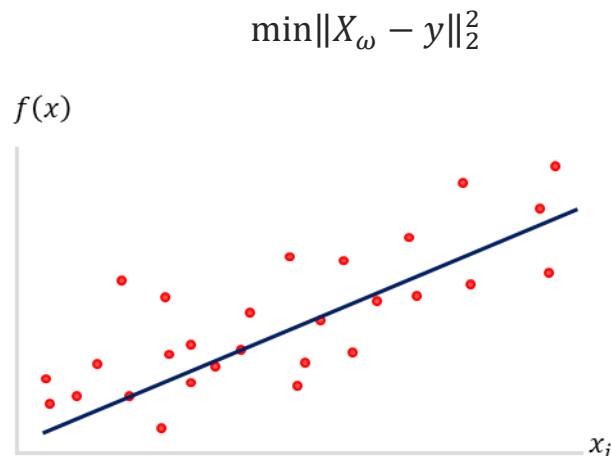


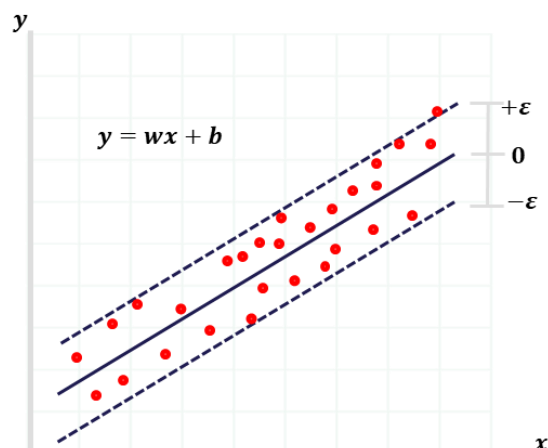
Figura 29. Ilustración Modelo Regresión Lineal

Fuente: Elaboración propia

b) Support Vector Regression (SVR)

SVR es un algoritmo de regresión basada en las máquinas de soporte vectorial utilizado para clasificar elementos de diferentes conjuntos. A continuación, explicaremos cómo funciona este el algoritmo SVR para realizar modelos predictivos de regresión, es importante señalar que este algoritmo funciona para datos lineales y no lineales.

Imaginemos que existe un conjunto de datos como el que se presenta a continuación, donde se observa que los datos presentan un comportamiento similar a un modelo lineal.



El primer paso que debemos realizar es encontrar el modelo lineal que mejor describa el comportamiento de este conjunto de datos. En el contexto del SVR a nuestra línea la llamaremos hiperplano. En el contexto de problemas donde los datos no son lineales el hiperplano que mejor representa el comportamiento de los datos será más complejo que una sencilla línea, pero la idea central seguirá siendo la misma.

La ecuación que represente esta línea será de la forma siguiente:

$$y = wx + b$$

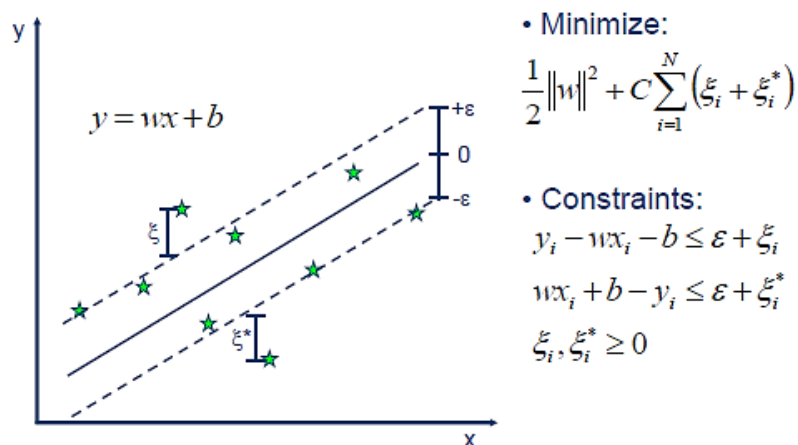


Figura 30. Ilustración del Modelo SVR

Fuente: Adaptado de *An Introduction to Data Science*, por Yasad, 2020

http://www.saedsayad.com/data_mining_map.htm.

El siguiente paso consiste en hallar las bandas paralelas al hiperplano donde es necesario que entre ambas cubran la mayor cantidad de datos. Estas bandas son conocidas como vectores de apoyo o soporte.

En la **Figura 30** se puede notar las bandas paralelas al hiperplano, estas bandas están representadas por las ecuaciones siguientes:

$$y = wx + b + \varepsilon$$

$$y = wx + b - \varepsilon$$

Como se puede notar en la gráfica anterior, es posible que las bandas propuestas no cubren todos los datos existentes. Estos datos son conocidos como errores y son datos que el algoritmo tomará en cuenta para mejorar su desempeño. En nuestro ejemplo donde se presenta datos lineales se utiliza la siguiente fórmula que controla el comportamiento.

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N (\xi + \xi^*)$$

Maximizar el margen

Minimizar el error

En la fórmula anterior se ve otras variables donde: w es la magnitud del hiperplano, C es una constante ($C > 0$) que se encarga de controlar el equilibrio entre la regularidad de la función y cuanto toleraremos las desviaciones fuera de las bandas de soporte y las variables X_i , las cuales controlan el error cometido por la función al aproximar las bandas.

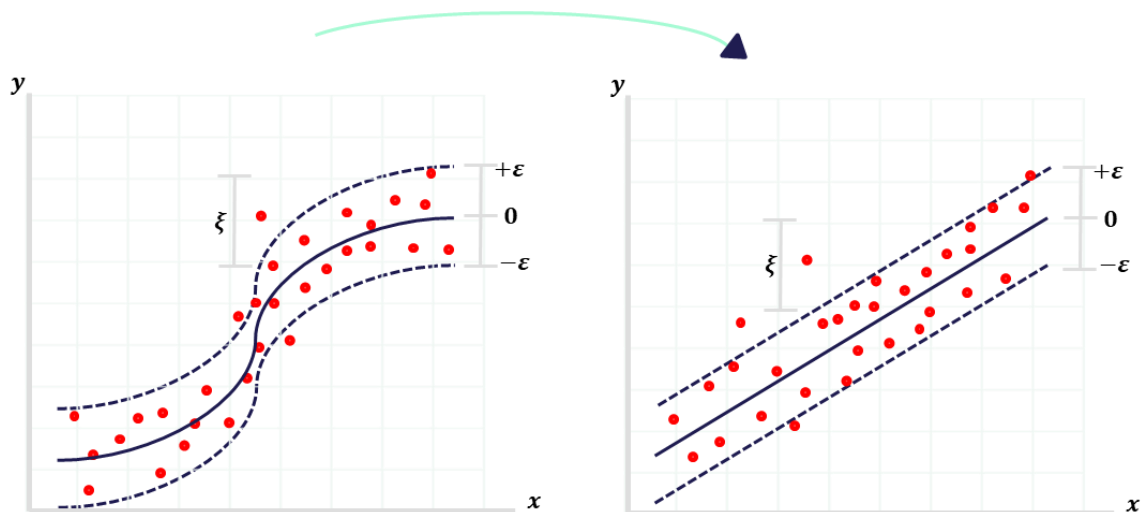
El ejemplo anterior representa un caso de un problema lineal, la cual el SVR está dado por:

$$y = \sum_{i=1}^N (a_1 - a_i^*) \cdot \langle x_i, x \rangle + b$$

En caso tengamos un problema no lineal, la función Kernel transforma los datos en una característica de espacio dimensional más alto para hacer posible ejecutar la separación lineal.

$$y = \sum_{i=1}^N (a_1 - a_i^*) \cdot \langle \varphi(x_i), \varphi(x) \rangle + b$$

$$y = \sum_{i=1}^N (a_1 - a_i^*) \cdot K(x_i, x) + b$$



Donde los Kernel(K) son:

Polinomial: $K(X_i, X_j) = (X_i, X_j)^d$

Gaussian Radial Basic Function (RBF): $K(X_i, X_j) = \exp\left(-\frac{\|X_i - X_j\|^2}{2\sigma^2}\right)$

c) Árbol de regresión

Los árboles de regresión son métodos predictivos basados en la segmentación, en donde el conjunto de datos se divide en diferentes categorías. Los datos que estén dentro de cada intervalo se les asigna directamente un valor establecido, en lugar de una etiqueta perteneciente a la categoría donde esté ubicado el dato. Lo cual, es la principal diferencia con respecto a su contraparte los árboles de clasificación.

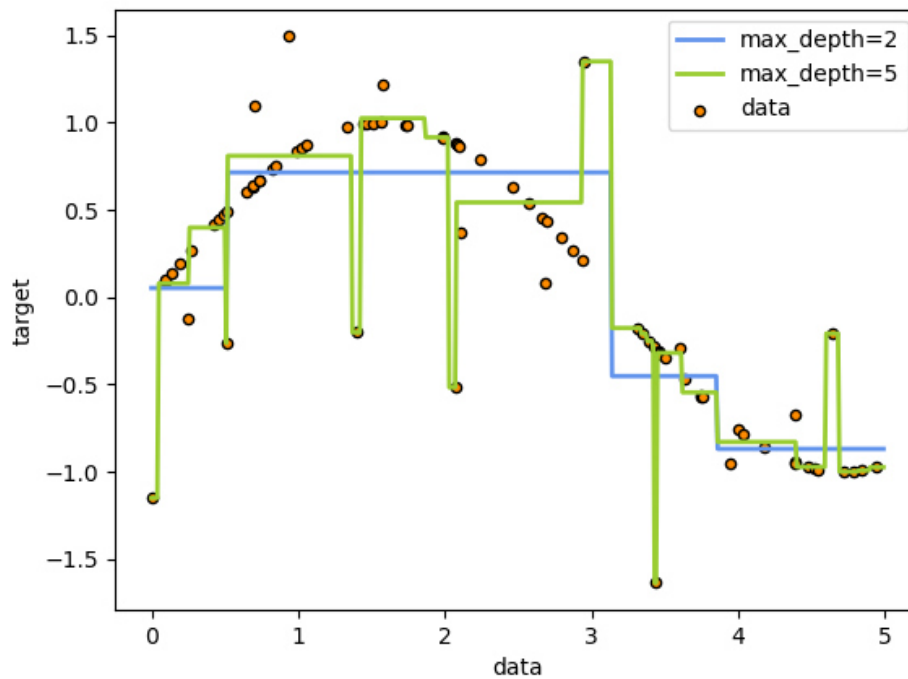


Figura 31. Ilustración del Modelo Árbol de Regresión

Fuente: Adaptado de (scikit-learn, 2021)

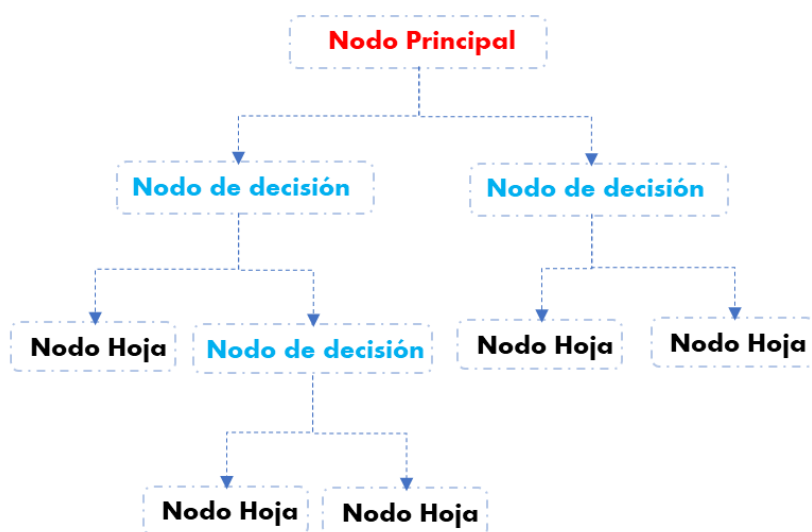
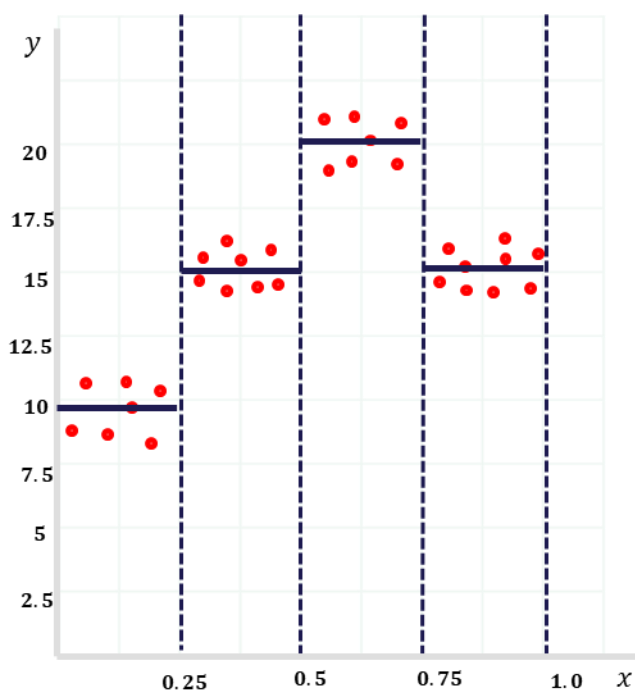


Figura 32: Partes de un diagrama de árbol de regresión.

Fuente: Elaboración propia

A continuación, se presentará un ejemplo más simple y detallado.

En la gráfica siguiente vemos un conjunto de datos diferentes donde la variable X se parte en intervalos más pequeños representados con líneas punteadas azules. En el ejemplo simple el agrupamiento de datos nos ayuda a identificar los siguientes intervalos.



Intervalo 1: $0 \leq x < 0.25$

Intervalo 2: $0.25 \leq x < 0.5$

Intervalo 3: $0.5 \leq x < 0.75$

Intervalo 4: $0.75 \leq x < 1.0$

Intervalo 1: $y = 10$

Intervalo 2: $y = 15$

Intervalo 3: $y = 20$

Intervalo 4: $y = 15$

Dentro de los intervalos del eje X también se puede identificar el valor más cercano de Y en cada una ellas. En la gráfica, estos valores promedios de Y están marcadas con líneas Azules gruesas. Sobre la base de estas observaciones podemos construir un conjunto de reglas basadas en el valor de X que tiene la variable Y representarlas en forma de un diagrama (Ver **figura 33**) como se muestra a continuación.

Este diagrama se conoce como árbol de regresión y se utiliza para asignarla un valor numérico a una entrada X de la cual no se conocía su valor de salida Y previamente.

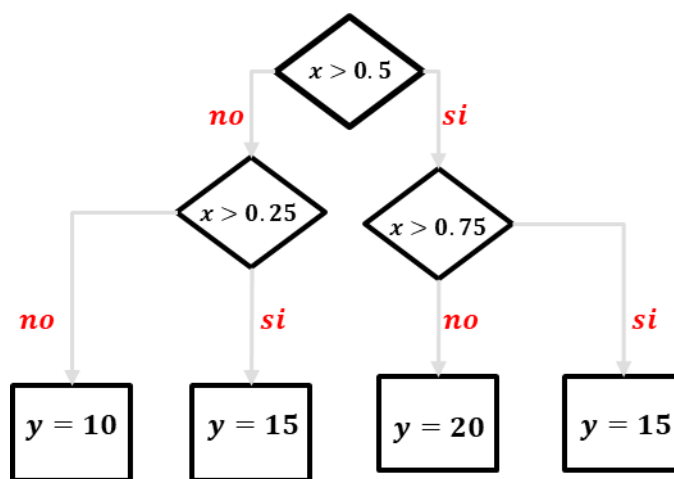


Figura 33. Diagrama del Modelo Árbol de regresión

5.1.2.6. Modelado

La finalidad de presentar tres modelos diferentes es para comparar el desempeño en cada uno de ellos con respecto a la base de datos recogidos. En los siguientes apartados se describirá los procesos que se llevaron a cabo para el entrenamiento de los modelos con cada uno algoritmos escogidos.

a) Modelo lineal

El modelo *regresión lineal* sirve con frecuencia para saber el rendimiento inicial que tiene los algoritmos de machine learning ante una base de datos nueva. Además, este modelo nos ayuda a tener una mejor idea de los datos que poseemos para definir o seguir trabajando en su preprocesamiento. Su uso es simple, de fácil entendimiento y no requiero ajustes adicionales como otros algoritmos.

Luego de haber intentado la optimización de variables con diferentes métodos como el *PCA*, *filtro* y *características recursivas RFE*, se optó por la utilización de todas las variables predictoras debido a que los resultados obtenidos fueron los mejores.

```

1 # ENTRENAMIENTO
2 from sklearn.linear_model import LinearRegression
3 reg = LinearRegression().fit(X_train,y_train)
4
5 # PREDICCIÓN
6 y_train_pred = reg.predict(X_train)
7 y_test_pred = reg.predict(X_test)
8
9 # EVALUACIÓN
10 from sklearn.metrics import r2_score
11 r2_score(y_train, y_train_pred)
12 r2_score(y_test, y_test_pred)

```

En la **Figura 34** se muestra el ajuste del modelo entrenado con el algoritmo regresión Multilineal comparado con el modelo real. El ajuste del modelo predictivo es bastante alto con el comportamiento real del precio de oro, excepto en los picos altos. Sin embargo, a pesar de la simplicidad de algoritmo y su fácil uso, se puede obtener buenos resultados con métricas de R^2 mayor a los 0.91, pero con un *RMSE* de 31.49 y un *MAE* de 17.75. Por lo que intentaremos obtener mejores resultados utilizando los otros algoritmos escogidos.

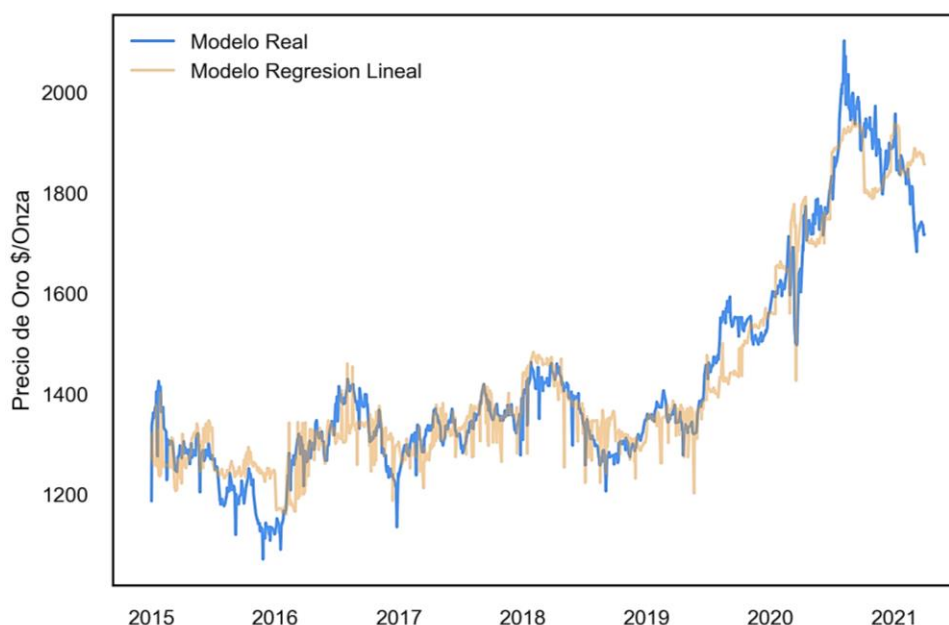


Figura 34. Modelo Multilineal para la predicción del precio de Oro.

Fuente: Elaboración propia

b) Support Vector Regressor (SVR)

El modelo SVR es más complejo que el modelo lineal debido a que en esta se necesita los ajustes de los hiperparámetros que posee. En la descripción de algoritmo para este ejemplo (Ver cap. V, 5.1.2.5, b.) se describe el uso de parámetros como *epsilon=define los vectores de soporte*, *C=constante* y *K= función Kernel*. Para desarrollar el ajuste en estos parámetros se usará los datos de validación cruzada. Esto se hace utilizando el método de *Búsqueda de cuadrícula* (GridsearchCV). El método consiste en probar un rango de valores de *C* y *gamma* en el conjunto de entrenamiento y validación para cada uno de los *Kernels*, el que tiene la puntuación más alta será elegido como parámetro apropiado.

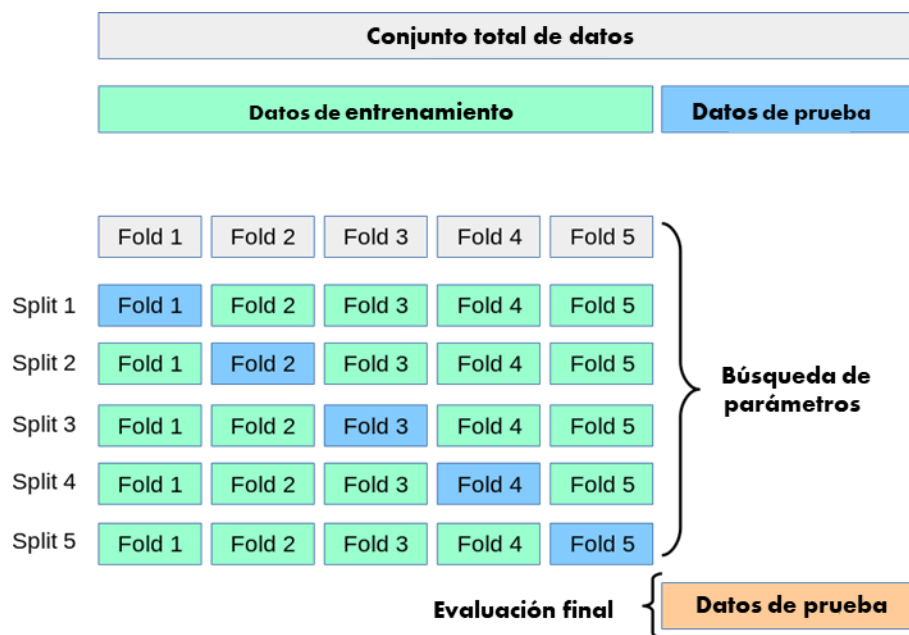


Figura 35. Esquema de validación cruzada

Fuente: Adaptado de (scikit-learn, 2021)

De manera más específica, los parámetros que exploraremos son los siguientes: para un *Kernel Lineal* tendremos un rango de $C = [1, 10, 100, 1000]$, y la segunda será para un *Kernel RBF* con un rango de $C = [1, 10, 100, 100]$ y valores *gamma* en $[0.01, 0.001]$. Finalmente, el valor de *Épsilon* se tomará como base 0.01 y el número de grupos en la que dividiremos los datos de entrenamiento será $K=5$.

```

1 from sklearn.model_selection import GridSearchCV
2 from sklearn.svm import SVR
3
4 parameters = [{'kernel':('linear', 'rbf'), 'C':[1, 10, 100, 1000], 'gamma':[0.01, 0.001]}]
5 K = 5
6 svr = GridSearchCV(SVR(epsilon = 0.01), parameters, cv = K)
7 svr.fit(X_train, y_train.reshape(-1))
8
9 print(svr.best_params_)
{'C': 1000, 'gamma': 0.01, 'kernel': 'rbf'}

1 print(svr.best_score_)
0.9625200210240967

```

Los resultados obtenidos arrojan como parámetros óptimos los siguientes datos: $C=1000$, $\gamma=0.01$, $\text{Kernel}=RBF$, además de incluir la $\epsilon=0.01$ y el mejor score en este proceso de optimización para estos parámetros fue de 0.96. Una vez obtenida estos parámetros óptimos para el algoritmo SVR según el criterio de búsqueda de cuadrícula, se utiliza para generar el modelo.

En la **Figura 36** se muestra el ajuste del modelo entrenado con el algoritmo SVR comparado con el modelo real. El ajuste es mucho mejor con respecto al modelo lineal. Los resultados obtenidos luego de optimización de los hiperparámetros mejoraron considerablemente respecto a los resultados que se había obtenido usando los hiperparámetros de manera aleatoria. Inicialmente, se hizo prueba de manera separada para los tres estimadores de Kernel obteniendo los siguientes resultados: para $K=\text{lineal}$ se obtuvo un $\text{Score}=0.890$; para $K=\text{Pol}$ se obtuvo un $\text{Score}=0.920$; para $K=\text{rbf}$ se obtuvo un $\text{Score}=0.940$; finalmente, con el uso de los hiperparámetros óptimos se obtuvo un $\text{Score}=0.9468$.

La utilización de *Búsqueda de cuadrícula* (GridsearchCV) no es el único método que existe para el ajuste de hiperparámetros. Existen otros métodos que también son conocidos y populares en su utilización como la *Búsqueda aleatoria* y de *Búsqueda de bayes*. El uso de estas diferentes técnicas dependerá generalmente del tiempo, el número de combinaciones, objetivo de trabajo y costo computacional.

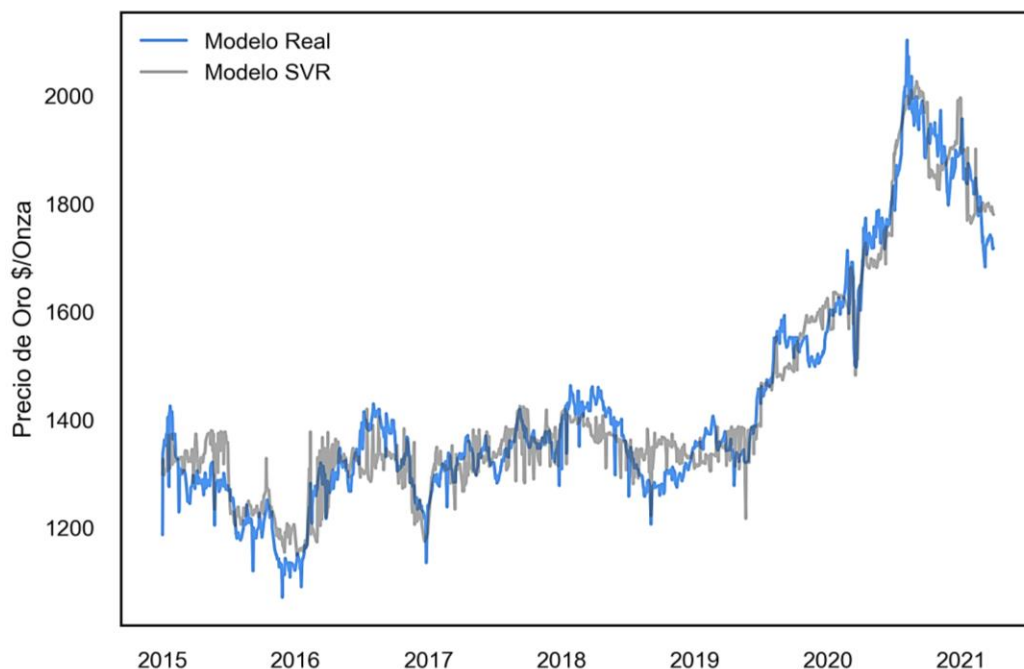


Figura 36. Modelo SVR para la predicción del precio de Oro.

Fuente: Elaboración propia

c) Árbol de regresión

Cuando desarrollamos modelos predictivos usando *árboles de decisión*, nos encontramos con diferentes bondades que tiene esta técnica. Dentro de las cuales está su fácil interpretación, modelos basados en un solo árbol, no necesita distribución específica, requiere menos limpieza de datos, no requiere estandarización, no se ven influenciados con datos atípicos, etc. Sin embargo, cuando se empieza con la construcción del modelo uno se puede encontrar con el *overfitting* o sobre entrenamiento, esto es debido a que durante el proceso de entrenamiento el error tiende a reducir rápidamente para ajustar las observaciones. Como consecuencia de ello, la capacidad predictiva se reduce con datos nuevos. La razón de este comportamiento radica en la facilidad con la que los árboles se ramifican adquiriendo estructuras muy complejas, que incluso pueden terminar ajustándose perfectamente con los datos de entrenamiento. Para evitar este tipo de problema usaremos las técnicas que ayudan a limitar el tamaño del árbol: *Pre-poda (pre-pruning)* y el *Post-poda (post-pruning)*.

Inicialmente se hizo una prueba de manera aleatoria obteniendo los resultados mostrados al final del código mostrado. Para el desarrollo de este modelo inicial no se tomaron en consideración todos los hiperparámetros necesarios para obtener un buen resultado. En este caso no sabemos si la profundidad del árbol escogido aleatoriamente es el óptimo, o los hiperparámetros usados de manera predeterminada son los correctos. Para tratar de mejorar los resultados utilizaremos las técnicas mencionadas anteriormente con la finalidad de obtener un árbol óptimo.

```

1 from sklearn.tree import DecisionTreeRegressor
2 from sklearn.metrics import r2_score
3 from sklearn.metrics import mean_squared_error
4 from sklearn.metrics import mean_absolute_error
5
6 # ENTRENANDO
7 tree_2 = DecisionTreeRegressor(max_depth=3)
8 reg = tree_2.fit(X_train,y_train.reshape(-1))
9
10 # PREDICCIÓN
11 y_train_pred = tree_2.predict(X_train)
12 y_test_pred = tree_2.predict(X_test)
13
14 # EVALUACIÓN
15 print("Score", r2_score(y_test, y_test_pred))
16 error = mean_squared_error(y_test,y_test_pred,squared=False)
17 print("RMSE {}".format(error))
18 error3 = mean_absolute_error(y_test,y_test_pred)
19 print("MAE {}".format(error3))

```

Score 0.9178041361898702
RMSE 59.83536652625277
MAE 47.73146635796088

Pre-pruning (pre – poda)

Pre-Pruning, también conocido como parada temprana o poda hacia delante, es una técnica que impide la generación de ramas no significativas. Se suele aplicar esta técnica antes de la construcción de un árbol. Esta utiliza algunas condiciones para decidir cuándo debe terminar la división de alguna de las ramas de forma prematura a medida que se genera el árbol.

La técnica de pre poda implica ajustar los hiperparámetros del modelo del árbol de decisión antes de la canalización de entrenamiento. Para este fin usaremos la implementación *sklearn* de la técnica *GridSearchCV* para encontrar el mejor conjunto de hiperparámetros en un modelo de árbol de decisión.

En la librería Scikit-Learn, que se está empleando para el desarrollo de este trabajo, existe una clase *DecisionTreeRegressor* dentro del módulo *sklear.tree* que nos permite entrenar árboles de decisión para problemas de regresión. En la cual, podemos encontrar hiperparámetros por

defecto como: *criterion*, *splitter*, *max_samples_split*, *min_samples_leaf*, *min_weight_fraction*, *max_features*, *random_state*, *max_leaf_nodes*, *min_impurity_decrease*, *min_impurity_split* y el *ccp_alpha*.

Para el caso del pre-pruning utilizaremos como condición valores para los siguientes hiperparámetros: *max_depth*, *max_samples_leaf*, *max_samples_split*. De todos los hiperparámetros estos son los que ayudan a evitar el crecimiento donde:

- ***Max_depth***: es la máxima profundidad que puede alcanzar un árbol.
- ***Max_samples_leaf***: es el número mínimo de observaciones que debe tener un nodo para que pueda dividirse.
- ***Max_samples_split***: número mínimo de observaciones que debe de tener cada uno de los nodos hijos para que se produzca la división.

A continuación, mostraremos los códigos ejecutados para la obtener el *Pre-pruning*

```
1 from sklearn.model_selection import GridSearchCV
2 from sklearn.tree import DecisionTreeRegressor
3
4 param_grid = {
5     "max_depth": [2,3,5],
6     "min_samples_split": [2,5,7,10],
7     "min_samples_leaf": [1,2,5]
8 }
9 tree = GridSearchCV(
10     estimator = DecisionTreeRegressor( random_state = 42),
11     param_grid = param_grid,
12     cv = 4,
13     refit = True,
14     return_train_score = True
15 )
16
17 tree.fit(X_train, y_train.reshape(-1))
```

```
1 tree_final = tree.best_estimator_
2 print(f"Profundidad del árbol: {tree_final.get_depth()}")
3 print(f"Número de nodos terminales: {tree_final.get_n_leaves()}")
4 print(tree.best_params_)
```

```
Profundidad del árbol: 5
Número de nodos terminales: 32
{'max_depth': 5, 'min_samples_leaf': 2, 'min_samples_split': 2}
```

El desempeño de la predicción obtenidos luego entrenar el modelo con los hiperparámetros óptimos con la técnica *pre-pruning* y los datos de prueba son los siguientes:

```

1 tree_final = tree.best_estimator_
2 tree_final.fit(X_train,y_train )
3
4 y_train_pred = tree_final.predict(X_train)
5 y_test_pred = tree_final.predict(X_test)
6
7 from sklearn.metrics import mean_squared_error
8 from sklearn.metrics import mean_absolute_error
9 print("Score", r2_score(y_test, y_test_pred))
10 error = mean_squared_error(y_test,y_test_pred,squared=False)
11 print("RMSE {}".format(error))
12 error3 = mean_absolute_error(y_test,y_test_pred)
13 print("MAE {}".format(error3))

```

```

Score 0.9621637330514261
RMSE 27.59636178558422
MAE 19.12436921555307

```

Post-pruning

El post-pruning, también conocido como Post-poda o poda hacia atrás, es el proceso donde primero se genera el árbol de decisión y luego se eliminan las ramas no significativas. Utilizamos esta técnica después de la construcción del árbol de decisión. Esta técnica se emplea cuando el árbol de decisión tiene una profundidad muy grande o infinita y muestra un sobre ajuste del modelo.

En la técnica anterior de pre poda hemos usado parámetros como '*max_depth*' y '*max_samples_split*', pero aquí podemos las ramas del árbol de decisión usando la técnica *cost_complexity_pruning*. Para aplicar el proceso de post-pruning es necesario indicar el argumento *ccp_alpha* que determina el grado de penalización por complejidad. A medida que este valor resulta mayor, más agresivo es el podado y menor el tamaño del árbol resultante. Para obtener los valores de *ccp_alpha*, primero entrenaremos el modelo con los parámetros predeterminados y luego calculamos su valor con la función *cost_complexity_pruning_path* (). A continuación, se muestra la ejecución del código para la obtención de los valores.

```

1 from sklearn.tree import DecisionTreeRegressor
2 tree_2 = DecisionTreeRegressor(random_state=42)
3
4 reg = tree_2.fit(X_train,y_train.reshape(-1))
5
6 path = reg.cost_complexity_pruning_path(X_train, y_train)
7 ccp_alphas, impurities = path.ccp_alphas, path.impurities
8 print(ccp_alphas)

```

```

[0.00000000e+00 7.51672780e-13 7.51672780e-13 ... 2.18493461e+03
 3.11035160e+03 3.52493485e+04]

```

El siguiente paso es introducir un rango de los valores de *ccp_alpha* para trazar puntajes de validación cruzada de entrenamiento y prueba para cada una de ellas. Esto se hace con la finalidad de obtener el valor óptimo de *ccp_alpha*.

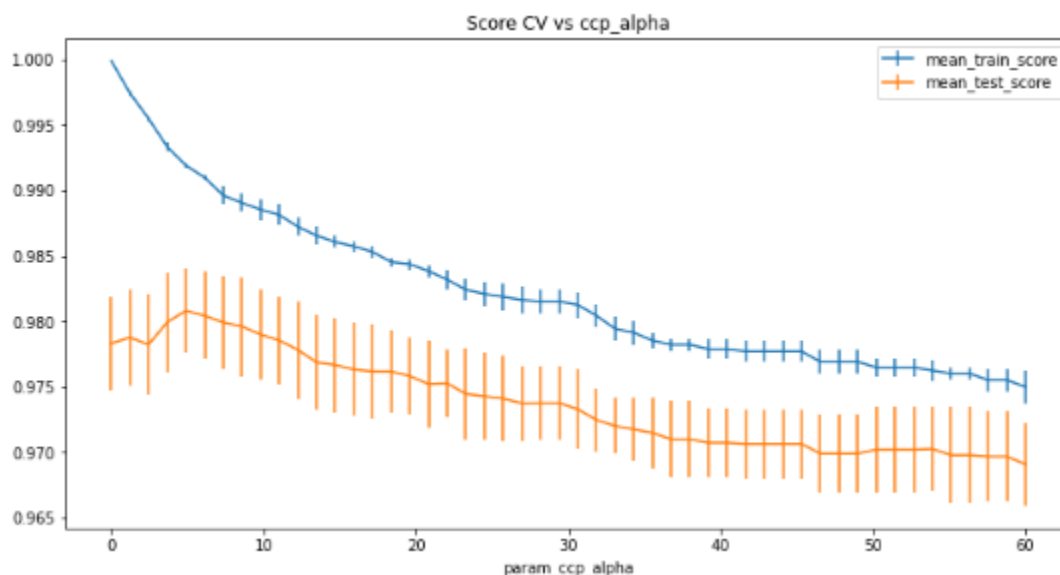


Figura 37. puntajes CV para diferentes *ccp_alpha*.

Fuente: Elaboración propia

Los resultados obtenidos se muestran a continuación

```
1 tree_final = grid.best_estimator_
2 print(grid.best_params_)
3 print(f"Profundidad del árbol: {tree_final.get_depth()}")
4 print(f"Número de nodos terminales: {tree_final.get_n_leaves()}")
5 |
```

{'ccp_alpha': 4.8979591836734695}
 Profundidad del árbol: 16
 Número de nodos terminales: 69

El desempeño de la predicción obtenidos luego entrenar el modelo con los hiperparámetros óptimos con la técnica *post-pruning* y los datos de prueba, son siguiente:

```

1 tree_final = grid.best_estimator_
2 tree_final.fit(X_train,y_train )
3
4 y_train_pred = tree_final.predict(X_train)
5 y_test_pred = tree_final.predict(X_test)
6
7 from sklearn.metrics import mean_squared_error
8 from sklearn.metrics import mean_absolute_error
9 print("Score", r2_score(y_test, y_test_pred))
10 error = mean_squared_error(y_test,y_test_pred,squared=False)
11 print("RMSE {}".format(error))
12 error3 = mean_absolute_error(y_test,y_test_pred)
13 print("MAE {}".format(error3))

```

Score 0.9821558828897646
RMSE 14.879212756932276
MAE 9.329107364353717

En la **Figura 38** se muestra el ajuste del modelo inicial entrenado con el algoritmo árbol de regresión comparado con el modelo real. Este modelo claramente no tiene un buen ajuste respecto al modelo real. Esto se debe principalmente a que los resultados que se había obtenido usando hiperparámetros predeterminados. Sin embargo, en las observaciones siguientes donde se ajustó algunos hiperparámetros utilizando dos técnicas de poda, se obtuvo mejores resultados. Finalmente, podemos decir que tenemos un modelo que no es sobre ajustable y el rendimiento en los datos de prueba ha mejorado mucho mejor con la técnica *post_pruning*.

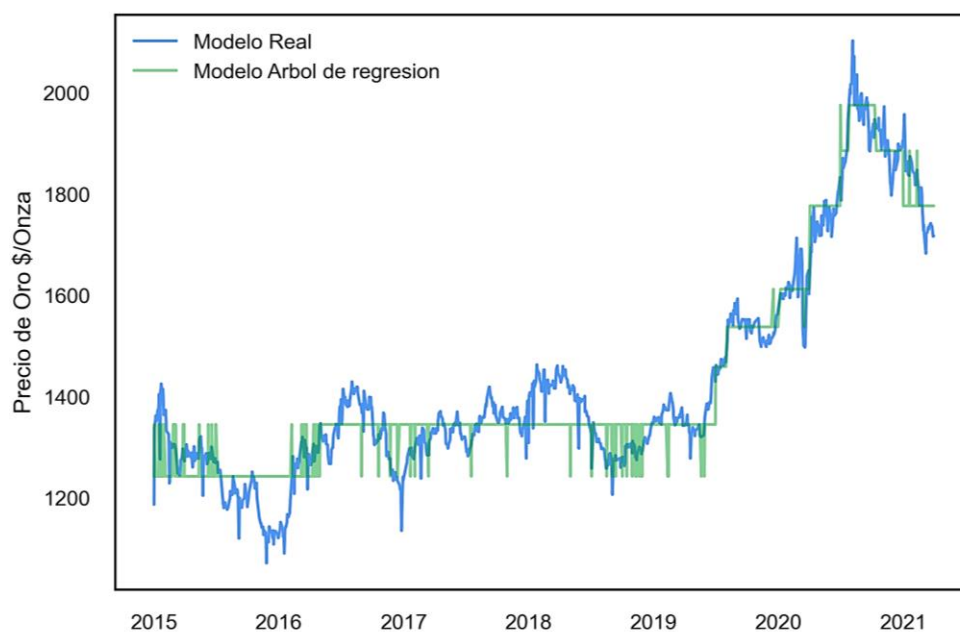


Figura 38. Modelo inicial Árbol de Regresión para la predicción del precio Oro.

Fuente: Elaboración propia

Dentro del modelo *Árbol de regresión* es importante señalar el *diagrama de árbol* que se obtiene para la obtención del modelo final. Este diagrama muestra la forma en que se obtiene la variable a predecir al ingresar las variables predictoras. La **Figura 39** muestra como manera de ejemplo el diagrama de árbol que se obtuvo para el modelo predictivo inicial del precio de Oro. Otra bondad del diagrama de árbol es que te permite identificar cuáles son las variables más influyentes en el modelo final. En la **Tabla 4** se muestra la importancia de las variables en el diagrama final de árbol de regresión. Según esta tabla, constatamos que algunas variables como el *costo de producción*, la *inflación en los estados unidos*, la *producción* y el *consumo mundial de oro* son los que más influyen en la generación del modelo, por ende, en la fluctuación del precio de oro.

```
1 from sklearn.tree import export_graphviz
2 print(export_graphviz(tree_2))
```

Diagrama de árbol de regresión.

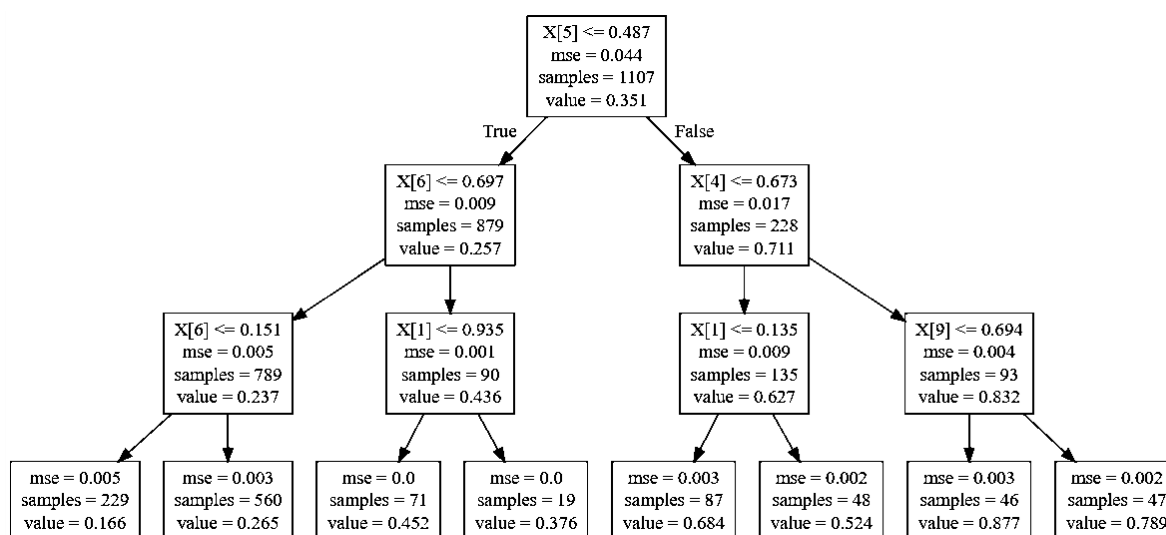


Figura 39. Diagrama del Árbol de Regresión del modelo inicial.

Fuente: Elaboración propia

Tabla 4. Importancia de los predictores en le diagrama árbol de regresión

	predictor	importancia
5	(c, o, s, t, o, _, p, r, o, d)	0.819955
6	(i, n, f, l, a, _, u, s, a)	0.105888
4	(p, r, o, d, u, c, c, i, o, n)	0.051031
8	(c, o, n, s, u, m, o)	0.017312
9	(p, r, e, c, _, p, e, t, r, i)	0.003905
1	(t, a, s, a, _, F, E, D)	0.001910
0	(u, s, d, x)	0.000000
2	(p, b, i, _, u, s, a)	0.000000
3	(p, b, i, _, c, h, i, n, a)	0.000000
7	(S, &, P, _, 5, 0, 0)	0.000000

Fuente: Elaboración propia

5.1.2.7. Evaluación

La evaluación de los modelos se realiza tomando en consideración el *Error Absoluto Medio*, *Error Cuadrático Medio* y el *Puntaje R^2* con respecto a los datos de prueba. En la **Tabla 5** se muestra un resumen de todos los resultados finales obtenidos en la elaboración de los diferentes modelos. En la **Figura 40** se muestra una comparación gráfica de los modelos finales con respecto a la variación real del precio de Oro.

Tabla 5. Resultados de modelos para la predicción del precio de Oro

MODELO	MÉTRICAS			
	MAE	RMSE	PUNTAJE (R2) Test	
Regresión Lineal	17.75	31.49	0.91	
Árbol de Regresión	9.32	14.32	0.98	
SVR	K=rbf	4.63	5.29	0.94
	K=lin	11.23	18.61	0.89
	K=pol	6.51	8.62	0.92

Fuente: Elaboración propia

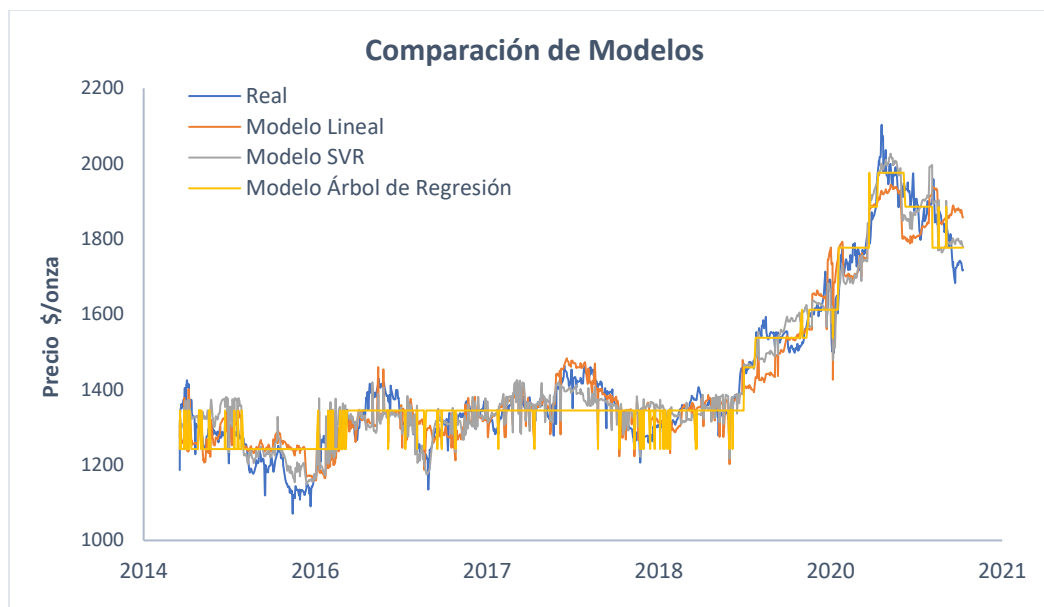


Figura 40. Comparación de modelos para la predicción del precio de Oro.

Fuente: Elaboración propia

5.1.2.8. Predicción

Esta fase implica la predicción del modelo luego de la implementación al negocio o al proceso para la cual fue desarrollado. Para fines de esta investigación, realizaremos la predicción solo para algunas muestras de datos nuevos.

Tabla 6. Predicción del precio de Oro

Predicción del precio de oro		
Muestras	Real \$/Onza	Predicción \$/Onza
1	1318.30	1316.94
2	1361.70	1360.64
3	1307.40	1306.33
4	1314.40	1314.45
5	1329.70	1329.44
6	1361.70	1360.64
7	1388.50	1388.58
8	1360.40	1361.09
9	1362.30	1360.75
10	1366.80	1365.46
11	1360.40	1361.09

5.1.3. Análisis de resultados

El alcance exploratorio de esta investigación nos permitió probar diferentes algoritmos para los casos explicados en esta investigación. Estos algoritmos fueron escogidos según la complejidad del problema a resolver. Además, nos permitió descubrir que algoritmo tiene mejor desempeño en cada caso en específico.

En el presente caso se optó por los algoritmos de regresión lineal múltiple, árbol de regresión y SVR. Todos con desempeños muy buenos al probar el modelo. El modelo que tuvo un desempeño mejor que los demás fue la que se generó con el método **SVR** obteniendo un puntaje **$R^2 = 0.94$, $MAE=4.63$ y $RMSE=5.29$** en la prueba. Estos resultados demuestran la efectividad de los algoritmos de machine learning para resolver problemas complejos de regresión. Sin embargo, para llegar a estos resultados se debe hacer un análisis cuidadoso de los pasos del flujograma explicado en el capítulo correspondiente, sobre todo, el primer paso que es entender y categorizar bien el problema que se quiere resolver. Esto es determinante para los pasos siguientes y, en última instancia, repercutirá el desempeño final del modelo.

Sin duda, los resultados obtenidos son muy buenos que incluso pueden tener mejor desempeño que los modelos tradicionales con estadística clásica. Estos modelos tienen un potencial enorme para ser utilizados por los gobiernos, los bancos, las empresas mineras, operadores de corto plazo y otros interesados. Además, estos algoritmos tienen el potencial de identificar cuáles son las variables predictoras que más influye en la variación del precio de oro, de esa forma podemos hacer un análisis de los riesgos potenciales cuando el precio de oro baja, ayudando a tomar decisiones estratégicas a las empresas mineras que producen oro.

Importancia de predictores

Cuando generamos modelos predictivos con algoritmos de machine learning, hay muchas formas de saber la importancia que tiene cada predictor respecto a la variable dependiente. No siempre el orden en la importancia de variables es igual para todos los modelos, en algunos casos una variable independiente tiene mayor relevancia en un modelo determinado que en otros. En este caso en particular se usó herramientas como la eliminación de características recursivas (RFE), método de filtrado y a través del modelo árbol de regresión.

Tabla 7. Importancia de las Variables en los modelos para predecir el precio de Oro

Importancia de Predictores	Método		
	RFE	Filtrado	Arbol de Regresión
1	infla_usa	costo_prod	costo_prod
2	tasa_FED	infla_usa	infla_usa
3	usdx	consumo	produccion
4	pbi_usa	S&P_500	consumo
5	pbi_china	usdx	prec_petr
6	costo_prod	tasa_FED	tasa_FED
7	produccion	produccion	usdx
8	consumo	prec_petr	pbi_usa
9	prec_petr	pbi_usa	pbi_china
10	S&P_500	pbi_china	S&P_500

Fuente: Elaboración propia

usdx	: Índice de precio de dólar
tasa_FED	: Tasa de interés reserva federal
pbi_china	: PBI china
pbi_usa	: PBI estados unidos
produccion	: Producción mundial de oro
infla_usa	: Inflación de Estados Unidos
S&P_500	: Índice S&P 500
consumo	: Consumo mundial de oro
prec_petr	: Precio del petróleo

Una de las variables que más influyen en la variación del precio de Oro es la **Inflación de los Estados Unidos** (Ver **tabla 7**). Esta influencia es debido a que los inversores suelen confiar más en el Oro en tiempos de incertidumbre económica y política. Esta estrategia lo utilizan a menudo como una herramienta de cobertura contra la inflación o la devaluación de divisas.

Otra variable de gran significancia en la fluctuación del Precio de Oro es el **índice del precio del dólar**. La influencia se debe a una relación inversamente proporcional y que pueden ser muy decisiva. La razón de esta correlación se debe principalmente al siguiente comportamiento:

cuando el tipo de cambio del dólar se deprecia en relación con las otras monedas de reserva, el poder adquisitivo de la compra de Oro en otras monedas se mantiene, la consecuencia lógica es el aumento de precio de Oro en relación con el dólar.

La variación del Precio de Oro también está influenciada por la oferta y la demanda, así como sucede con cualquier tipo de mercancía. El **Consumo** de oro tiene diferentes propósitos como la industria de la joyería, tecnología y principalmente como reserva para los inversionistas y los bancos centrales. Por otro lado, es importante destacar que el precio de Oro depende de la **Producción** mundial. La producción cayó un 26% aproximadamente entre el 2011 y 2019. Esta tendencia junto con la dificultad de encontrar nuevos yacimientos puede incentivar el alza del precio si es que los otros factores se mantienen constantes.

El oro como cualquier otro mineral pasa por distintas etapas y sub procesos hasta convertirse en un bien comercial, pueden ser lingotes, monedas de oro, etc. Este proceso conlleva a un **Costo de Producción** hasta obtención del producto final. Este costo es inherente al tratamiento del producto, por lo tanto, a su precio de venta. Si bien es cierto las empresas junto con la academia han ido trabajando para poder reducir sus costos considerablemente logrando explotar yacimientos de baja ley y poca profundidad, muchos de los yacimientos en explotación se están agotando. La explotación de yacimientos a gran profundidad podría tener un incremento considerable de los costos haciendo que el precio del metal también se incremente.

5.2. Modelo predictivo para estimar sílice en el concentrado de hierro

5.2.1. Compresión del problema

Dentro del diagrama de flujo para el procesamiento del mineral de hierro (Ver **figura. 41**) existen varias etapas dentro de las cuales están el chancado, molienda y la separación magnética. Si bien es cierto que los métodos de concentración magnética son muy eficientes para la recuperación alta en los minerales de hierro, este se vuelve ineficientes cuando se requiere la producción de concentrado de hierro con menos de 8% a 6% de sílice. Es ahí cuando los resultados obtenidos con el proceso de flotación son más eficientes. Previa a la etapa de flotación se realiza la limpieza con equipos magnéticos que puede llegar a producir concentrados de hasta 63% a 64 % de hierro con una pulpa de 50% a 55% de sólidos. Este producto es enviado al circuito de flotación para la remoción de sílice hasta 4% a 6% en el concentrado final (Metallurgist911, 2017).

Obtener de manera continua los resultados anteriores después de la flotación es una tarea muy difícil para los operadores del proceso. Las decisiones que toman están sujetas al tiempo que demora en brindar los resultados del análisis del laboratorio, que incluso pueden ser muchas horas. Este tiempo muerto impide tomar acciones inmediatas para cumplir las especificaciones de salida de manera continua.

Una solución para resolver el problema anterior sería estimar el concentrado de SiO_2 en un tiempo más corto. Esta solución podría ayudar en la tomar decisiones más rápidas como: aumentar el flujo de reactivos, reducir la velocidad de alimentación o algún otro parámetro. Esto se puede lograr usando modelos predictivos construidos con algoritmos de machine learning.

Objetivo: desarrollar un modelo predictivo para la estimación de SiO_2 en el concentrado de hierro a partir de datos históricos.

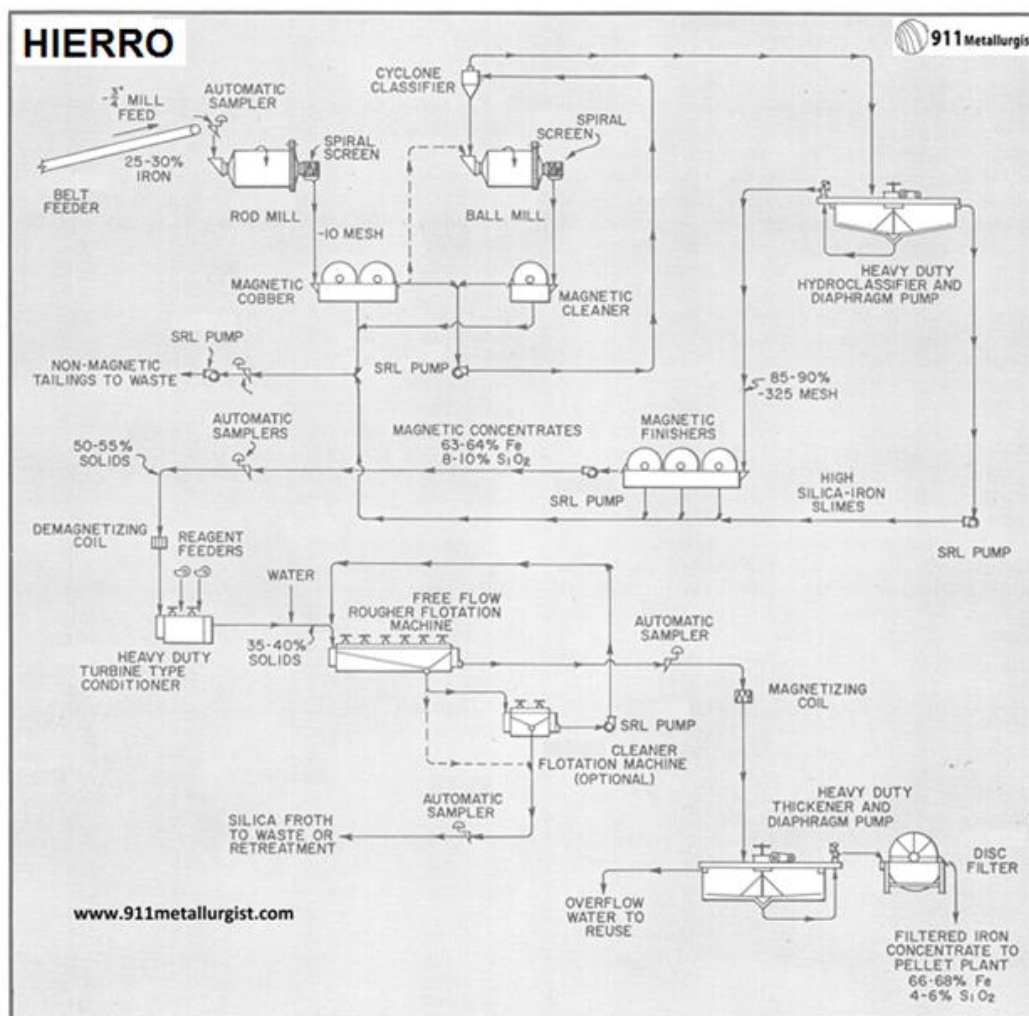


Figura 41. Diagrama de flujo para la obtención del concentrado de Hierro

Fuente: adaptado de (Metallurgist911, 2017)

5.2.2. Aplicación del flujograma para el desarrollo del modelo

5.2.2.1. Categorización del problema

Según lo expuesto en el apartado Categorización del problema (Ver cap. IV, 4.2.1). Para predecir el concentrado de sílice aplicaremos el método de aprendizaje supervisado debido a que se está trabajando con datos continuos y conocidos tanto para los datos de entrada como para los datos de salida. Por otro lado, nuestro problema tendrá como dato de salida un valor numérico, por lo que será un problema de regresión.

Las variables descritas en la base de datos se mencionan en la tabla siguiente, donde la variable a predecir es el concentrado de sílice (% Silice Concentrate).

Variable	Descripción
% Iron Feed	El contenido de hierro en la pulpa antes de la alimentacion a la planta de flotación
% Silice Feed	El contenido de sílice en la pulpa antes de la alimentacion a la planta de flotación
Starch Flow	Variables importantes
Amina Flow	
Ore Pulp Flow	Flujo de pulpa de mineral
Ore Pulp pH	pH de pulpa de mineral
Ore Pulp Density	Densidad de la pulpa de mineral
Flotation Flow	Flujo de aire en la flotación
Flotation Level	Nivel de aire en flotación
% Silice Concentrate	Concentrado de sílice

5.2.2.2. Métricas de medición

Según nuestros objetivos planteados y la categorización de nuestro problema, las métricas a tomar en consideración para este caso será el RMSE, MAE, R^2 .

5.2.2.3. Recolección de datos

La base de datos con las que estamos trabajando se tomaron cada 20 segundos, algunos cada 1 hora y con fechas entre marzo de 2017 a setiembre del 2017 con un total de 700k datos.

Para fines de este ejemplo se hicieron varias pruebas con 10 000 datos para cada entrenamiento. Además, esta base de datos consta con 22 columnas repartidas entre las variables mencionadas en el apartado de selección de variables. La **Tabla 8** muestra la presentación de algunas de las variables y sus principales indicadores estadísticos.

Tabla 8. Indicadores estadísticos de las variables

	% Iron Feed	% Silica Feed	Starch Flow	Amina Flow	Ore Pulp Flow	Ore Pulp pH	Ore Pulp Density	Flotation Column 01 Air Flow	Flotation Column 01 Level	% Silica Concentrate
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
mean	56.324172	14.642105	2885.001842	485.831770	397.485708	9.765803	1.679146	280.199668	521.783943	2.299640
std	5.198152	6.841805	1216.035036	91.855674	9.632551	0.387922	0.070118	29.978729	131.565213	1.113268
min	42.740000	1.310000	0.311761	241.699000	376.254000	8.753391	1.519832	175.820296	157.636355	0.600000
25%	52.655000	8.870000	2082.805000	426.562000	394.077000	9.527128	1.645525	250.320750	417.421000	1.430000
50%	56.080000	13.950000	3020.665000	502.319000	399.240500	9.793670	1.697020	299.357500	492.041500	1.960000
75%	59.720000	19.600000	3734.092500	552.190500	402.892250	10.033800	1.728075	300.148250	595.145000	2.932500
max	65.780000	33.400000	6293.142347	739.369400	418.550714	10.807492	1.851990	369.885072	661.949833	5.530000

Fuente: Elaboración propia

5.2.2.4. Análisis de datos

Visualización de datos

En la **Figura 42** se puede observar la variabilidad del concentrado de SiO_2 en el concentrado de hierro obtenido después del proceso de flotación. En ella observamos que hay concentrados que poseen de 6% hasta 1% de sílice, esto nos lleva a pensar que controlar el proceso para tener un concentrado continuo es muy difícil.

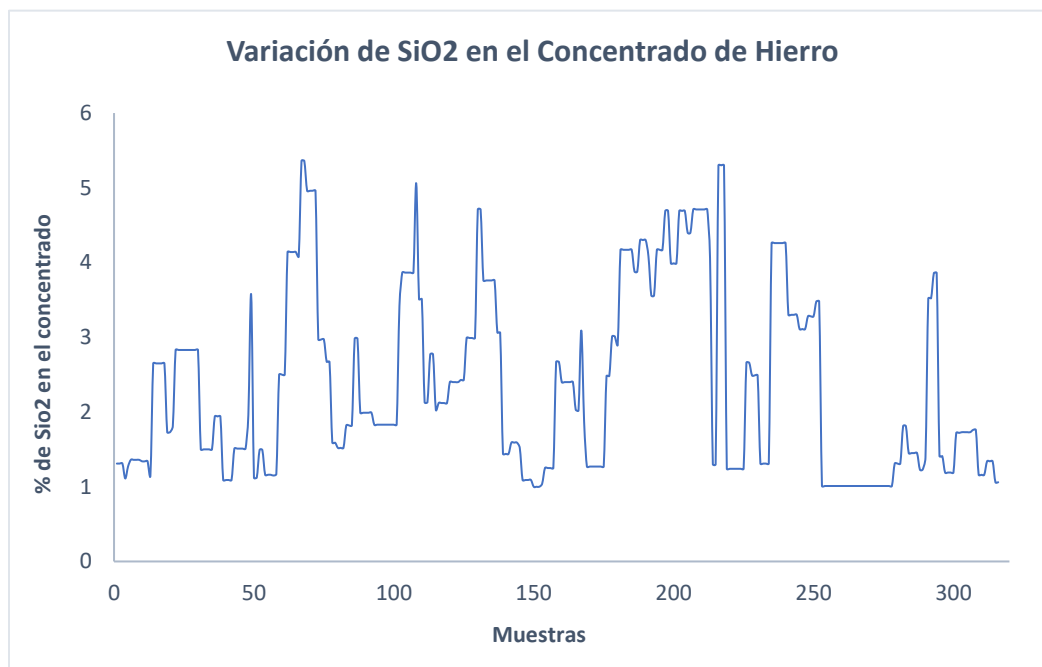


Figura 42. Variación del Concentrado de Sílice

Fuente: Elaboración propia

Generar un modelo que nos ayude a predecir en poco tiempo el SiO_2 en el concentrado final de hierro, sin duda ayudaría tomar acciones rápidas para tener un producto final dentro de los estándares. Además, con ayuda de los algoritmos podremos identificar cuáles son las variables que más influyen en la variación de SiO_2 en el concentrado final y con ello generar un modelo más robusto.

Correlacion: Variables Vs Concentrado de Silice

```

Flotation Column 03 Air Flow  -0.218074
Flotation Column 01 Air Flow  -0.214420
Flotation Column 05 Level     -0.179482
Flotation Column 02 Air Flow  -0.165844
Ore Pulp pH                   -0.163897
Flotation Column 07 Level     -0.152757
Flotation Column 04 Level     -0.143244
Flotation Column 06 Level     -0.094819
% Iron Feed                   -0.078602
Starch Flow                   -0.072270
Flotation Column 07 Air Flow  -0.067182
Flotation Column 06 Air Flow  -0.032640
Flotation Column 04 Air Flow  -0.018035
Flotation Column 05 Air Flow  -0.009543
Ore Pulp Flow                  0.010912
Flotation Column 01 Level      0.015228
Flotation Column 02 Level      0.015401
Flotation Column 03 Level      0.020434
Ore Pulp Density               0.062236
% Silica Feed                  0.074291
Amina Flow                    0.174046
% Silica Concentrate           1.000000
Name: % Silica Concentrate, dtype: float64

```

A pesar de que existen muchas variables predictoras, la correlación de estas con la variable a predecir son bastantes bajas.

El flujo de aire en la flotación y el nivel de aire en la flotación son los que mayormente influyen en la concentración final de la SiO_2 .

En menor medida los que también influyen el pH de la pulpa y la Amina.

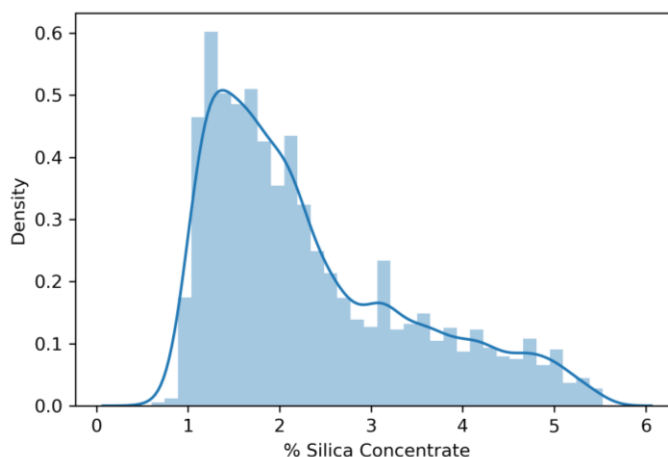


Figura 43. Histograma del Concentrado Sílice

Fuente: Elaboración propia

En el histograma del concentrado de SiO_2 (Ver **figura. 43**), la distribución normal del concentrado de Sílice parece estar lejos, más bien podemos notar una alta variación estándar debido principalmente a que los resultados obtenidos no tienen una continuidad durante el proceso.

Limpieza de datos

Para fines de este ejemplo se desarrolló pruebas con grupos de 10 mil datos, dentro de las cuales se desarrolló los siguientes procesos con el fin de tener datos más enriquecidos.

- Eliminación de valores atípicos
- Eliminación de datos variables colineales
- Selección de variables finales con la técnica de eliminación de características recursivas (RFE).
- Eliminación de datos duplicados

División de datos

Para la validación del modelo la división de los datos fue:

- 67% de datos para el entrenamiento.
- 33% de datos para las pruebas.

Estandarización de datos

En el caso actual para usar estimadores de machine learning en la predicción de sílice en el concentrado de hierro, escalaremos las características para que se encuentren entre un valor mínimo y máximo, específicamente entre 0 y 1, esto se puede lograr usando MinMaxScaler de la librería Scikit-Learn.

5.2.2.5. Elección de algoritmo

Para encontrar el mejor modelo para los datos seleccionados, primero se hizo pruebas con los métodos más simples como la Regresión Lineal Múltiple, Árboles de Regresión y SVR. Luego hicimos pruebas con algoritmos de ensamble que son más sofisticados como es el caso de Random Forest Regressor, Gradient Boosting Regressor y Votin Regressor. La elección de estos algoritmos se hizo con la finalidad mejorar los resultados obtenidos con algoritmos menos complejos.

Random Forest Regressor

Random Forest es un conjunto de árboles de decisión que se caracteriza por ser flexible y fácil de usar. Generalmente, produce resultados muy buenos incluso sin ajuste de los parámetros. Este algoritmo es uno de los más utilizados debido a que se emplea para realizar tareas de clasificación como de regresión.

La funcionalidad de este algoritmo de aprendizaje supervisado se basa en generar múltiples árboles de decisión y los combina para obtener una predicción más precisa y estable. En la mayoría de los casos el modelo se vuelve más robusto mientras más árboles hay, pero se debe tomar en consideración el costo computacional.

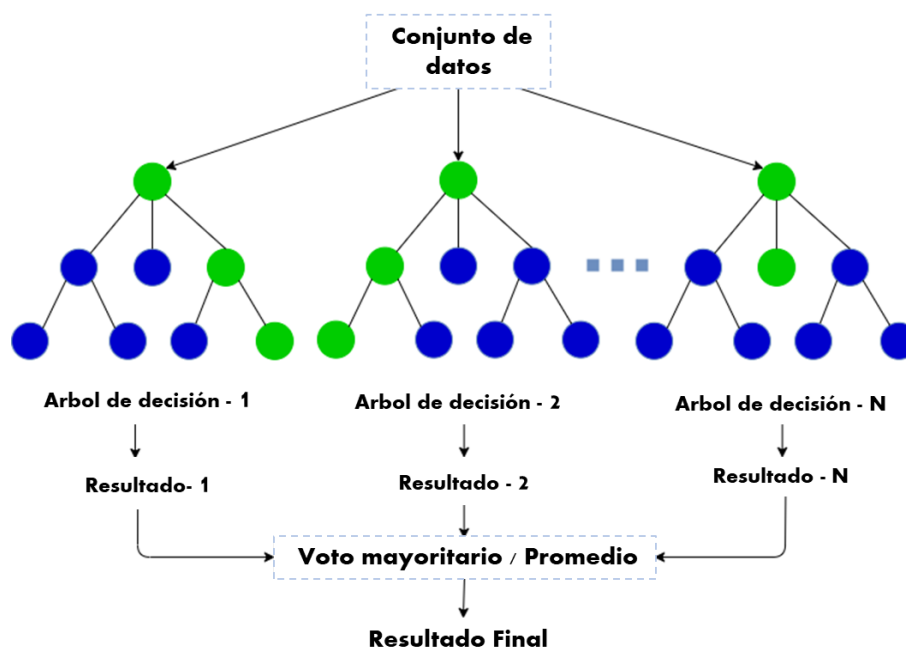


Figura 44. Ilustración del Modelo Random Forest

Fuente: Adaptado de *Random Forests Understanding*, por Ampadu, 2021, <https://ai-pool.com/a/s/random-forests-understanding>.

En este algoritmo se agrega aleatoriedad adicional al modelo mientras crece los árboles. En lugar de buscar la característica más importante al dividir un nodo, busca la mejor característica entre un subconjunto aleatorio de características. Esto da como resultado una amplia diversidad que generalmente resulta en un mejor modelo. Por lo tanto, en Random Forest (Bosques Aleatorios), el algoritmo para dividir un nodo solo tiene en cuenta un subconjunto aleatorio de las

características. Incluso puede hacer que los árboles sean más aleatorios mediante el uso adicional de umbrales aleatorios para cada función; en lugar de buscar los mejores umbrales posibles, como lo hace un árbol de decisión normal.

Gradient Boosting Regressor

El Gradient Boosting (Aumento de gradiente) es otro de los algoritmos populares en la comunidad de científicos de datos. Este es una técnica que se puede usar tanto en problemas de regresión como en problemas de clasificación. La funcionalidad se basa en construir un conjunto de árboles sucesivos poco profundos y débiles con el fin ir aprendiendo y mejorando lo anterior. Este proceso de combinar muchos árboles sucesivos débiles produce un conjunto poderoso que muy pocas veces es superado con otros algoritmos.

Las ventajas de utilizar este algoritmo es la precisión de predicción que a menudo es insuperable, es muy flexible, funciona bien con datos sin pre procesamiento y maneja los datos faltantes.

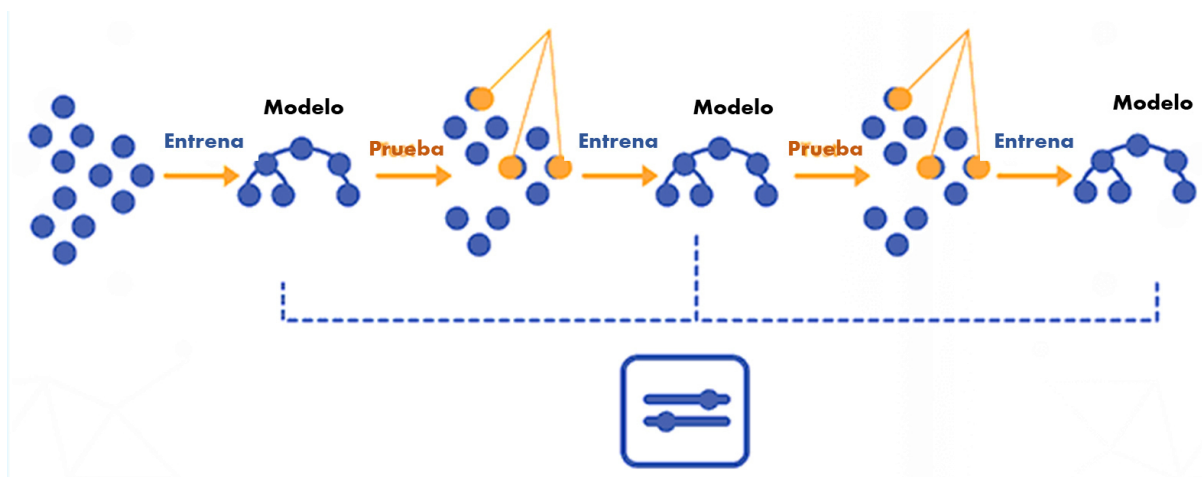


Figura 45. Ilustración del Modelo Gradient Boosting

Fuente: adaptado de *Gradient Boosting*, por Akira AI, 2020, <https://www.akira.ai/glossary/gradient-boosting>

Votin Regressor

El Votin Regressor es un metaestimador de conjunto que se ajusta a varios regresores de base, cada uno en el conjunto de datos completos. Luego, promedia las predicciones individuales para formar una predicción final. Un regresor de este tipo puede ser útil para un conjunto de modelos igualmente buenos con el fin de equilibrar sus debilidades individuales.

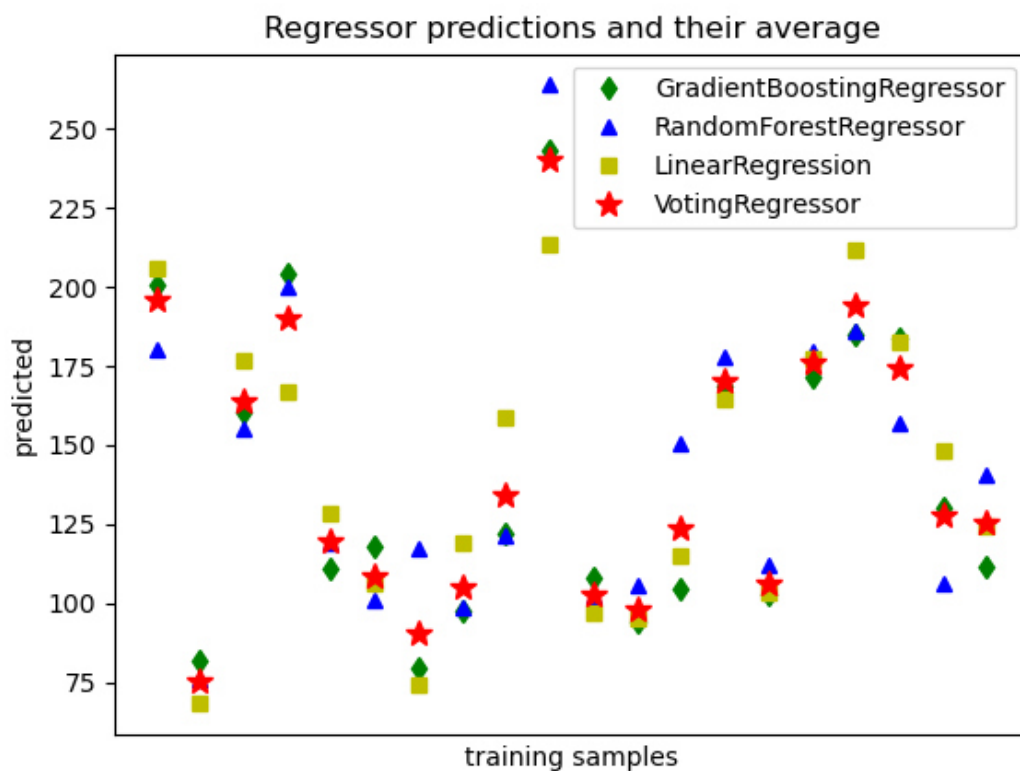


Figura 46. Ilustración del Modelo Votin Regressor

Fuente: Adaptado de (scikit-learn, 2021)

5.2.2.6. Modelado

Ante la baja correlación de nuestras variables predictoras frente a la variable a predecir, incluido la falta de otras variables importante que influyen en la variación del concentrado final de Sílice. Se hizo pruebas con algoritmos más sofisticados obteniendo mejoras considerables.

Support Vector Regressor

Para el desarrollo de este modelo se realizó el mismo proceso que en el primer ejemplo. En la **Figura 47** muestra la comparación del modelo SVR frente a la variación real del concentrado de sílice. En esta imagen se nota la gran diferencia que existe en el ajuste de modelo predictivo frente al modelo real. En el ejemplo 1 (Ver cap V, 5.1) se obtuvo resultados muy buenos al desarrollar un modelo con el algoritmo SVR, sin embargo, ante problemas más complejos se nota la debilidad de los modelos menos complejos con respecto a los algoritmos de ensamble.

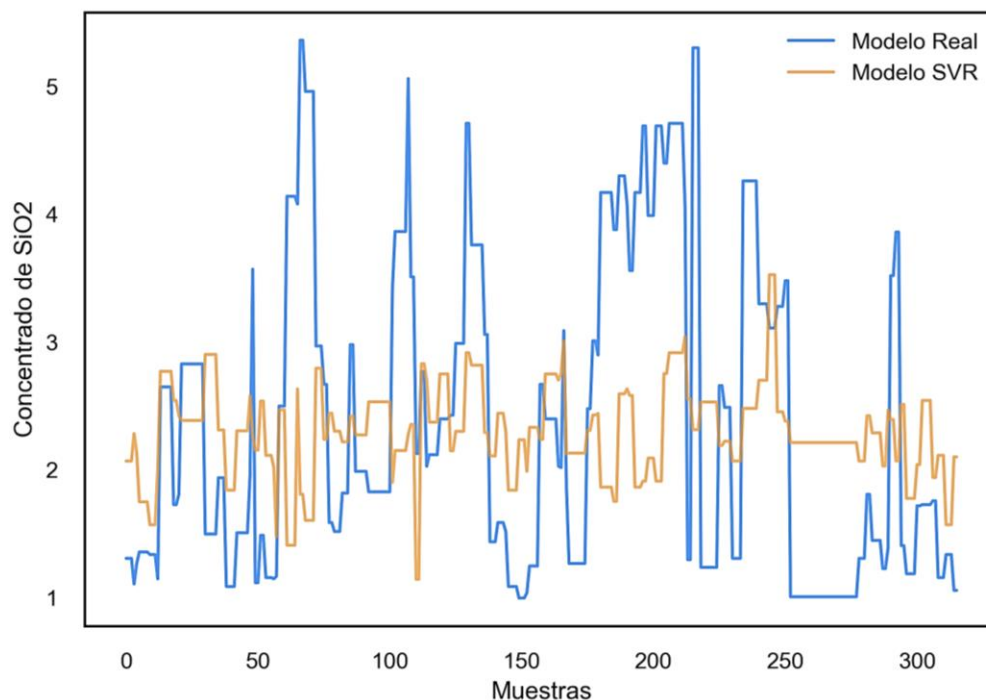


Figura 47. Modelo SVR para la predicción de Concentrado de Sílice

Fuente: Elaboración propia

Random Forest

Random forest es un algoritmo de aprendizaje supervisado que se caracteriza por su fácil uso. Actualmente, la librería Scikit-learn tiene este algoritmo implementado para resolver tareas tanto de regresión como de clasificación. Para este fin usaremos el módulo *sklearn.ensemble*, el cual contiene la clase ***RandomForestRegressor()***. Como la mayoría de los algoritmos que posee hiperparámetros, esta necesita ajustar su modelo para tener mejor rendimiento. Este proceso puede

ser muy trabajoso y llevar bastante tiempo cuando se realiza de manera separada o probando aleatoriamente los hiperparámetros que contienen. Para evitar este tipo de problemas al ajustar nuestro modelo random forest, utilizaremos **GridSearchCV** para optimizar los hiperparámetros.

El uso de los hiperparámetros que se usa dentro de la cuadrícula de **GridSearchCV** tiene que ser revisado y evaluado para su utilización. Tener muchos hiperparámetros dentro de la cuadrícula puede causar demoras en su cálculo y tener mayor costo computacional. Para este fin tomaremos en consideración parámetros como **max_features**, **max_depth**, **criterion**, **Bootstrap** y el más importante de todos, **n_estimators** donde:

n_estimators: Es el número de árboles en el bosque

max_features: Es el número de características a considerar la mejor división.

criterion: Función para medir la calidad de una división. Los criterios admitidos son "squared_error" para el error cuadrático medio, que es igual a la reducción de varianza como criterio de selección de características, "absolute_error" para el error absoluto medio y "poisson" que utiliza la reducción en la desviación de Poisson para encontrar divisiones.

max_depth: la profundidad máxima del árbol

bootstrap: Si se utilizan muestras de bootstrap al crear árboles. Si es False, se usa todo el conjunto de datos para generar cada árbol.

```

1  from sklearn.model_selection import GridSearchCV
2  from sklearn.ensemble import RandomForestRegressor
3
4
5  parameters = {
6      'n_estimators': [100, 200, 400, 600],
7      'max_features': ['auto', 'sqrt', 'log2'],
8      'criterion': ['mse', 'mae'],
9      'max_depth': [2,3,4],
10     'bootstrap': [True, False]
11 }
12 random_f = GridSearchCV(estimator=RandomForestRegressor(random_state=42),
13                         param_grid=parameters,
14                         scoring='accuracy',
15                         cv=5,
16                         n_jobs=-1,
17                         refit = True,
18                         return_train_score = True)
19
20 random_f.fit(X_train,y_train.ravel())

```

Dentro de los resultados obtenidos se encontró como mejor **n_estimators = 100**. Ajustado y entrenado el modelo, los resultados obtenidos fueron los siguientes.

```
1 print(random_f.best_params_)
{'n_estimators': 100}

1 rf_final = random_f.best_estimator_
2 rf_final.fit(X_train,y_train )
3
4 y_train_pred = rf_final.predict(X_train)
5 y_test_pred = rf_final.predict(X_test)
6
7 from sklearn.metrics import mean_squared_error
8 from sklearn.metrics import mean_absolute_error
9 print("Score", r2_score(y_test, y_test_pred))
10 error = mean_squared_error(y_test,y_test_pred,squared=False)
11 print("RMSE {}".format(error))
12 error3 = mean_absolute_error(y_test,y_test_pred)
13 print("MAE {}".format(error3))

Score 0.45088277131050625
RMSE 0.1627292350272771
MAE 0.12050698395264428
```

La **Figura 48** muestra la comparación del modelo Random Forest frente a la variación real del concentrado de sílice.

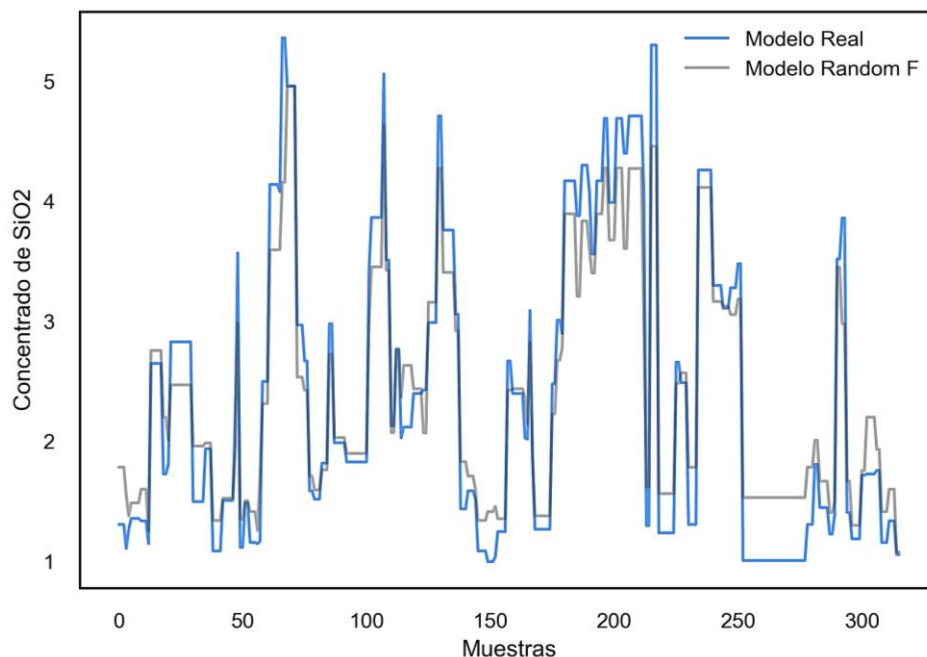


Figura 48. Modelo Random Forest para la predicción de Concentrado de Sílice

Fuente: Elaboración propia

Para una mejor visualización del comportamiento de los datos estimados a partir del modelo predictivo frente a los datos reales, se graficó el ajuste lineal de cada una de ellas. La **Figura 49** muestra el ajuste de los datos para el modelo **SVR** y la **Figura 50** muestra el ajuste de los datos para el modelo **Random Forest**. En ambos casos los datos de prueba estimada y los datos de prueba reales están lejos de alcanzar una relación lineal. Sin embargo, los datos obtenidos a partir del modelo de ensamble, muestra una mejoría considerable frente a los datos estimados a partir de modelos tradicionales.

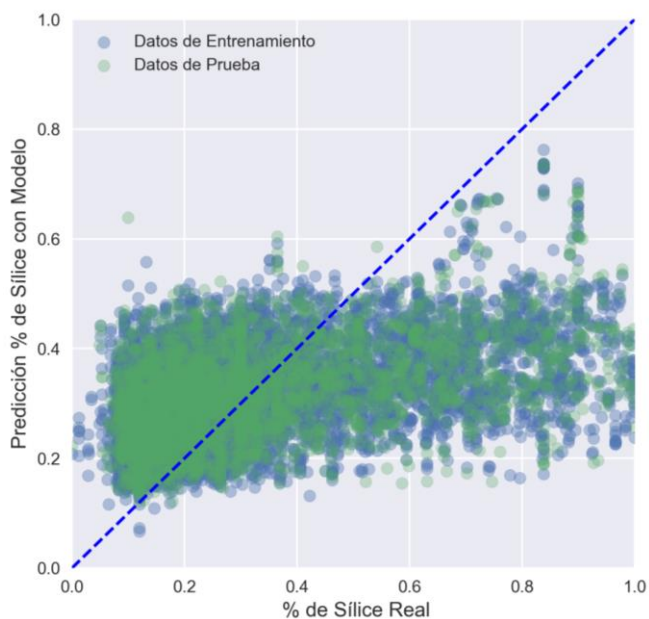


Figura 49. Ajuste de los datos con Algoritmo Normal SVR

Fuente: Elaboración propia

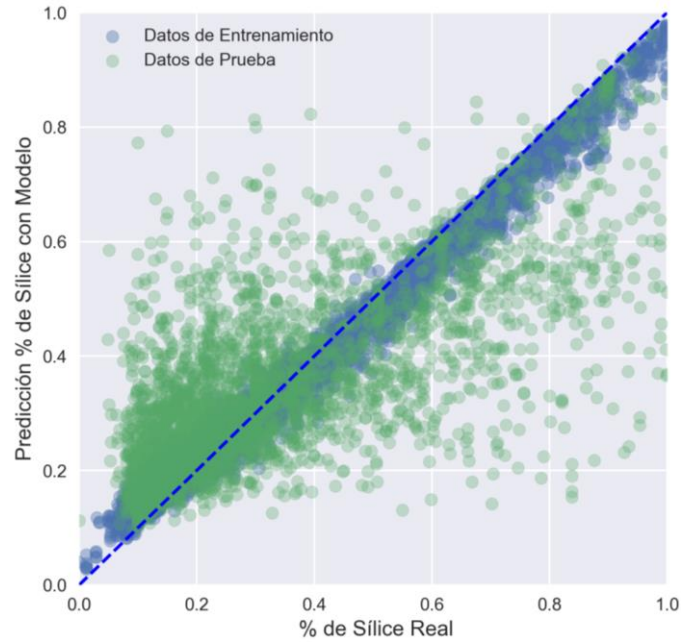


Figura 50. Ajuste de los datos con Algoritmo de Ensamble GB

Fuente: Elaboración propia

5.2.2.7. Evaluación

Los resultados finales de la evaluación de los modelos se muestran en la **Tabla 9**. Gráficamente la **Figura 51** muestra la diferencia significativa de ajuste que existe entre los modelos tradicionales de machine learning y los modelos de ensamble.

Tabla 9. Resultados de los Modelos para la predicción de Concentrado de Sílice

MODELO	METRICAS		
	MAE	RMSE	PUNTAJE (R2) Test
Árbol de Regresión	1.01	1.01	0.21
SVR K=rbf	1.03	1.03	0.16
Randon Forest	0.12	0.16	0.45
Gradient Boosting	0.81	0.81	0.51
Voting Regressor	0.86	0.86	0.43

Fuente: Elaboración propia

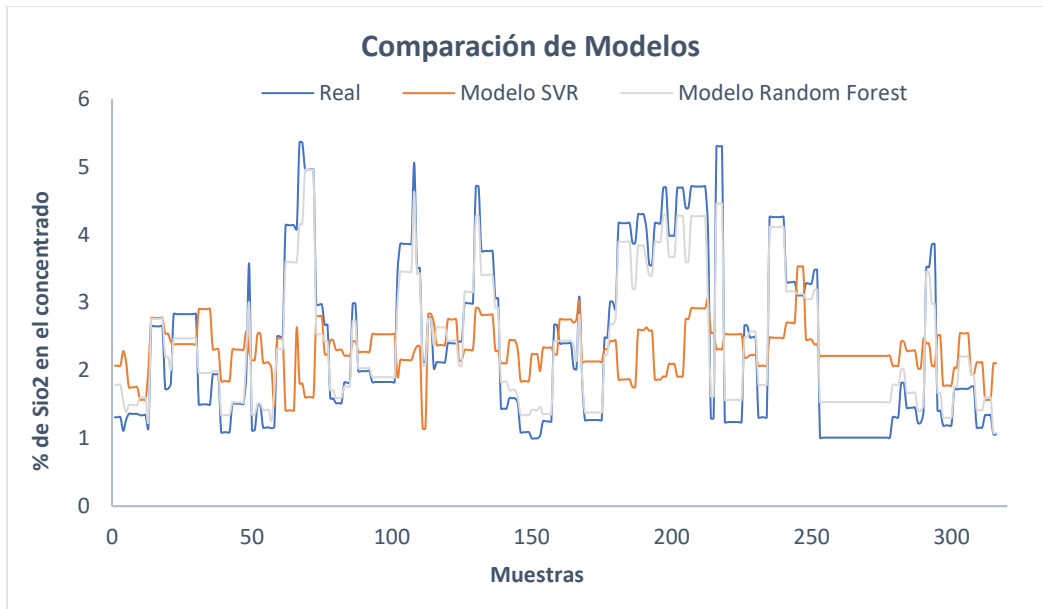


Figura 51. Comparación de modelo para la predicción de Concentrado de Sílice

Fuente: Elaboración propia

5.2.2.8. Predicción

Esta fase implica la predicción con el modelo luego de la implementación al negocio o al proceso para la cual fue desarrollado. Para fines de esta investigación, realizaremos la predicción solo para algunas muestras de datos nuevos.

Tabla 10. Predicción del contenido de Sílice

Predicción del contenido de sílice		
Muestras	Real	Predicción
	% de Sílice	% de Sílice
1	4.96	4.95
2	4.96	4.95
3	4.96	4.95
4	2.77	2.76
5	2.77	2.76
6	2.13	2.08
7	1.82	1.76
8	1.82	1.76
9	1.82	1.76
10	3.52	3.45
11	3.52	3.45

5.2.3. Análisis de resultados

En el ejemplo anterior se optó por hacer una comparación entre los algoritmos tradicionales (**Regresión lineal múltiple, árbol de regresión y SVR**) y los algoritmos más complejos o llamados también de ensamble (**Random Forest Regressor, Gradient Boosting Regressor y Votin Regressor**). La finalidad de usar estos algoritmos más sofisticados fue mostrar capacidad para generar modelos predictivos aun cuando las variables predictoras tenían una correlación baja con respecto a la variable a predecir. El desempeño de los modelos generados a partir de los algoritmos de ensamble fue superior significativamente a los otros modelos. La intención en hacer esta comparación era básicamente mostrar la robustez de los algoritmos de ensamble aplicables a algunos problemas más complejos que se quiere resolver.

Dentro de los modelos generados a partir de los algoritmos de ensamble, el modelo que tuvo un mayor desempeño frente a los datos de prueba fue el **Gradient Boosting Regressor** (**$R^2=0.51$, MAE= 0.81, RMSE=0.81**). Si bien es cierto que los resultados obtenidos al aplicar algoritmos más sofisticados mejoran los modelos, estos están limitadas a las variables disponibles en la base de datos. Para un mejor desempeño del modelo desarrollado hace falta otros datos que tienen mucha relevancia dentro del proceso de flotación como alimentación total, algunos parámetros de molienda, total de concentrado producido, leyes de los relaves, ton de relaves y algunos otros parámetros que influyen en el proceso de flotación.

5.3. Modelo predictivo para estimar el consumo de combustible en camiones mineros de alto tonelaje

5.3.1. Compresión del problema

Uno de los procesos productivos de mayor costo en una operación minera a cielo abierto es el transporte, este proceso se caracteriza principalmente por tener una gran cantidad de equipo y alto grado de mecanización. Esta característica hace que el costo de transporte oscila entre los 35 y un 45% del costo total en una operación minera, siendo el consumo de combustible uno de los parámetros que influye en la elevación de este costo. Por otro lado, sabemos que el consumo excesivo de combustible podría causar una elevación de emisión de gases de efecto invernadero como el dióxido de carbono.

Esta problemática se puede solucionar optimizando el consumo de combustible con el uso de un modelo predictivo desarrollados con algoritmos de Machine Learning. En muchas ocasiones, el consumo de combustible está influenciado por caminos defectuosos, mala maniobras de operadores, demoras en las colas, vehículos defectuosos, incluso puede haber desvío de combustible. Desarrollar un modelo para estimar el consumo de combustible y compararlo con el consumo real, ayuda identificar picos altos de consumo real respecto al estimado por el modelo. Con este resultado podemos identificar a las variables que más influyen en el consumo excesivo a través de diferentes herramientas de machine learning, y tomar acciones preventivas para mejorar el estado o desempeño en las variables identificadas.

Dado la importancia de los efectos que podría tener un consumo excesivo de combustible, hacer un modelo que nos ayude a predecir el consumo de combustible es de gran importancia tanto en el aspecto económico como ambiental.

Generalmente, el cálculo de consumo de combustible se basa en parámetros técnicos que derivan de la comprensión del sistema físico del equipo y la dinámica de sus componentes en cada paso del tiempo. Este tipo de análisis podrían tomar demasiado tiempo. Para evitar la demora en este tipo de análisis, desarrollaremos modelos con algoritmos de machine learning con base en la

información histórica y diaria de camiones Cat 797F de una operación minera a tajo abierto en el sur del Perú.

Objetivo: desarrollar un modelo predictivo para calcular el consumo de combustible diario en los camiones mineros de gran tonelaje.



Figura 52. Camión Minero de gran tonelaje CAT 797E

5.3.2. Aplicación del flujograma para el desarrollo del modelo

5.3.2.1. Categorización del problema

Como hemos venido desarrollando en los otros casos y según la caracterización del problema, para desarrollar este modelo aplicaremos el método de aprendizaje supervisado. Además, nuestro problema requiere como dato de salida un valor numérico, por lo cual utilizaremos algoritmo de regresión. Para un análisis comparativo desarrollaremos modelos con técnicas tradicionales o simples de regresión y con técnicas más sofisticadas como son los algoritmos de ensamble.

Las variables descritas en la base de datos se mencionan en los puntos siguientes.

Variable a predecir

- Consumo de combustible diario

Variables predictoras

- Physical availability

- Use
- Use of availability
- Delay
- Hour
- Satnd By
- Productividad ciclo de acarreo
- Tkph vacío
- Tkph cargado
- Velocidad cargada
- Velocidad vacía
- Toneladas sensor
- Payload
- Tiempo retrocediendo
- Tiempo Acarreando
- Queue en minutos total
- Hang en minutos total

5.3.2.2. Métricas de medición

Según los objetivos planteados y la categorización del problema, las métricas a tomar en consideración para este caso será el RMSE, MAE, R^2 .

5.3.2.3. Recolección de datos

Los datos obtenidos fueron del registro diario del área de despacho y los registros de consumo de combustibles de los camiones, entre los periodos de enero de 2020 hasta setiembre de 2021, con un total de 20 000 datos. Para este fin se utilizaron 29 camiones Cat 797F de una operación minera a tajo abierto.

Tabla 11. Lista de camiones usados para el desarrollo del modelo

Camión	Consumo Gal/Hora	Camión	Consumo Gal/Hora	Camión	Consumo Gal/Hora
FC100	33.19	FC120	35.86	FC72	21.25
FC102	34.40	FC121	32.55	FC73	19.10
FC103	33.12	FC122	35.37	FC75	21.98
FC104	33.57	FC126	33.74	FC76	20.79
FC105	33.61	FC127	34.41	FC78	21.86
FC106	32.04	FC29	34.26	FC86	19.77
FC107	36.20	FC30	33.63	FC88	35.53
FC109	36.93	FC58	34.54	FC94	33.04
FC111	33.90	FC59	35.88	FC95	33.65
FC112	33.19	FC61	35.72	FC96	33.03
FC116	34.18	FC63	35.99	FC97	34.48
FC117	33.17	FC70	21.75	FC98	33.75
FC119	35.41	FC71	19.28	FC99	34.62

Fuente: Elaboración propia

5.3.2.4. Análisis de datos**Visualización de datos**

La **Figura 53** muestra la variación diaria del consumo de combustible en un camión. Si bien es cierto que este consumo puede estar sujeto al factor de utilización o a la suma total de los tiempos en un ciclo de acarreo, los picos pueden estar también influenciados por otros factores que contribuye al consumo innecesario de combustible. La gráfica muestra una incoherencia entre los picos de consumo con la productividad diaria. Esta incoherencia puede estar sujeto a motivos como las colas de los camiones en las palas, habilidad del conductor, condición del camión, mala asignación y distribución de camiones, etc.

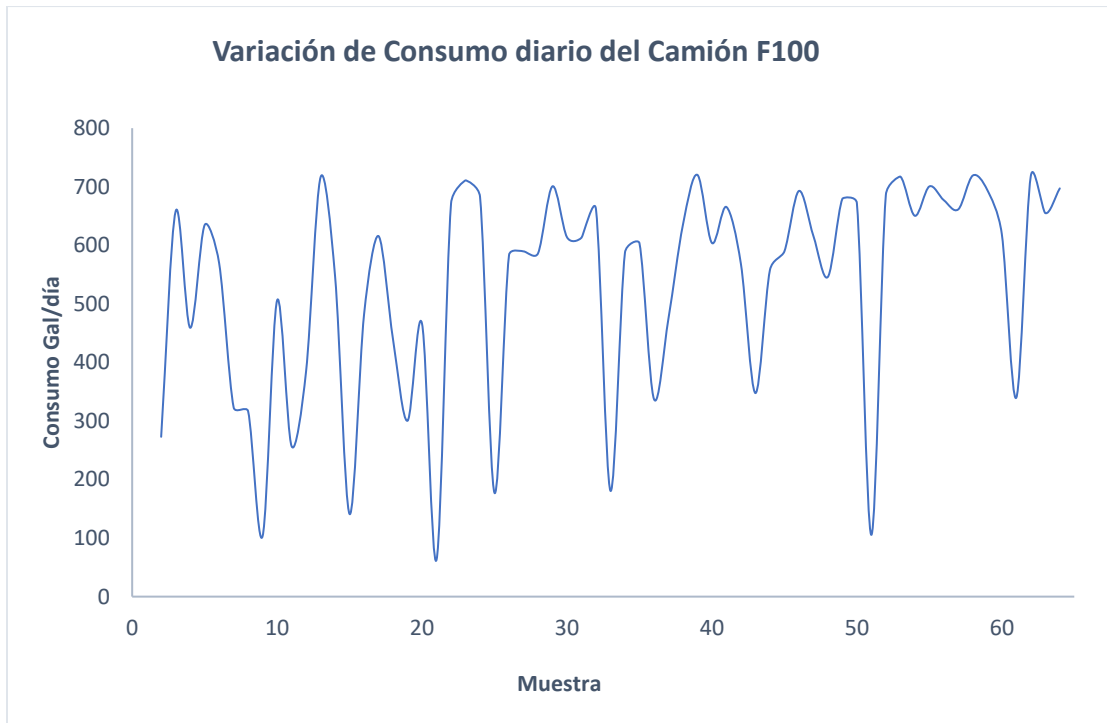


Figura 53. Variación de consumo de combustible diario para del camión F100

Fuente: Elaboración propia

Stand by	-0.576161
Horometro	-0.133672
Suma de Delay	-0.058142
Hang en minutos total	-0.024646
Velocidad cargado	0.041052
Queue en minutos total	0.058552
Tiempo Retrocediendo	0.138470
Suma de Hour	0.169355
Velocidad vacio	0.217636
Use	0.276996
Productividad ciclo acarreo	0.383074
tkph cargado	0.424500
Physical availability	0.492529
tkph vacio	0.545851
Tiempo Cuadrando	0.578643
Use of availability	0.713693
Payload	0.715608
Tiempo Acarreando	0.841393
Consumo	1.000000

Name: Consumo, dtype: float64

Al observar la correlación del consumo de combustible con las variables predictoras, se puede notar que el *Tiempo Acarreando* es determinante en el consumo de combustible. Esta variable está relacionada con la distancia de camino de acarreo, la pendiente, las curvas, las velocidades máximas y mínimas que alcanza los camiones. Por lo tanto, es importante las buenas condiciones de los caminos y sus componentes para evitar un consumo excesivo de combustible.

Otro parámetro no mecánico que influye bastante en la variación del consumo de combustible es el *Tiempo Cuadrando*. Esto puede ser indicio de la mala ubicación para el posicionamiento de los camiones, la superficie en el trabajo no está limpia, la poca habilidad del operador o las malas condiciones meteorológicas. Por lo tanto, implementar un modelo predictivo que ayude a identificar el comportamiento de cada una de estas variables, al mismo tiempo hacer una comparación del consumo real con el predicho, será de gran utilidad para reducir el consumo excesivo de combustible y los efectos negativos que puedan ocasionar.

En el análisis de los datos históricos se reveló que el consumo de combustible por hora (Ver **figura 54**) tenía una variación muy alta llegando a picos de 36 gal/hora hasta un mínimo de 31 gal/hora, una variación significativa a pesar de tratarse de un mismo modelo en los camiones y realizando trabajos similares en la misma operación.

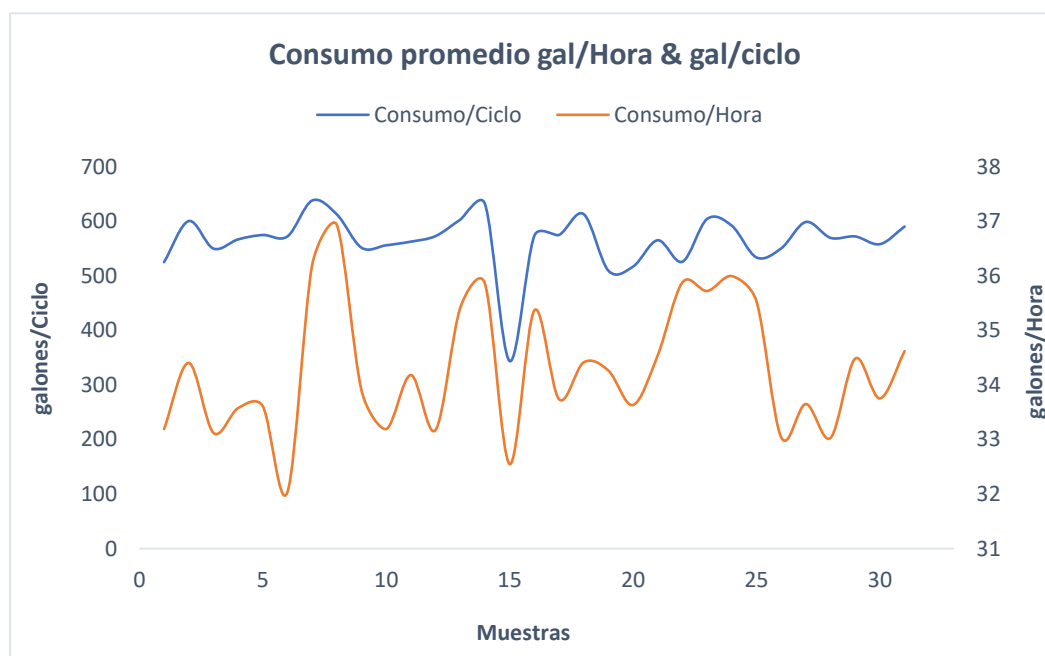


Figura 54. Consumo promedio por hora de los camiones.

Fuente: Elaboración propia

La **Figura 55** nos muestra el histograma del consumo diario de combustible, en ella se observa que la distribución normal parecer estar lejos, más bien podemos notar una alta variación estándar debido principalmente a que el consumo de combustible no tiene una continuidad.

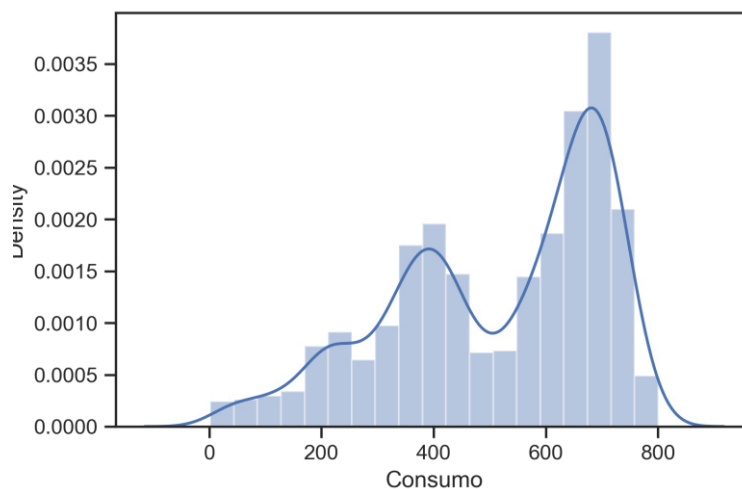


Figura 55. Histograma del consumo diario de combustible de los camiones

Fuente: Elaboración propia

Limpieza de datos

En la actualidad existe muchos métodos para recolectar datos en tiempo real, como es el caso de los sistemas de despacho que nos brindan mucha información acerca de los equipos. Dado la enorme cantidad de datos disponibles que se obtuvo inicialmente, se hizo un análisis riguroso para determinar cuáles serán las variables predictoras para que finalmente desarrollemos el modelo.

Inicialmente se hizo varias pruebas con más de 40 variables predictoras obteniendo resultados poco prometedores. A través de un análisis se identificó que muchas variables tenían el problema de colinealidad afectando el desempeño de los algoritmos. Por lo cual decidimos eliminar variables que estaban directamente relacionadas con la variación de consumo de combustible, y a la vez tenían correlación entre estas mismas. Finalmente se optó por 19 variables predictoras para el desarrollo del modelo.

División de datos

Para la validación de nuestro modelo la división de nuestros datos fue

- 67% de datos para el entrenamiento.
- 33% de datos para las pruebas.

Estandarización de datos

En el caso actual para usar estimadores de machine learning en la predicción de consumo de combustible, escalaremos las características para que se encuentren entre un valor mínimo y máximo, específicamente entre 0 y 1, esto se puede lograr usando MinMaxScaler de la librería Scikit-Learn.

5.3.2.5. Elección de algoritmo

El alcance de esta investigación indica que tiene un alcance exploratorio, con base a ello probaremos algunos algoritmos que hay bajo la categorización de nuestro problema. Para empezar, se hizo prueba con los métodos más simples como es caso del Árbol de Regresión y SVR. Luego se hizo pruebas con algoritmos de ensamble que son más sofisticados como es el caso de Random Forest Regressor, Gradient Boosting Regressor.

5.3.2.6. Modelado

Como punto de partida para el desarrollo de este ejemplo se optó por utilizar el modelo lineal debido a su fácil uso para construir modelos y no requiere ajustes. Como hemos visto en los ejemplos anteriores el modelo lineal sirve para tener una idea general del comportamiento de un conjunto de datos frente a algoritmos de machine learning. Sin embargo, en algunos casos pueden llegar alcanzar desempeños muy altos. Para este ejemplo, los modelos SVR, Random forest y Árbol de regresión se desarrollaron usando los mismos procedimientos descritos en el primer (Ver cap. V, 5.1.2) y segundo (Ver cap. V, 5.2.2) ejemplo.

La **Figura 56** muestra la comparación del modelo SVR frente a la variación real del consumo de combustible. La **Figura 57** muestra la comparación del modelo Random Forest frente a la variación real del consumo de combustible.

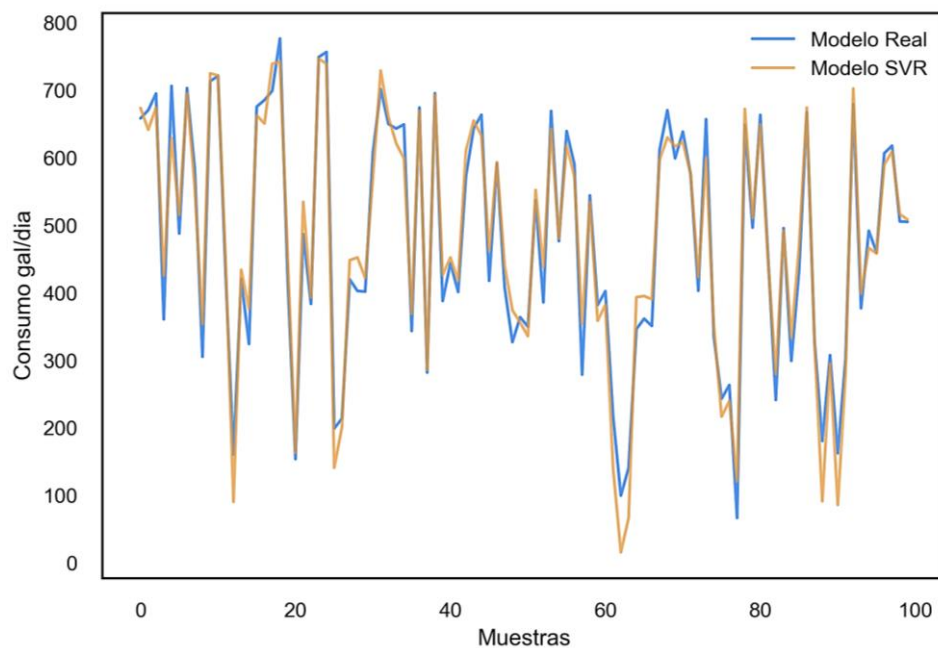


Figura 56. Modelo SVR para la predicción del consumo de combustible

Fuente: Elaboración propia

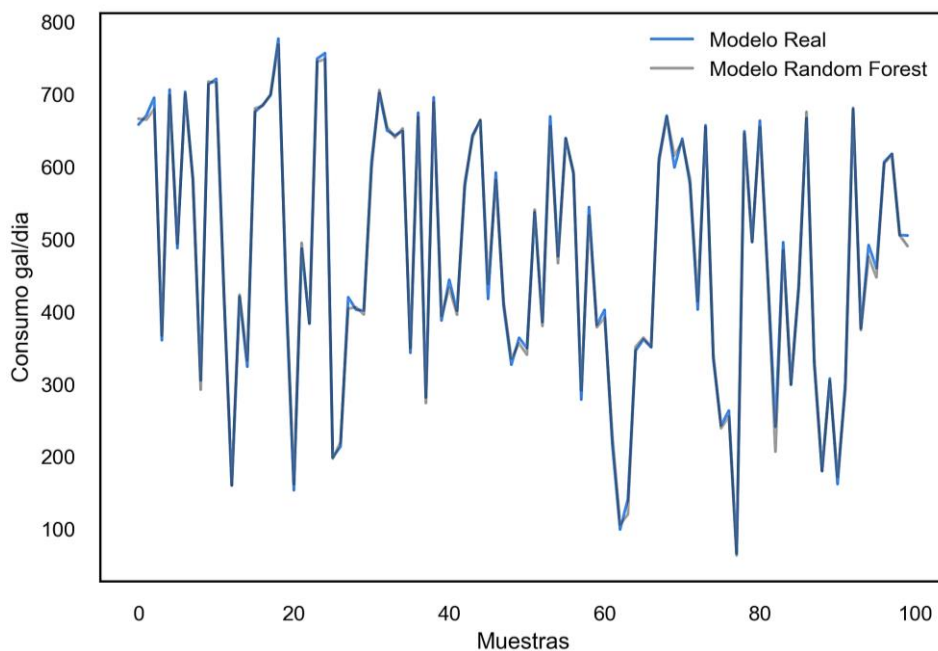


Figura 57. Modelo Random Forest para la predicción del consumo de combustible

Fuente: Elaboración propia

5.3.2.7. Evaluación

Los resultados obtenidos en la evaluación de los modelos se muestran en la **Tabla 12**.

Tabla 12. Resultado de los modelos para la predicción del consumo de combustible

MODELO	METRICAS		
	MAE	RMSE	PUNTAJE (R2) Test
Árbol de Regresión	4.12	4.94	0.9
SVR K=rbf	3.98	4.63	0.94
Randon Forest	0.87	0.91	0.98
Gradient Boosting	0.93	0.95	0.97

Fuente: Elaboración propia

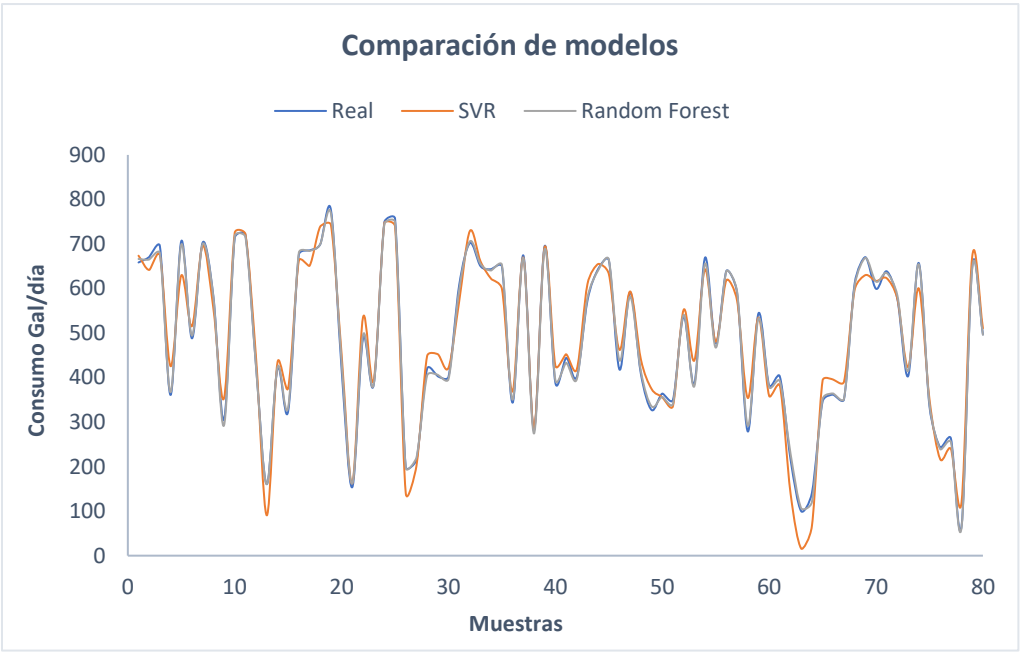


Figura 58. Comparación de modelos para la predicción del consumo de combustible

Fuente: Elaboración propia

5.3.2.8. Predicción

Esta fase implica la predicción con el modelo luego de la implementación al negocio o al proceso para la cual fue desarrollado. Para fines de esta investigación, realizaremos la predicción solo para algunas muestras de datos nuevos.

Tabla 13. Predicción del consumo de combustible

Predicción del consumo de combustible			
Muestras	Volquete	Real gal/día	Predicción gal/día
1	FC76	383.51	384.45
2	FC94	527.88	528.85
3	FC105	588.23	589.21
4	FC126	299.20	300.18
5	FC76	410.03	411.01
6	FC30	607.20	608.21
7	FC105	693.70	694.71
8	FC104	711.18	712.19
9	FC109	45.07	46.09
10	FC86	384.42	385.45
11	FC127	620.34	621.37

5.3.3. Análisis de resultados

La predicción del consumo de combustible de los camiones mineros de alto tonelaje proporcionó excelentes resultados y, sin duda, tiene el potencial para resolver el problema planteado. Según las métricas de evaluación, el modelo que tuvo mejor desempeño respecto a los otros fue el **Random Forest (R2=0.98, MAE= 0.87, RMSE=0.91)** mostrando una vez más la robustez de los algoritmos de ensamble.

Este resultado es una muestra de cómo se pueden generar información valiosa a partir de datos que en muchas ocasiones no suelen ser usados. La gran cantidad de datos que genera una operación minera puede brindar información que ayuda a mejorar el rendimiento de toda la operación minera, desde los procesos unitarios hasta los resultados generales.

En el análisis de datos, realizado para el desarrollo del modelo, se identificó las fluctuaciones del consumo diario de combustible de los camiones, así como las variables que más influyeron en su fluctuación. La implementación de este modelo en una operación tiene el potencial de verificar las proyecciones del consumo de combustibles con respecto al consumo real. De manera que podemos identificar que camión ha tenido un consumo excesivo, además, identificar las variables más influyeron en su variación. El siguiente paso es tomar acciones inmediatas para mejorar el desempeño o las condiciones de las variables que influyen en el consumo excesivo de combustible y hacer un seguimiento de ello.

Optimizar el consumo de combustible de camiones mineros usando modelos predictivos desarrollados con algoritmos de machine learning no solo puede tener un impacto económico en el costo de operaciones, sino también tiene el potencial de reducir significativamente la emisión del CO₂ ayudando a desempeñar el desarrollo de las operaciones mineras de manera responsable ambientalmente.

En las figuras siguientes (Ver **figura 59, 60, 61**) se muestra el ajuste lineal de los datos reales frente a los datos estimados a partir de los modelos.

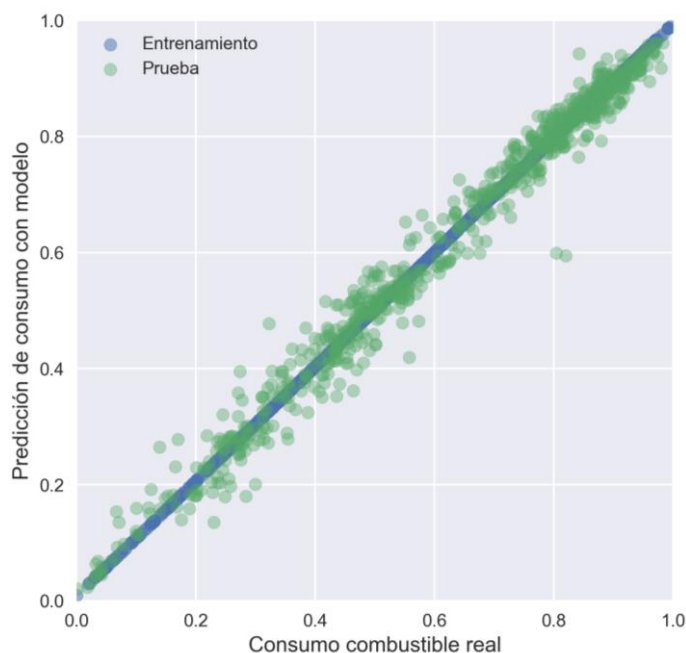


Figura 59. Ajuste de los datos con el modelo GB

Fuente: Elaboración propia

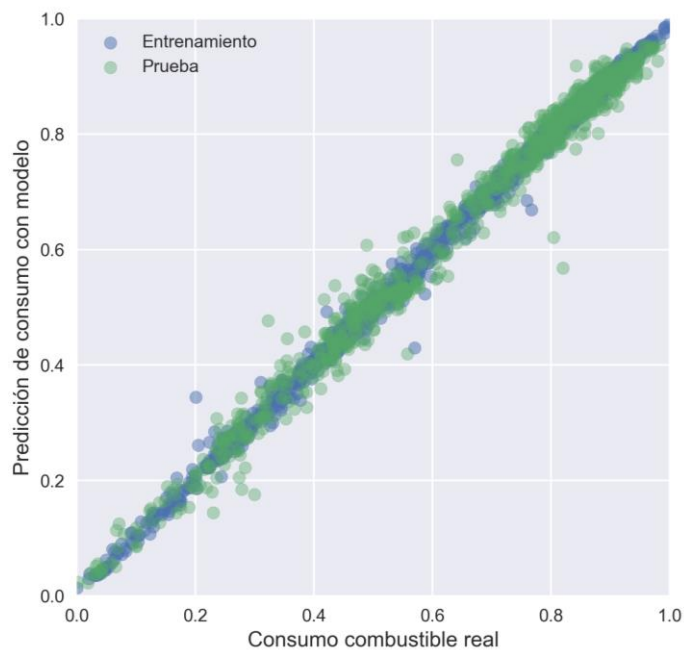


Figura 60. Ajuste de los datos con el modelo RF

Fuente: Elaboración propia

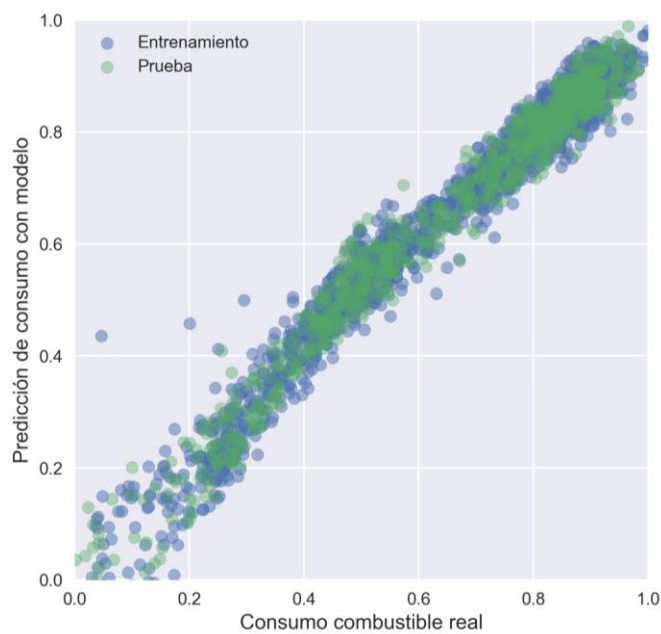


Figura 61. Ajuste de los datos con el modelo SVR

Fuente: Elaboración propia

Capítulo VI: Conclusiones

El principal objetivo de este trabajo ha sido el desarrollo de modelos predictivos en casos de la industria minera mediante el uso de algoritmos de Machine Learning. En particular, el trabajo se ha centrado en la aplicación de algoritmos para realizar tareas de regresión.

Para un mejor entendimiento en el desarrollo de los modelos, se construyó un flujograma a partir de diferentes aplicaciones similares que se identificaron durante la revisión bibliográfica. Por lo cual, se optó por usar términos más conocidos y que se adecuen de manera intuitiva y didáctica para su mejor comprensión en su desarrollo. Finalmente, Quedando *la categorización, determinación de métricas, recolección de datos, análisis de datos, elección de algoritmos, modelado, evaluación del modelo y predicción* como los pasos del flujograma.

La comprensión del problema que se quiere solucionar aplicando Machine Learning es un sub capítulo tan importante y que se debe discutir desde el punto de vista de los diferentes actores. A través de un análisis de diferentes casos y entrevistas a cerca de tecnologías basadas en machine learning, llegamos a la conclusión que las nuevas tecnologías deben cumplir los siguientes objetivos: *Ayuden a mejorar la productividad, Ayude a mejorar la toma de decisiones en un tiempo más corto y Que tenga un retorno de la inversión que se hace en estas aplicaciones.*

Mediante el uso de diferentes algoritmos de machine learning, pudimos identificar las fortalezas y debilidades de estos algoritmos de regresión frente a problemas desarrollados en los ejemplos. Inicialmente, aplicamos algoritmos menos complejos como el SVR, Árbol de Regresión y Regresión Multilineal obteniendo muy buenos resultados. Sin embargo, ante problemas más complejos vimos la efectividad y robustez de los algoritmos de ensamble mejorando los resultados significativamente.

A continuación, se describe los resultados obtenidos en el desarrollo de los casos:

1. Caso 1: Objetivo principal, desarrollar un modelo predictivo para estimar el precio de oro. El modelo desarrollado con la técnica **SVR** tuvo mejor desempeño con las siguientes métricas de evaluación: $R^2=0.94$, $MAE=4.63$ y $RMSE=5.29$.
2. Caso 2: Objetivo principal, desarrollar un modelo predictivo para estimar el contenido de sílice en el concentrado de hierro. El modelo que tuvo mejor desempeño fue el modelo desarrollado con la técnica de **Gradient Boosting Regressor** con las siguientes métricas de evaluación: $R^2=0.51$, $MAE=0.81$ y $RMSE=0.81$.
3. Caso 3: Objetivo principal, desarrollar un modelo predictivo que ayude a predecir el consumo de combustible de camiones mineros de gran tonelaje. Los resultados obtenidos para el modelo desarrollado con el algoritmo de ensamble **Random Forest Regressor** fue superior frente a los otros **algoritmos normales**, obteniendo las siguientes métricas de evaluación: $R^2=0.98$, $MAE=0.87$, $RMSE=0.91$.

La aplicación de algoritmos de machine learning para desarrollar modelos predictivos no se limita al conocimiento único de los procesos de minería, sino también al conocimiento de ciencia de datos. Durante el desarrollo de este estudio identificamos y buscamos herramientas informáticas para su desarrollo y que finalmente se describieron en un subcapítulo (Ver cap. IV, 4.3). Es muy importante ir revisando la descripción, buscar más información y descubrir nuevas herramientas según las necesidades particulares.

Los resultados obtenidos en los ejemplos junto con los casos de implementaciones descritos en el Capítulo 2.1, son una muestra del potencial que tienen las tecnologías basadas en inteligencia artificial y machine learning para la industria minera. Estas aplicaciones no se limitan a los casos presentados, sino que más bien abre un campo de oportunidades en los diferentes procesos que hay en la industria minera en general. Además, las operaciones mineras actuales no siempre comparten las mismas características, abriendo un campo de investigación y desarrollo aún más amplio.

Capítulo VII: Recomendaciones

- Esta investigación fue desarrollada para aquellos interesados que no cuentan con conocimiento previo de la aplicación de machine learning en minería. Es importante que tengan una idea básica de cómo se puede aplicar estas tecnologías en minería, para después buscar ideas y conocimientos más avanzados y que puedan ser implementadas en su trabajo.
- La revisión teórica de Inteligencia Artificial, Machine learning y Deep Learning no se describieron a profundidad, debido a que nuestro principal objetivo no es revisar las matemáticas que llevan, sino de comprenderlas y aplicarlas a los problemas que se pueden presentar dentro de la industria minera. Sin embargo, hace falta conocimientos de estadística y matemáticas para realizar este tipo de trabajo. En la parte de bibliografía se mencionará algunos artículos e informes acerca de los temas mencionados.
- Para obtener los resultados finales se desarrolló un trabajo iterativo con base al flujograma del trabajo presentado. Es importante analizar cada paso para ir investigando y buscando más información acerca del apartado.
- Uno de los pasos más cruciales del flujograma es la categorización del problema, esta nos ayudará determinar las variables que usaremos para generar el modelo. Para ello, se debe tener un claro conocimiento de la industria y del problema que se quiere resolver.
- Los problemas potenciales para la aplicación de algoritmos de machine learning dentro de la industria minera no solo son problemas de regresión. Existen otros casos en donde se pueden aplicar algoritmos de clasificación y de agrupamiento como hemos descrito en los antecedentes del tema.
- Las variables no siempre se comportan de la misma manera al usar diferentes algoritmos, si analizamos la importancia de las variables en cada modelo es más probable que el orden sea diferente.
- Los resultados obtenidos en los casos 2 y 3 no son definitivos para todas las operaciones. Para realizar una implementación se tiene que realizar con datos y características propias de cada operación en particular.

- Para mayor detalle de los resultados obtenidos en cada ejemplo, revisar el apartado de **Análisis de Resultado** preliminar que se hizo para cada una de ellas.
- Es difícil abarcar en su totalidad las diferentes técnicas de machine learning dentro de un solo documento. Es posible que los ejemplos desarrollados sean mejorados utilizando otras técnicas.
- Para las futuras investigaciones nos hemos planteado implementar en alguna operación minera modelos mejorados basados en los ejemplos desarrollados.

Bibliografía

- Akira.AI. (2020). *Akira.AI*. Obtenido de Gradient Boosting: <https://www.akira.ai/glossary/gradient-boosting>
- Almaliki, Z. A. (19 de Mayo de 2019). *Towards Data Science*. Obtenido de Do you know how to choose the right machine learning algorithm among 7 different types?: <https://towardsdatascience.com/do-you-know-how-to-choose-the-right-machine-learning-algorithm-among-7-different-types-295d0b0c7f60>
- Ampadu, H. (01 de Mayo de 2021). *AI Pool*. Obtenido de Random Forests Understanding: <https://ai-pool.com/a/s/random-forests-understanding>
- Bhoraskar, A. (2019). *Prediction of Fuel Consumption of Long Haul Heavy Duty Vehicles using Machine Learning and Comparison of the Performance of Various Learning Techniques*. Delft University of Technology-Delft Mechanical, Maritime and Materials Engineering.
- Castro, Á. (26 de Octubre de 2021). *El gas noticias*. Obtenido de Digitalización: Minería peruana ahorraría US\$500 mil en energía: <https://elgasnoticias.com/digitalizacion-mineria-peruana-ahorraria-us500-mil-en-energia/>
- Collado, C. F., & Baptista Lucio., P. (2014). *Metodología de la Investigación*. México: McGraw W-Hill / Interamericana Editores, S.A.
- Deloitte, & NORCAT. (2020). *Future of mining with AI*. Canadá: Deloitte Design Studio.
- Díaz, M. I. (2019). *Análisis predictivo de activos mineros para la obtención de intervalo de falla mediante algoritmos de machine learning*. Santiago: Universidad de Chile.
- Dindarloo, S., & Irdemoosa, S. (2016). Determinants of fuel consumption in mining trucks. *Department of mining and Nuclear Engineering, Missouri University of Science and Technology*.
- Gamarra, A. F. (2020). *Desarrollo de un modelo para la predicción del precio del cobre empleando herramientas de Machine Learning*. Lima: Pontificia Universidad Católica del Perú- Facultad de Ciencias e Ingeniería.

- Guozhong. (1996). The effects of adding noise during backpropagation training on a generalization performance. En *Neural computation*.
- Harlalka, R. (2018). *Hackernoon*. Obtenido de Choosing the Right Machine Learning Algorithm: <https://hackernoon.com/choosing-the-right-machine-learning-algorithm-68126944ce1f>
- IBM.Cloud.Education. (3 de Junio de 2020). *IBM*. Obtenido de Inteligencia Artificial: <https://www.ibm.com/pe-es/cloud/learn/what-is-artificial-intelligence>
- Jesús Alberto Torres Guerra, D. M. (2021). Geometalurgia y el futuro de la minería digital en el Perú. *Revista del Instituto de investigación de la Facultad de minas, metalurgia y ciencias geográficas de la Universidad Nacional Mayor de San Marcos*.
- Linca, M., & Nanescu, M. (s.f.). *Dacian Cosulting*. Obtenido de Artificial Intelligence Is Changing The Mining Industry: <https://daciandata.tech/artificial-intelligence-is-changing-the-mining-industry-examples-of-successful-applications/>
- McCarthy, J. (2014). What is Artificial Inteligence. *Computer Science Deparment Stanford University*.
- Mercurio, E. (11 de Marzo de 2019). *Minería Chilena*. Obtenido de Análisis de datos para optimizar procesos en la minería: <https://www.mch.cl/2019/03/11/analisis-datos-optimizar-procesos-la-mineria/#>
- Metallurgist911. (17 de Diciembre de 2017). *911metallurgist*. Obtenido de <https://www.911metallurgist.com/metalurgia/procesamiento-minerales-hierro/#CHANCADO-Y-MOLIENDA>
- Osnayo, C. J. (2018). *Aplicación de redes neuronales artificiales para la prediccion de la recuperacion de la planta concentradora de Minsur. S.A - Unidad San Rafael*. Arequipa: Universidad Nacional San Agustín de Arequipa.
- P.Murphy, K. (2012). *Machine Learning a Propabilistic Perspective*. Cambridge,Massachusetts:: The MIT press.

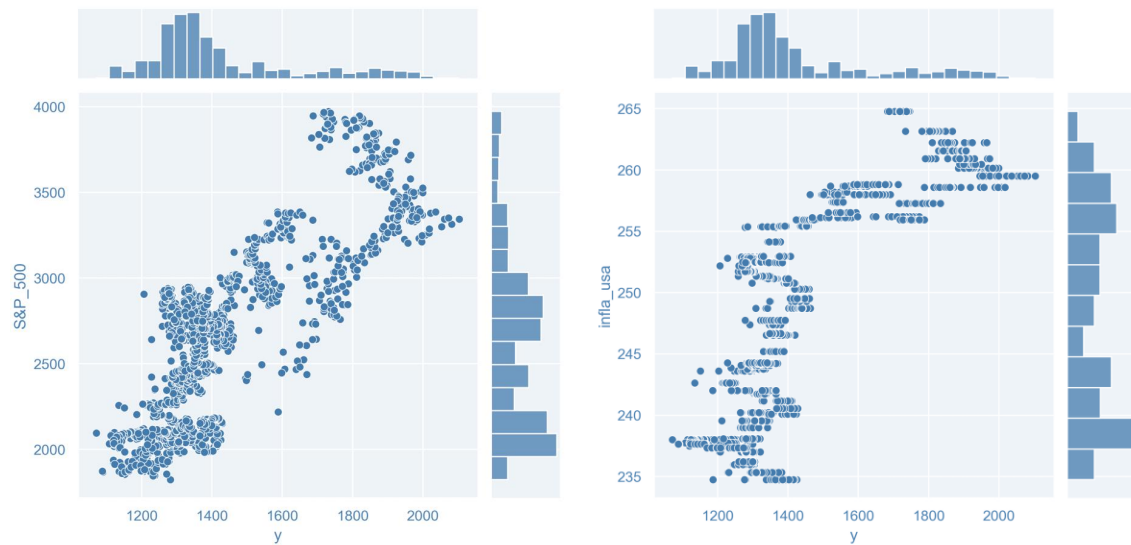
- Palomino, C. T. (2021). *Aplicación de técnicas de análisis de regresión y aprendizaje automático para la estimación de sobre dilución en el método de sub level Stoping - Compañía minera condestable*. Lima: Universidad Peruana de Ciencias Aplicadas.
- Rouse, M. (2021). *TechTarget*. Obtenido de Inteligencia Artificial o IA: https://www.techtarget.com/es/contribuidor/Margaret-Rouse?_gl
- Salgado, J. A. (2018). *Modelo predictivo basado en machine learning de órdenes de trabajo riesgosas para mantenimiento de equipos mineros*. Santiago: Universidad de Chile.
- scikit-learn. (2021). *Choosing the right estimator*. Obtenido de https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html
- Sepúlveda, L. (29 de Octubre de 2019). *Horizonte Minero*. Obtenido de Inteligencia Artificial en minería y energía: <https://www.horizonteminero.com/inteligencia-artificial-en-mineria-y-energia/>
- Technology, M. (s.f.). *Mining Technology*. Obtenido de <https://www.mining-technology.com/artificial-intelligence-in-mining/>
- Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind* 49, 433-460.
- Underwood, C. (17 de Mayo de 2019). *Emerj Artificial Intelligence*. Obtenido de The Future of AI in Heavy Industry – Agriculture, Construction, Mining, and Beyond.: <https://emerj.com/ai-podcast-interviews/future-ai-in-heavy-industry/>
- Walker, J. (02 de Febrero de 2019). *Emerj Artificial Intelligence*. Obtenido de AI in Mining – Mineral Exploration, Autonomous Drills, and More.: <https://emerj.com/ai-sector-overviews/ai-in-mining-mineral-exploration-autonomous-drills/>
- Wikipedia, C. d. (febrero de 2021). *Cross Industry Standard Process for Data Mining*. Obtenido de <https://es.wikipedia.org/w/index.php?title=Cross-Industry-Standard-Process-for-Data-Mining>
- Yasad, S. (2020). *Saed Yasad*. Obtenido de An Introduction to Data Science: http://www.saedsayad.com/data_mining_map.htm

Anexo

Anexo 1. Correlación de las variables del Ejemplo 1

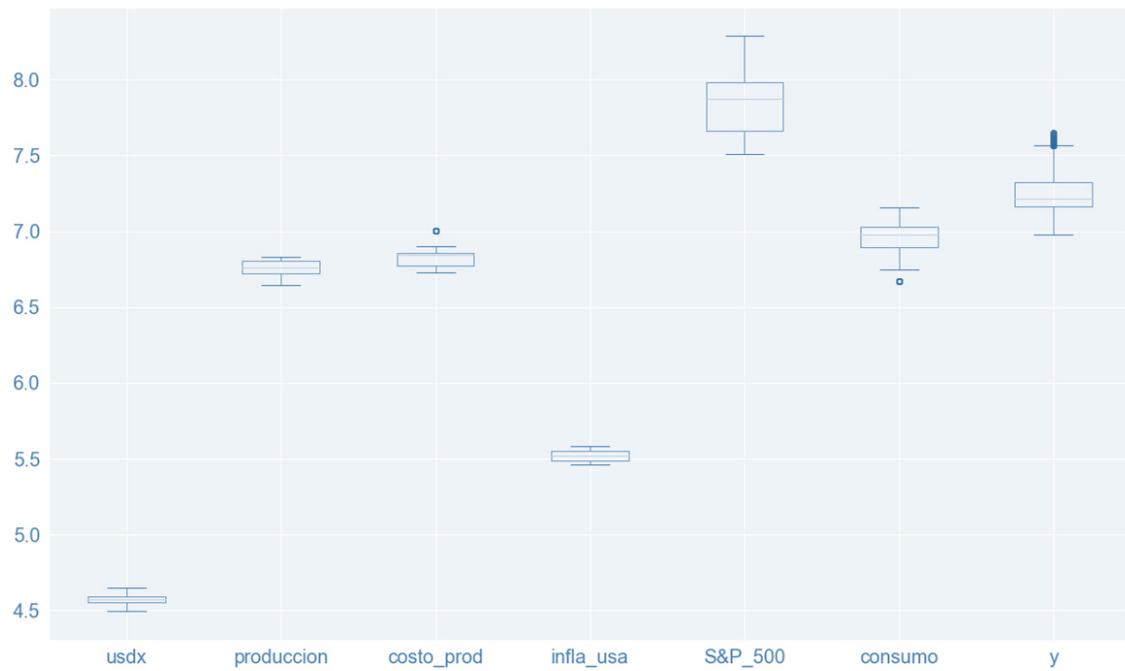
	usdx	tasa_FED	pbi_usa	pbi_china	produccion	costo_prod	infla_usa	S&P_500	consumo	prec_petri	y
usdx	1.000000	0.007501	-0.112943	-0.350504	-0.159978	-0.231443	-0.216947	-0.357181	0.376528	-0.408400	-0.323143
tasa_FED	0.007501	1.000000	0.100197	-0.109225	0.369350	-0.094484	0.362753	0.220329	0.412559	0.487413	-0.195782
pbi_usa	-0.112943	0.100197	1.000000	0.144544	0.077565	0.005552	-0.007291	0.036545	-0.029936	0.199439	-0.145644
pbi_china	-0.350504	-0.109225	0.144544	1.000000	0.178206	0.099714	-0.097428	0.061008	-0.138260	0.502606	-0.107453
produccion	-0.159978	0.369350	0.077565	0.178206	1.000000	-0.175940	0.258302	0.237981	0.001799	0.283294	0.027420
costo_prod	-0.231443	-0.094484	0.005552	0.099714	-0.175940	1.000000	0.676172	0.735606	-0.500076	-0.034916	0.731337
infla_usa	-0.216947	0.362753	-0.007291	-0.097428	0.258302	0.676172	1.000000	0.951625	-0.402424	0.166137	0.779797
S&P_500	-0.357181	0.220329	0.036545	0.061008	0.237981	0.735606	0.951625	1.000000	-0.523759	0.234091	0.816325
consumo	0.376528	0.412559	-0.029936	-0.138260	0.001799	-0.500076	-0.402424	-0.523759	1.000000	-0.054708	-0.698177
prec_petri	-0.408400	0.487413	0.199439	0.502606	0.283294	-0.034916	0.166137	0.234091	-0.054708	1.000000	-0.110938
y	-0.323143	-0.195782	-0.145644	-0.107453	0.027420	0.731337	0.779797	0.816325	-0.698177	-0.110938	1.000000

Anexo 2. Histograma y gráfica de dispersión de las variables Ejemplo 1





Anexo 3. Diagrama de Caja de las variables para el Ejemplo 1

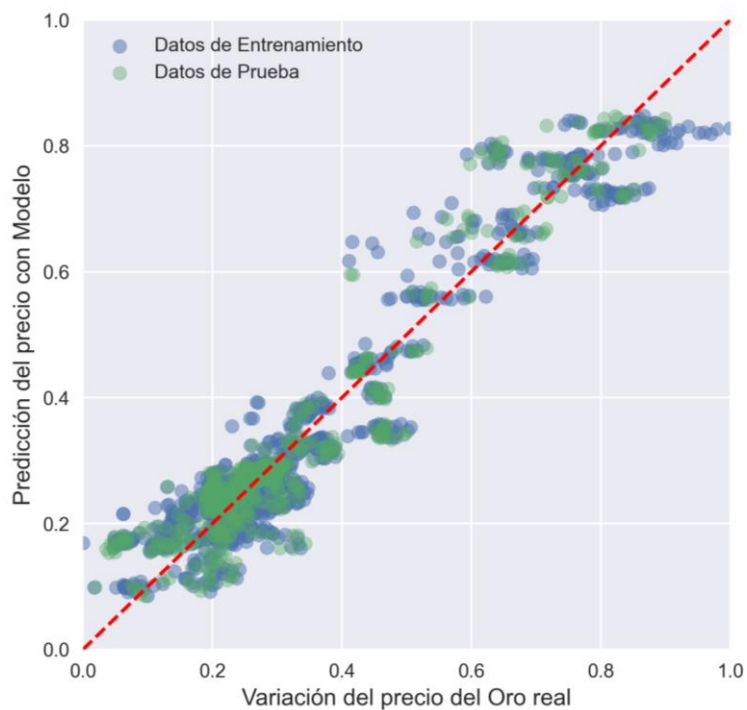
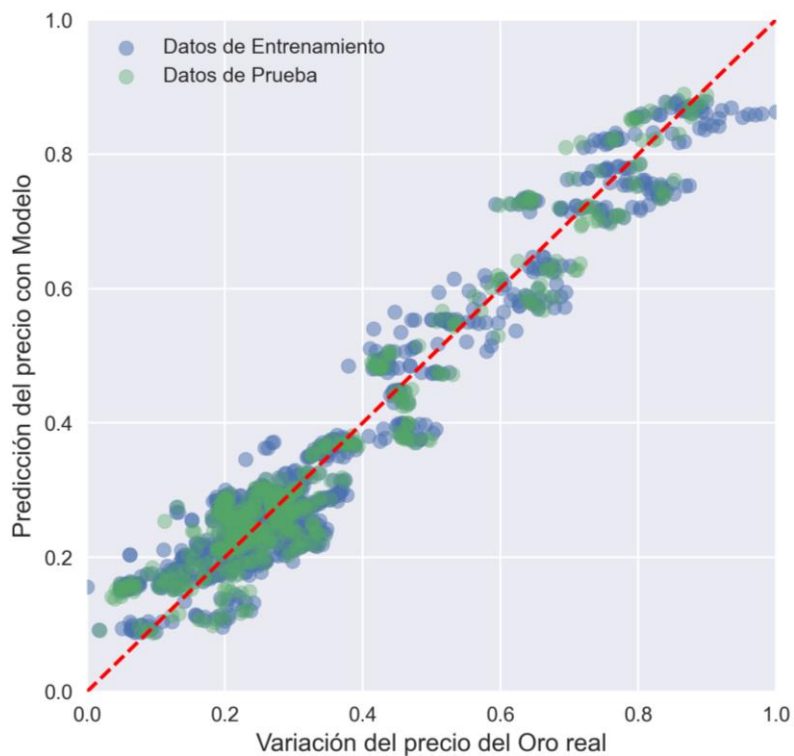


Anexo 4. Principales indicadores estadísticos de las variables para el Ejemplo 1

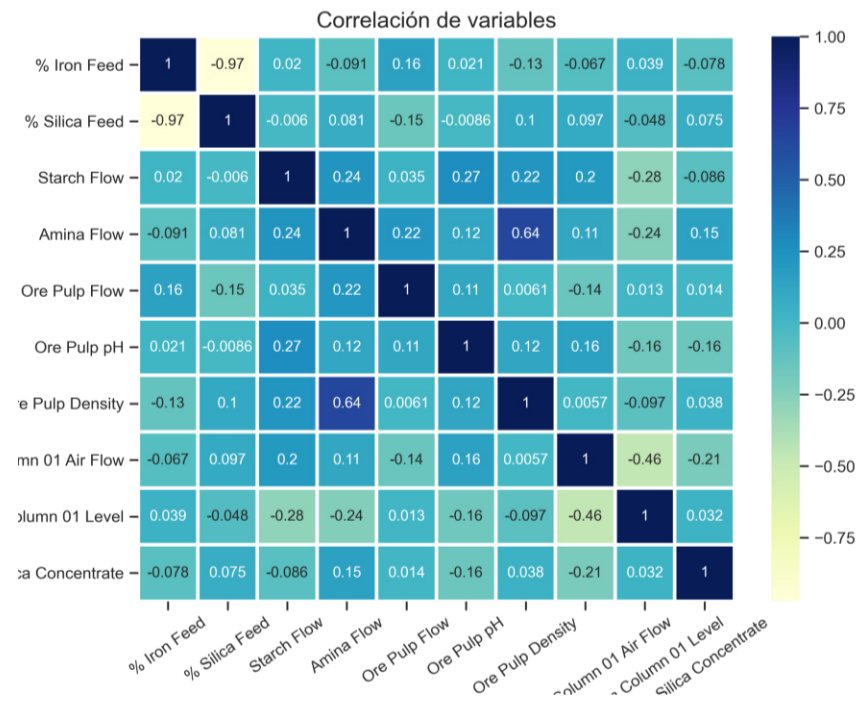
	usdx	tasa_FED	pbi_usa	pbi_china	produccion	costo_prod	infla_usa	S&P_500	consumo	prec_petr1	y
count	1653.000000	1653.000000	1653.000000	1653.000000	1653.000000	1653.000000	1653.000000	1653.000000	1653.000000	1653.000000	1653.000000
mean	95.832002	0.947078	1.702420	6.275378	864.300901	918.823351	248.577217	2628.703745	1056.075644	51.292946	1427.845856
std	2.964747	0.827448	9.316173	3.507625	44.119909	58.473450	8.488296	510.851722	114.177139	10.912866	213.246643
min	88.510000	0.050000	-32.900000	-6.800000	768.170000	837.000000	234.750000	1824.500000	789.030000	-37.630000	1070.800000
25%	93.810000	0.120000	1.400000	6.400000	828.750000	875.000000	240.220000	2128.000000	985.070000	44.980000	1291.900000
50%	96.090000	0.650000	2.200000	6.700000	862.220000	935.000000	249.300000	2632.000000	1069.660000	51.560000	1354.900000
75%	97.700000	1.690000	3.100000	6.800000	903.880000	950.000000	256.120000	2923.750000	1125.240000	58.940000	1515.800000
max	103.610000	2.420000	33.400000	18.300000	927.730000	1100.000000	264.790000	3974.000000	1284.070000	76.410000	2103.200000

Anexo 5. Principales indicadores estadísticos de las variables para el ejemplo 2

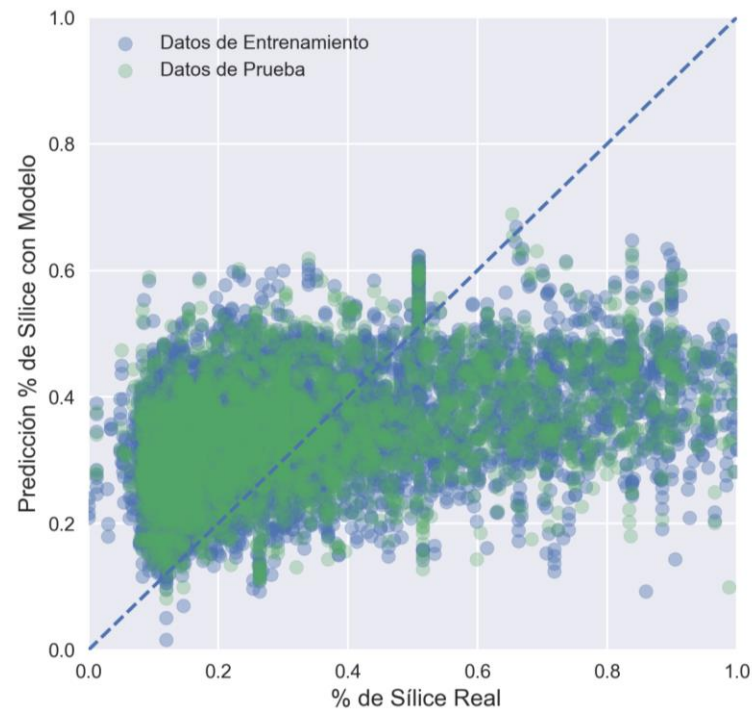
	% Iron Feed	% Silica Feed	Starch Flow	Amina Flow	Ore Pulp Flow	Ore Pulp pH	Ore Pulp Density	Flotation Column 01 Air Flow	Flotation Column 01 Level	% Silica Concentrate
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
mean	56.324172	14.642105	2865.001842	485.831770	397.485708	9.765803	1.679146	280.199668	521.783943	2.296640
std	5.198152	6.841805	1216.035036	91.855674	9.632551	0.387922	0.070118	29.978729	131.565213	1.113268
min	42.740000	1.310000	0.311761	241.699000	376.254000	8.753391	1.519832	175.820296	157.636355	0.600000
25%	52.655000	8.870000	2062.805000	426.562000	394.077000	9.527128	1.645525	250.320750	417.421000	1.430000
50%	56.080000	13.950000	3020.685000	502.319000	399.240500	9.793670	1.697020	299.357500	492.041500	1.960000
75%	59.720000	19.600000	3734.092500	552.190500	402.892250	10.033800	1.728075	300.148250	595.145000	2.932500
max	65.780000	33.400000	6293.142347	739.369400	418.550714	10.807492	1.851990	369.885072	861.949833	5.530000

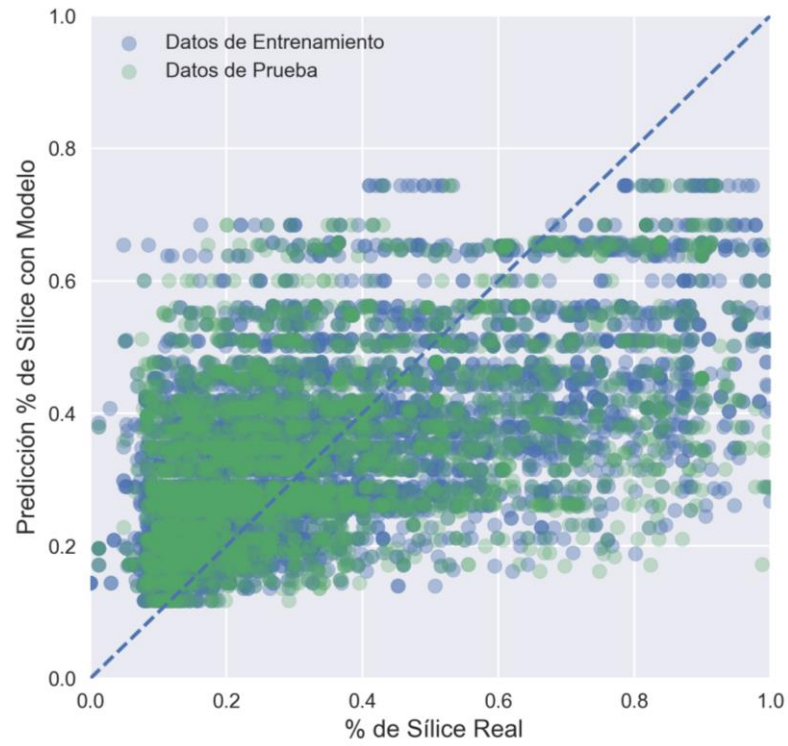
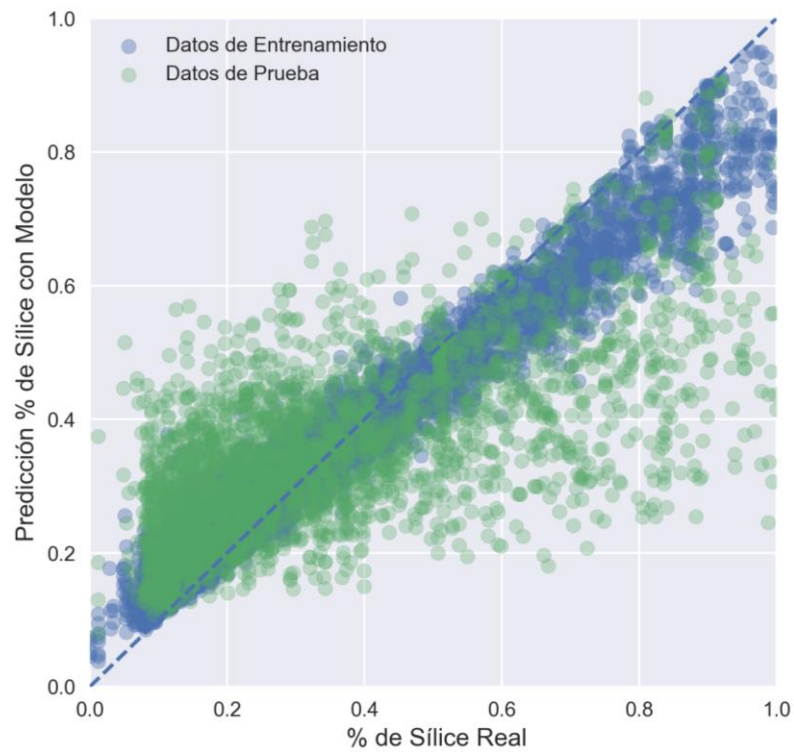
Anexo 6. Ajuste de los datos con el Modelo Lineal Ejemplo 1**Anexo 7. Ajuste de los datos con el Modelo SVR Ejemplo 1**

Anexo 8: Correlación de las variables del concentrado de Hierro.

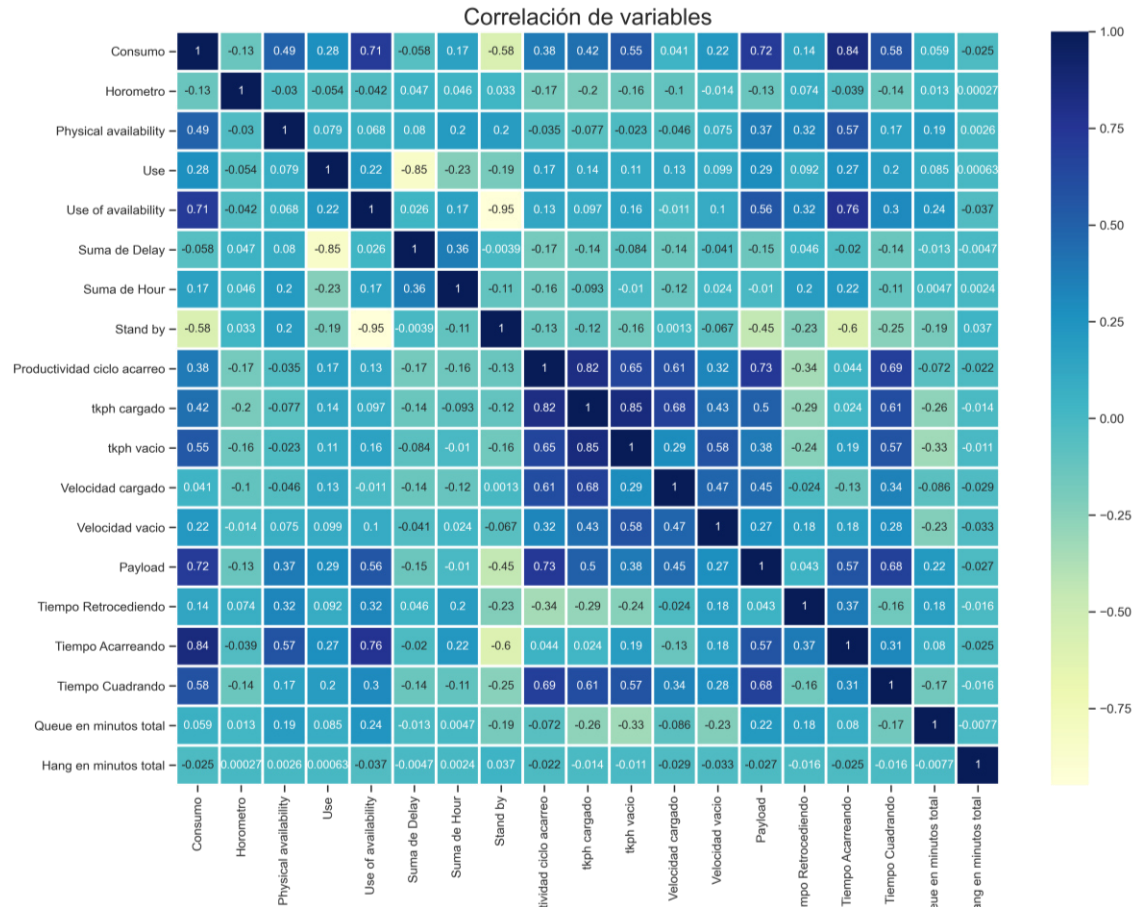


Anexo 9: Ajuste de los datos con el Modelo Multilineal Ejemplo 2



Anexo 10: Ajuste de los datos con el Modelo Árbol de regresión Ejemplo 2**Anexo 11: Ajuste de los datos con el Modelo Random Forest Ejemplo 2**

Anexo 12: Correlación de las variables en el consumo de combustible.



Anexo 13. Principales indicadores estadísticos de las variables para el ejemplo 3

	Consumo	Horometro	Physical availability	Use	Use of availability	Suma de Delay	Suma de Hour
count	2667.000000	2667.000000	2667.000000	2667.000000	2667.000000	2667.000000	2667.000000
mean	515.648931	34687.854218	0.931393	0.976882	0.750611	0.347387	464.455568
std	189.041445	17037.120397	0.163999	0.038454	0.174481	0.548782	93.350668
min	2.099821	5284.200000	0.034000	0.136000	0.002000	0.000000	276.000000
25%	374.402372	25550.900000	0.963000	0.970000	0.693500	0.049000	402.000000
50%	575.644995	31207.100000	1.000000	0.987000	0.821000	0.212000	447.000000
75%	678.177134	39949.700000	1.000000	0.997000	0.872000	0.467500	510.000000
max	800.410541	283821.300000	1.000000	1.000000	1.000000	11.628000	1130.000000

	Stand by	Productividad ciclo acarreo	tkph cargado	tkph vacio	Velocidad cargado	Velocidad vacio
count	2667.000000	2667.000000	2667.000000	2667.000000	2667.000000	2667.000000
mean	5.517846	326.214333	2278.963111	4026.402628	17.563378	31.016381
std	4.009399	134.837536	495.766902	767.138573	2.191639	2.439630
min	0.000000	87.588292	1128.065168	1054.501200	8.784500	7.002000
25%	2.870479	210.794962	1890.294689	3785.264513	15.953852	29.574874
50%	3.966581	290.805663	2327.823916	4279.418029	17.650000	31.128111
75%	6.614584	462.218044	2664.711140	4546.050708	19.139164	32.585786
max	23.941662	633.240751	3463.896017	5476.727459	23.390754	40.773618

	Payload	Tiempo Retrocediendo	Tiempo Acarreando	Tiempo Cuadrando	Queue en minutos total	Hang en minutos total
count	2667.000000	2667.000000	2667.000000	2667.000000	2667.000000	2667.000000
mean	4126.977510	1936.037194	25889.751940	870.233199	46.563161	129.302524
std	2030.074566	1286.543800	8631.581933	864.819678	53.081630	4433.222996
min	91.594203	0.000000	290.025000	0.000000	0.033333	0.000000
25%	2661.485508	1150.114000	19802.134845	2.988034	19.828953	10.890157
50%	3804.710145	1627.208000	28340.350950	709.655000	35.022050	27.297700
75%	5932.210145	2331.353139	32179.131155	1431.448088	56.713021	55.009692
max	9109.492754	9168.837568	45096.128630	5347.626000	807.890327	228965.751800