

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INGENIERÍA METALÚRGICA Y DE MATERIALES



**ESTUDIO CONCEPTUAL SOBRE PROYECTO DE
OPTIMIZACIÓN EN LA SUPRESIÓN DEL POLVO MINA Y
PLANTEO DE ALGORITMOS DE APRENDIZAJE DE
MAQUINA PARA PREDECIR CONDICION DE MATERIAL
PARTICULADO EN CMDIC**

MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL DE MINAS

Presentado por

Carlos Ignacio González Arenas

Profesor Guía

Gastón Adriazola González

Profesor patrocinante

Rodrigo Estay Huidobro

Profesor invitado

Víctor Encina Montenegro

Marzo 2022

Agradecimientos

Agradezco a todas las personas que me han rodeado durante el inicio de mi vida y durante su desarrollo, por darme amor y apoyo incondicional. Mis padres, quienes han entregado todo de si para que yo y mis hermanas podamos crecer y ser personas de bien. Mis hermanas, macarena quien siempre ha sido un ejemplo y una guía, aconsejándome en todo momento. Catalina, quien sin duda es mi partner máximo. Mi nina y mi abuela, quienes siempre han cuidado de mi como si fuera su hijo, dándome cariño y apoyo incondicional, y claro, como olvidar las tantas veces que me daban una luquitas y comida para que me fuera a la universidad. Mis amigos del colegio, con quienes he crecido desde segundo básico, amistad que sin duda perdurara en el tiempo. Mis amigos de la universidad, con quien vivimos duras noches de estudio, de trabajos e informes, pero también de muy buenos momentos de ocio en el cerrito de la universidad. Mi tío oscar, tía janny, tía Ivonne y tío Jaime, quienes siempre han cuidado de mi y me han tratado como uno de los suyos. Agradecer a mi casa de estudios y mis profesores, lugar donde pude aprender una gran cantidad de conocimientos y sobre todo lugar que me permitió desarrollar muchas atrás actividades como el deporte y mis primeros trabajos como promotor de la UTFSM y como coordinador de algunos eventos. También agradecer a Don Gastón y a la compañía, por darme la oportunidad de realizar mi primera practica y quienes me propusieron el tema de memoria a desarrollar, razón por la cual actualmente me desempeño felizmente en este lugar. No olvidar a María Ignacia, con quien viví una muy linda y amorosa etapa, y me apoyo en mis momentos más difíciles. Finalmente agradecer a Dios, ya que siempre he sentido una energía que siempre me ha acompañado y me ha dado fuerza para seguir luchando por mis sueños. Muchas gracias a todo el que se dé el tiempo de leer esto, siempre estarán en mi corazón.

Resumen Ejecutivo

En el presente trabajo se plantea la conceptualización de un nuevo modelo de trabajo el cual propone el monitoreo dinámico de la concentración de polvo en los distintos caminos y sectores productivos de la Compañía Minera Doña Inés de Collahuasi. Por otra parte, se proponen dos metodologías de aprendizaje automático, LSTM (Long short time memory) y SVM (Support vector machine), las cuales predicen los niveles futuros de material particulado en base a datos históricos de concentración de polvo y variables meteorológicas. Los mejores resultados se obtienen con *support vector machine* con el modelo de kernel radial o RBF, el cual es capaz de predecir un día de datos con un cien por ciento de certeza. Esta herramienta permite anticiparse a las condiciones desfavorables de polución de material particulado para la operación, a través de alertas preventivas, lo que permite optimizar la gestión de las herramientas de abatimiento de polvo fino que dispone la compañía.

Abstract

In the present work, the conceptualization of a new work model is proposed, which proposes the Dynamic monitoring of dust concentration in the different roads and productive sectors of the Doña Inés de Collahuasi Mining Company. Furthermore, two machine learning methodologies, LSTM (Long short time memory) and SVM (Support vector machine), are proposed, which predict the future level of particulate matter based on historical dust concentration data and meteorological variables. The best results are obtained with support vector machine with the radial kernel or RBF, which is able to predict a data day with a hundred percent certainty. This tool allows to anticipate the unfavourable conditions of pollution of particulate material for the operation through preventive alerts, which allows to optimize the management of the fine dust abatement tools available to the company.

Glosario

- **ML:** Machine Learning
- **PM:** Particulate Matter
- **ANN:** Artificial Neural Network
- **RNN:** Recurrent Neural Network
- **LSTM:** Long Short Time Memory
- **SVM:** Support Vector Machine
- **CGR:** Ciclo de gestión de riesgos
- **MSE:** Mean squared error
- **RMSE:** Root Mean Squared Error
- **MAE:** Mean absolute error
- **AOD:** Optical aerosol Depth
- **ESED:** Empresas de servicios
- **CAEX:** Camiones de extracción
- **EMRP:** Estación de monitoreo con representatividad poblacional

Contenido temático

Capítulo 1. Contextualización y planteamiento del problema	9
1.1 Planteamiento de la problemática.....	10
1.2 Hipótesis.....	10
1.3 Objetivos	11
1.3.1 Objetivo Principal	11
1.3.2 Objetivos Secundarios.....	11
Capítulo 2. Generalidades sobre la compañía y el proyecto	12
2.1 Historia.....	12
2.2 Ubicación y accesos	13
2.3 Clima	14
2.4 Ciclo de gestión de riesgos.....	14
2.4.1 Fundamentos del CGR	15
2.4.2 El proyecto dentro del CGR.....	17
2.5 Generalidades sobre el polvo	19
2.5.1 Fuentes de emisión de polvo	20
2.5.2 Problemas a la operación asociados al polvo	21
2.5.3 Afecciones a la salud por material particulado	22
2.5.4 Normas medioambientales sobre material particulado.....	22
2.6 Proceso de regadío en CMDIC.....	24
2.6.1 Flota de equipos regadores	25
2.6.2 Supresores de polvo utilizados.....	26
2.6.3 Criterios de operación de camión supresor	26
2.6.4 Criterios de aplicación de supresor de polvo.....	27
2.6.5 Consumo de agua del proceso de regadío	27
2.7 Compromisos medioambientales	28
Capítulo 3. Conceptualización	30
3.2 Conceptualización proyecto de optimización.....	31
3.2 Modelo de trabajo	34
3.2.1 Descripción modelo de trabajo.....	34
3.2.2 Variables a medir	35
3.2.3 Tipos de redes a implementar.....	35
3.2.4 Variables objetivo del plan piloto	35
Capítulo 4. Marco Teórico	36

4.1 Estudios previos	36
4.2 Series de tiempo	40
4.3 Tipos de aprendizaje automático.....	41
4.4 Redes neuronales.....	42
4.5 Redes neuronales recurrentes	43
4.6 Long-Short Time Memory (LSTM).....	45
4.7 Support Vector Machine (SVM).....	48
4.8 Medidas de rendimiento	51
Capítulo 5. Metodología y Datos	53
5.1 Conjunto de datos.....	53
5.2 Análisis de datos.	60
5.3 Implementación de las metodologías	63
5.4 Entorno de ejecución.....	68
Capítulo 6. Experimentos y resultados.....	69
6.1 LSTM	69
6.2 SVM	72
Capítulo 7. Conclusiones y recomendaciones	76
Referencias bibliográficas	78
Anexos.....	81
1. Código LSTM	81
2. Código SVM	88

Índice de figuras

Figura 1. Imagen satelital Google Earth. Fuente: Google.....	13
Figura 2. Representación esquemática del ciclo de gestión de riesgos. Fuente: CMDIC.	17
Figura 3. Línea de acción del proyecto en el CGR. Fuente: elaboración propia.....	18
Figura 4. Camión regador Komatsu HD 785-7. Fuente: CMDIC.	25
Figura 5. Imagen referencial de camión regador. Fuente: CMDIC.....	27
Figura 6. Imagen referencial de CAEX en tránsito. Fuente: CMDIC	32
Figura 7. Esquema flujo de información. Fuente: elaboración propia.	34
Figura 8. Diagrama modelo de trabajo. Fuente: elaboración propia.	34
Figura 9. Estructura generalizada de una red neuronal. Fuente: elaboración propia.....	43
Figura 10. Representación esquemática de red neuronal recurrente.	44
Figura 11. Celda de memoria del modelo LSTM.....	47
Figura 12. Escenario sencillo de un conjunto de datos. Fuente: elaboración propia.....	48
Figura 13. Puntos en espacio unidimensional. Fuente: elaboración propia.	49
Figura 14. Hiperplano en dos dimensiones. Fuente: elaboración propia.	49

Figura 15. Estación 460 - Conc MP10 [UG/M3]. Fuente: elaboración propia.	53
Figura 16. Estación 460 - Temperatura [°C]. Fuente: elaboración propia.	54
Figura 17. Estación 460 - Presión atmosférica [PA]. Fuente: elaboración propia.	54
Figura 18. Estación 460 - Humedad relativa [%]. Fuente: elaboración propia.	54
Figura 19. Estación 460 - Velocidad del viento [m/s]. Fuente: elaboración propia.	55
Figura 20. Estación 460 - Dirección del viento [°]. Fuente: elaboración propia.	55
Figura 21. Estación Llama - Conc MP10 [UG/M3]. Fuente: elaboración propia.	56
Figura 22. Estación Llama - Temperatura [°C]. Fuente: elaboración propia.	56
Figura 23. Estación Llama - Presión atmosférica [Pa]. Fuente: elaboración propia.	56
Figura 24. Estación Llama - Humedad relativa [%]. Fuente: elaboración propia.	57
Figura 25. Estación Llama - Velocidad del viento [M/s]. Fuente: elaboración propia.	57
Figura 26. Estación Llama - Dirección del Viento [°]. Fuente: elaboración propia.	57
Figura 27. Estación Puma - Conc MP10 [UG/M3]. Fuente: elaboración propia.	58
Figura 28. Estación Puma - Temperatura [°C]. Fuente: elaboración propia.	58
Figura 29. Estación Puma - Presión atmosférica [Pa]. Fuente: elaboración propia.	58
Figura 30. Estación Puma - Humedad relativa [%]. Fuente: elaboración propia.	59
Figura 31. Estación Puma - Velocidad del viento [M/s]. Fuente: elaboración propia.	59
Figura 32. Estación Puma - Dirección del viento [°]. Fuente: elaboración propia.	59
Figura 33. Matriz correlación datos estación 460.	62
Figura 34. Matriz correlación datos estación Llama.	62
Figura 35. Matriz correlación datos estación Puma.	63
Figura 36. Predicción LSTM datos 460.	71
Figura 37. Predicción LSTM datos Llama.	71
Figura 38. Predicción LSTM datos Puma.	72
Figura 39. Predicción SVM datos 460.	74
Figura 40. Predicción SVM datos Llama.	74
Figura 41. Predicción SVM datos Puma.	75

Índice de tablas

Tabla 1. Matriz de riesgos. Fuente: CMDIC.	18
Tabla 2. Controles sobre el riesgo. Fuente: CMDIC.	19
Tabla 3. Niveles de emergencia ambiental de MP2,5 en 24 horas. Fuente: Biblioteca del Congreso Nacional, 2017.	23
Tabla 4. Niveles de emergencia ambiental de MP10 en 24 horas. Fuente: Biblioteca del Congreso Nacional, 2017.	23
Tabla 5. Tabla comparativa revisión bibliográfica.	40
Tabla 6. Estadísticas globales Estación 460.	60
Tabla 7. Estadísticas globales Estación Llama.	61
Tabla 8. Estadísticas globales Estación Puma.	61
Tabla 9. Grado a punto cardinal.	65
Tabla 10. Valor categórico para rangos de MP10.	66
Tabla 11. Medidas de rendimiento LSTM datos 460.	69
Tabla 12. Medida de rendimiento LSTM datos Llama.	69

Tabla 13. Medidas de rendimiento LSTM datos Puma.....	70
Tabla 14. Medidas de rendimiento SVM datos 460.....	72
Tabla 15. Medidas de rendimiento SVM datos Llama.....	73
Tabla 16. Medidas de rendimiento SVM datos Puma.....	73

Capítulo 1. Contextualización y planteamiento del problema

En las operaciones de transporte de material en Compañía Minera Doña Inés de Collahuasi, la roca fragmentada se mueve desde las distintas fases activas del rajo hasta el exterior por medio de camiones de extracción, los cuales circulan por los caminos mineros interiores del rajo y por caminos exteriores que conducen al tranque, botaderos, stocks y los chancadores.

Producto del roce entre los neumáticos y el suelo de roca, se levanta una gran cantidad de material particulado, el cual genera una condición desfavorable para la operación de transporte, disminuye considerablemente la visibilidad de los operadores de CAEX, representa un agente de riesgo para la salud de todos los trabajadores ya que puede adentrarse en las estructuras más pequeñas del sistema respiratorio humano, y además es un riesgo para el medio ambiente en términos de contaminación del aire y por el uso de recursos hídricos.

Dependiendo de las condiciones meteorológicas características de la zona altiplánica, siendo abundantes precipitaciones, nevazones y tormentas eléctricas, es como varía el comportamiento del material particulado. Sin embargo, es en los meses más secos donde se presenta la mayor problemática.

Producto de esto, en la compañía se realizan medidas de control para suprimir el polvo, implementando el riego con agua en los distintos caminos donde transitan vehículos livianos y de alto tonelaje, e implementando riego con sal de magnesio en los caminos exteriores al pit. La disposición de los camiones regadores es discreta y no tiene mayor planificación que regar constantemente los caminos y fases activas.

En este contexto, la Superintendencia de Desarrollo y Servicios Mina, busca mejorar la gestión del proceso, en específico la actividad de regadío mina, apuntando a ser más eficientes en el control del polvo. Para esto se implementa un plan piloto con una empresa externa, la cual, con la integración de distintas tecnologías, permite realizar una mayor cantidad de mediciones de material particulado y dar cuenta de ello a un sistema que almacena y procesa la información en tiempo real. De esta manera es posible levantar alertas sobre las condiciones de material particulado y tomar decisiones que permitan controlar esta condición desfavorable.

La presente memoria realiza una revisión sobre las estrategias de control que actualmente se usan para mitigar el polvo en CMDIC y una conceptualización del nuevo modelo de trabajo. Por medio de aprendizaje de máquina, machine learning, se propone un modelo basado en distintos algoritmos computacionales que permitan procesar la información proveniente de la operación y de los instrumentos de medición de material particulado. De esta manera la “máquina”, basada en el entrenamiento realizado con datos históricos multivariantes, podrá predecir la condición futura de material particulado, generando alertas preventivas tempranas para los supervisores. Esto permitirá gestionar de forma más rápida y automática los riesgos y las estrategias de control del polvo, siendo más eficientes en la reducción de material particulado en el aire, en la utilización de recursos, el cuidado de la salud de los trabajadores y en el cumplimiento de los compromisos medioambientales acordados.

1.1 Planteamiento de la problemática

Actualmente la compañía cuenta con solo 2 estaciones fijas de monitoreo de material particulado, las cuales realizan las mediciones y almacenan los datos en el mismo dispositivo, por lo que periódicamente es necesario ir a la estación de monitoreo a recolectar los datos. Estas mediciones solo sirven para elaborar reportes sobre las emisiones atmosféricas pasadas mediante modelos de calidad de aire en conjunto con modelos meteorológicos.

Conforme a la condición actual, los datos recopilados de las estaciones no tienen un propósito mayor más que generar información de forma periódica. Estos no se utilizan para generar algún tipo de aprendizaje en base al comportamiento de las variables.

Frente a esto, la compañía busca ser más eficiente en el control y mitigación del material particulado, para ello se busca optimizar el proceso de regadío actual por medio del monitoreo en tiempo real de las variables críticas y el procesamiento de datos a través de aprendizaje de máquina.

1.2 Hipótesis

Por medio de la conceptualización del nuevo modelo de trabajo y monitoreo en línea de variables críticas desarrollar una herramienta computacional por medio de Machine Learning, que permita predecir la condición de material particulado en el futuro. Anticipando

la condición desfavorable, sería posible optimizar las estrategias de control de polvo. Contribuyendo en la mejora de la calidad del aire que se respira en faena.

1.3 Objetivos

1.3.1 Objetivo Principal

Conceptualizar el nuevo modelo de trabajo que monitorea en tiempo real la concentración de material particulado en los distintos sectores productivos, proponiendo modelos de aprendizaje automático que permitan predecir la condición futura del polvo fino. La finalidad es anticipar las condiciones desfavorables de polución para optimizar la gestión de las herramientas de control de polvo fino.

1.3.2 Objetivos Secundarios

- Analizar el proceso actual de supresión de polvo con camiones regadores de la compañía, para identificar cuáles son las variables principales que determinan el comportamiento del proceso y como este se adaptaría con la implementación de las tecnologías que realizan el monitoreo dinámico.
- Proponer una metodología basada en algoritmos de aprendizaje de máquina que permitan predecir la concentración de material particulado en base a datos históricos multivariados, para establecer alertas preventivas al proceso.
- Verificar la eficiencia de predicción de los modelos con datos de las estaciones de monitoreo actuales y comparar su comportamiento en base a métricas de desempeño como el *Mean Squared Error (MSE)* y el *Root Mean Square Error (RMSE)*.

Capítulo 2. Generalidades sobre la compañía y el proyecto

2.1 Historia

La actividad comercial del distrito minero Collahuasi se inicia en el año 1880, con la explotación de los sistemas de vetas de cobre-plata de alta ley, dicha operación se interrumpe en el año 1930 a raíz de la crisis económica mundial. Las actividades en el área se reiniciaron en 1978, para luego, posterior al año 1991, una combinación de estudios basados en imágenes satelitales, levantamientos aerofotogramétricos, terrestres y actividades de perforación de sondajes, entregaron como resultado el descubrimiento de nuevos yacimientos en el sector.

Se inaugura su operación comercial en abril de 1999, Compañía Minera Doña Inés de Collahuasi explota los yacimientos de cobre pertenecientes al Distrito Minero Collahuasi, ubicados en el altiplano chileno de la Región de Tarapacá, 185 km al sureste de la ciudad de Iquique, entre los 3.800 y 4.600 metros sobre el nivel del mar. Los estudios de factibilidad y de impacto ambiental del Proyecto Collahuasi fueron aprobados en 1995. A fines de 1996, habiéndose suscrito los acuerdos relativos al financiamiento y comercialización, se inició la etapa de desarrollo y construcción del proyecto minero.

Inicialmente, la operación minera (EIA “Proyecto Collahuasi”, 1995) poseía una capacidad nominal de procesamiento para 60 kilotoneladas por día de minerales sulfurados y 15 ktpd de minerales oxidados, aprobado ambientalmente a través del EIA “Proyecto Collahuasi” en el año 1995, el cual fue definido como proyecto original. Posteriormente, tras la aprobación del EIA “Proyecto Expansión 110 ktpd Planta Concentradora Collahuasi” (2001), la producción de la línea de minerales sulfurados aumentó a 110 ktpd. Esta capacidad experimentó un nuevo aumento a 133 ktpd después de un proceso de optimización de la operación aprobado ambientalmente mediante la DIA “Proyecto Optimización Collahuasi” del año 2003. Finalmente fue aprobado un procesamiento de 170 ktpd, mediante la DIA “Proyecto Optimización a 170 ktpd” del año 2010. Actualmente, la faena cuenta con dos yacimientos en explotación de minerales sulfurados de cobre, los rajes Ujina donde se extrae principalmente óxidos y rajo Rosario aportando con la producción de concentrados.

Collahuasi está autorizado a producir anualmente del orden de 600.000 toneladas de cobre fino, contenido en concentrados. Además, produce aproximadamente 13.730 toneladas de

concentrado de molibdeno, promedio anual, el que se obtiene a partir del concentrado colectivo cobre-molibdeno en la planta de Molibdeno en el sector Puerto Patache.

El concentrado colectivo de cobre-molibdeno es enviado como pulpa desde el Sector Cordillera, a través de dos minero ductos de 203 km de longitud, hasta una planta de filtrado y terminal de embarque ubicados en Puerto Collahuasi, en Punta Patache, unos 65 km al sur de Iquique.

2.2 Ubicación y accesos

Los yacimientos de la compañía ocurren en el altiplano andino, en el desierto de Atacama, a una altura promedio de 4400 msnm, una zona caracterizada por tener climas muy variados, desde intensas lluvias en los meses de verano y fuertes nevadas en la época de invierno continental. A 40km de las operaciones mineras se ubica el poblado de Huatacondo, y a 130 y 135km se encuentran las localidades de Pica y Matilla respectivamente. Ambos poblados se encuentran rodeados de zonas de alto valor por su diversidad biológica, flora y fauna característica de la zona, geografías con salares, humedales y bofedales.



Figura 1. Imagen satelital Google Earth. Fuente: Google.

Para acceder a la propiedad minera desde la ciudad de Iquique, se debe salir de la ciudad por la autopista Humberstone-Iquique/ Ruta 16, hasta la Ruta 5 en dirección hacia el sur. Luego, tomar el acceso a la ruta A-687, la cual se transforma en la ruta A-651 hasta llegar a la garita norte del control de acceso a Collahuasi.

Desde la ciudad de Antofagasta, se debe salir de la ciudad por la Ruta 26 hasta la Ruta 5 en dirección al norte, por esta se debe tomar el acceso a la Ruta 25, la cual se transforma en la Ruta 21 hasta llegar a la comuna de Ollagüe, en este lugar tomar el camino B-15-A el cual conduce hasta el sector cordillera de la compañía.

Otra forma de acceder al área cordillerana es por vía aérea, ya que la compañía cuenta con un aeródromo. Sobre el minero ducto, cuenta con un camino de servicio con múltiples accesos entre el sector cordillera y el puerto, es utilizado principalmente para labores de observación, inspección y mantención de la tubería.

2.3 Clima

Las operaciones del sector cordillera de la compañía se ven afectadas por las particularidades del clima altiplánico y continental. Por esto, la faena se ve enfrentada a dos tipos de invierno:

- Invierno continental, el cual se presenta entre los meses de mayo a septiembre y se caracteriza por intensas lluvias y bajas temperaturas que propician nevazones.
- Invierno altiplánico, el cual se presenta entre los meses de noviembre a abril y se caracteriza por ser un período de constantes tormentas eléctricas, lluvia e incluso nevazones eventuales.

Ambos períodos son definidos por el Comité de Operación Invierno (COI), el cual determina la fecha de inicio y termino de cada temporada de invierno.

2.4 Ciclo de gestión de riesgos

El ciclo de gestión de riesgos es un modelo de trabajo propio de la compañía, se fundamenta en definiciones técnicas del círculo de Deming, la norma ISO 9001, OSHAS 18001, ISO 1400, entre otras normas técnicas asociadas, buscando calidad y simpleza en su aplicación, y que involucre de manera activa a todos los trabajadores. Comenzó a implementarse el 2013, su objetivo era definir una base que permita gestionar efectivamente el riesgo en todos los procesos de la cadena de valor de la compañía.

Esta metodología asume que los procesos están enfrentados continuamente a incidentes, los que generan la variabilidad de los procesos y actividades, teniendo una diferencia significativa entre los resultados reales y presupuestados. La clave del éxito es que el 100%

de la organización este focalizada en los procesos, planificándolos, controlándolos y generando aprendizajes, para sistemáticamente ir mejorando y disminuyendo la incertidumbre de cada proceso.

En la compañía, el ciclo de gestión de riesgos, CGR, se ha convertido en la filosofía de cómo hacer las cosas, en conjunto con los colaboradores y las empresas de servicios externos (ESED), entienden como reducir la incertidumbre de los procesos a nivel de la seguridad de los trabajadores y la integridad de sus activos. Esto se traduce en cero horas perdidas para todo el año 2018 por incidentes hacia los trabajadores de la compañía y externos.

2.4.1 Fundamentos del CGR

El CGR se fundamenta en 4 grandes ámbitos: planificación, ejecución, verificación y comunicación y aprendizaje.

- **Planificación**

Se define en tres partes

1. Mapa de procesos, que define la cadena de valor de la compañía y sus procesos de apoyo.
2. Matriz de riesgos, permite identificar y evaluar riesgos a nivel de personas, negocio, medio ambiente y comunidad.
3. Controles y mitigación, están insertos en los estándares de prevención de fatalidades y de seguridad, medio ambiente, salud y operaciones que permiten controlar y minimizar los riesgos en los procesos.

- **Ejecución**

Se especifica el qué y cómo debe considerar un trabajador al momento de realizar las actividades y tareas que componen un proceso, con especial énfasis en la aplicación de controles de prevención y mitigación. Se establecen procedimientos de trabajo donde el dueño del proceso deberá asegurar al inicio de la actividad la documentación requerida por el CGR, asegurando la materialización de los compromisos y garantizando el entendimiento por parte del personal que ejecuta la actividad

En la ejecución se debe considerar lo entregado en los documentos integrantes y el material de desarrollo de ingeniería y diseño, complementándose con las experiencias del proponente en buenas prácticas, legislación y normativa vigente como todo aquello que permita asegurar la confiabilidad del proceso de construcción.

Dentro de las herramientas de gestión requeridas y solicitadas por la compañía se cuentan:

- RITUS → Reunión de inicio de turno seguro
- GRT → Gestión de riesgos en terreno

- **Verificación**

Se ejecuta a través de la herramienta que integra el cumplimiento de estándares antes de efectuar una actividad. El dueño del área deberá aplicar la herramienta de verificación y de control necesarias para dar fiel cumplimiento a lo contratado y que es reflejo de la planificación previa como ejecución posterior de los trabajos.

Se utilizan los documentos V.A.T.S, verificación y autorización de trabajo seguro, por medio de este documento, la supervisión solicitante, el ejecutor y dueño del área y/o equipos verifican que los controles establecidos en la planificación del CGR permiten la ejecución de la actividad con los riesgos controlados y dentro de las herramientas de gestión requeridas y solicitadas por la compañía.

- **Comunicación y aprendizaje**

El aprendizaje constante que se genera permite retroalimentar el modelo constantemente, este se nutre de la investigación de incidentes, la gestión de riesgo en terreno (GRT) y reunión de inicio de turno seguro (RITUS), estas últimas se realizan diariamente por la supervisión y trabajadores. Es de vital importancia aplicar las herramientas de esta etapa en cada actividad, debe ser constante y metódico, de tal manera de asegurar la identificación de mejoras y las correctas prácticas.

Una herramienta preventiva, donde se puede difundir el aprendizaje y los controles pertinentes es en la reunión de inicio de turno seguro (RITUS), herramienta clave del

CGR, con foco en el control de actividades y tareas derivadas de la planificación, con una visión más preventiva y de identificación de riesgos y sus mejoras en prácticas similares.

La aplicación de este ciclo, en un ambiente de seguimiento permanente, ha permitido progresar considerablemente en la protección de personas y el negocio. En la siguiente ilustración se representa esquemáticamente el modelo:



Figura 2. Representación esquemática del ciclo de gestión de riesgos. Fuente: CMDIC.

2.4.2 El proyecto dentro del CGR

Dentro del ciclo de gestión de riesgos el proyecto se enmarca en la etapa de planificación, en las cajas 1, 2 y 3. La caja número uno corresponde al mapa de procesos, el proyecto se encuentra dentro de los procesos de la Super Intendencia de Desarrollo y Servicios Mina de la compañía. Dentro del subproceso de Desarrollo Mina, específicamente en la actividad de supresión de polvo con camión regador.

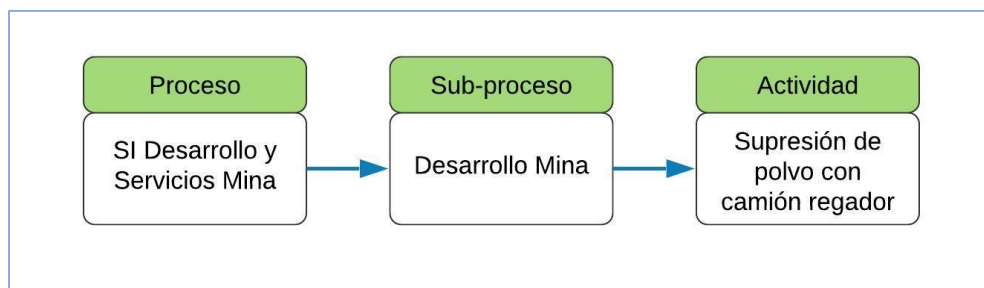


Figura 3. Línea de acción del proyecto en el CGR. Fuente: elaboración propia.

En esta actividad se identifican las siguientes tareas a cumplir diariamente:

1. Chequeo sistema eléctrico, mecánicos y tecnológicos.
2. Revisión de equipo.
3. Carguío del equipo con supresor.
4. Operación del equipo.

En la caja número dos del ciclo de gestión de riesgos, se tiene la matriz de riesgos del subproceso Desarrollo Mina, donde se identifica una serie de peligros relacionados con la actividad de regadío, sin embargo, los que guardan estrecha relación con el proyecto y sus objetivos son:

Tabla 1. Matriz de riesgos. Fuente: CMDIC.

Actividad	Rama del peligro	Familia peligros	Peligro de seguridad y salud
Supresión de polvo con camiones regadores	Salud	Químicos	Polvo fracción respirable
	Ambiental	Consumo de agua	Agotamiento recursos hídricos
		Afectación a la calidad del aire	Emisión de material particulado por tránsito de vehículos en caminos no pavimentados

Para cada uno de estos riesgos se consideran controles administrativos, controles específicos y en algunos casos controles duros o de ingeniería:

Tabla 2. Controles sobre el riesgo. Fuente: CMDIC

Controles administrativos	Controles específicos	Controles Duros
Monitoreo	Vigilancia de salud, exámenes ocupacionales	Respirador doble vía
Administrativos operacionales	Planificar el uso y demanda del recurso hídrico para en lo posible no desperdiciarlo	N/A
Monitoreo	Medición de variables ambientales	N/A

2.5 Generalidades sobre el polvo

El termino polvo no tiene un significado científico muy preciso, pero se suele definir como una pequeña partícula sólida creada como consecuencia de la fractura de otras partículas más grandes. Dependiendo de su tamaño, las partículas pueden ser peligrosas para la salud de los trabajadores, particularmente aquellas que son capaces de suspenderse en el aire. Formado en su mayoría por desintegración, abrasión o fractura de materiales sólidos, estas partículas sólidas tienen diámetros del orden de las micras. Las partículas en el rango de las 60 a 2000 μm pueden suspenderse en el aire a no más de un metro de la superficie, precipitando rápidamente. Los tamaños por sobre los 2000 μm se mantienen en el suelo y solo ruedan por la acción del viento.

Los tamaños más pequeños son capaces de permanecer suspendidos en el aire por horas y presentan un mayor riesgo para el sistema respiratorio cuando estas son inhaladas, causando complicaciones como frecuencia cardiaca irregular, tos, irritación de las vías respiratorias, función anormal del pulmón, dificultad al respirar, ataques cardiacos, asma, silicosis (Xing, et. al. 2016). En general, mientras más pequeñas sean las partículas, mayor es su capacidad para penetrar en los aparatos más pequeños del sistema respiratorio humano.

Se distinguen dos categorías de polvo:

- a) Polvo inhalable - MP10

Es aquella fracción de polvo que entra en el cuerpo, pero es retenida en la nariz, la garganta, y en las vías respiratorias superiores. La medida del diámetro de este polvo es aproximadamente 10 μm y todos los tamaños comprendidos entre 10 y 2,5 μm .

b) Polvo respirable - MP2,5

Son aquellas partículas lo bastante pequeñas como para penetrar la nariz y el sistema superior respiratorio y llegar hasta las estructuras funcionales del pulmón. Las partículas que penetran superan la capacidad natural de despacho del organismo teniendo una mayor probabilidad de quedarse permanentemente en el sistema respiratorio.

2.5.1 Fuentes de emisión de polvo

El polvo mineral procede de las variadas operaciones mineras como resultado de la abrasión o trituración de roca por los distintos procesos mineros. En general, el polvo proviene de las actividades como la tronadura de la mina, traslado y chancado de mineral son algunas de las fuentes de emisión. Por ende, su composición depende del mineral o material tratado en las operaciones o explotación minera.

- Explotación de rajo, polvo proveniente de la tronadura, la extracción del mineral y el carguío de mineral o lastre.
- Transporte de mineral, proveniente de caminos no pavimentados, producto del tránsito de camiones de alto tonelaje y vehículos livianos, emisión desde las ruedas y por el material cargado en las tolvas. Riego con agua por camiones aljibes es el método de control más común en la industria, el uso de supresores de polvo son soluciones complementarias. Esta actividad representa la principal fuente de emisiones de polvo en las faenas, con una contribución que varía entre el 78-97% del total de emisiones de material particulado menor a 10 micrones, según Dust control handbook for industrial mineral mining and processing, 2019.
- Planta de procesamiento, desde los chancadores primarios y por la descarga de mineral en la planta. Sobre la molienda y harneros pueden tener una participación importante dependiendo de las condiciones específicas de cada faena. En la industria se suele usar una combinación de métodos de control, por vía seca (encapsulamiento)

y húmeda (agua pulverizada). Las cantidades emitidas van a depender del tipo de material a trabajar y los requerimientos en el producto de salida.

- Erosión eólica, conocida como fuente fugitiva, afectadas por la acción del viento los depósitos de lastre, acopios de baja ley, pilas, correas transportadoras y depósitos de relave. Algunas prácticas de control son humectación con aspersores y barreras contra el viento como mallas, cercas o vegetación.
- Descarga de material, incluye descarga de lastre en depósito y descarga de mineral de baja ley y oxidado. Como medidas de mitigación común es la humectación con aspersores o recubrimiento con mallas durante la descarga.

Sobre el proyecto en cuestión, se enfoca principalmente en el polvo producido por el transporte de material por camiones de alto tonelaje.

2.5.2 Problemas a la operación asociados al polvo

En contaminación por polvo, se define como material particulado sedimentable, sea cual sea su tamaño, a aquel captado sobre una unidad de superficie en una unidad de tiempo. Las partículas que permanecen en suspensión se clasifican en polvo inhalable MP10 y polvo respirable MP2,5. El primer grupo incorpora la fracción gruesa del polvo, bajo 10 μm y sobre 2,5 μm . Estas partículas son capaces de permanecer en el aire por un periodo de horas hasta días. El MP2,5 representa la fracción fina de polvo, bajo los 2,5 μm , estas son capaces de permanecer suspendidas por periodos de días y hasta semanas, debido a que son capaces de recorrer largas distancias producto de la acción del viento.

Por lo anterior, el material particulado presenta una serie de desventajas para la operación minera a cielo abierto, en lo que concierne a la actividad de transporte por medio de camiones de extracción, el material particulado suspendido en el aire reduce la visibilidad de los operadores, debido a que el polvo en suspensión tiene propiedades de adsorción y refracción de luz. Si bien no existe una correlación definida entre las condiciones del polvo e incidentes de CAEX o vehículos livianos, sin embargo, la observación y experiencia dictan que condiciones extremas de polvo en suspensión aumentan la probabilidad de que se produzca una colisión o algún incidente provocado por esta condición (Cecala, et al, 2019)

2.5.3 Afecciones a la salud por material particulado

Conforme a lo establecido por la Agencia de Protección Medioambiental de Estados Unidos (EPA), el tamaño de las partículas se encuentra directamente vinculado con el potencial de provocar problemas de salud, donde las partículas con tamaños inferiores a 10 micrones de diámetro suponen los mayores problemas, ya que pueden llegar a la profundidad de los pulmones e incluso alcanzar el torrente sanguíneo.

Los problemas asociados a la salud de las personas son variados, múltiples estudios científicos vinculan la exposición a la contaminación por partículas a una variedad de problemas que incluyen:

- Muerte prematura en personas con enfermedades cardíaca y pulmonares.
- Infartos de miocardio no mortales.
- Frecuencia cardíaca irregular.
- Asma agravada.
- Función pulmonar reducida.
- Síntomas respiratorios aumentados, como irritación en las vías respiratorias, tos o dificultad para respirar.

2.5.4 Normas medioambientales sobre material particulado

El decreto supremo N° 12, en su artículo tercero, establece la norma primaria de calidad del aire para material particulado fino MP2,5, instaura una concentración límite de veinte microgramos por metro cúbico ($20 \mu\text{g}/\text{m}^3$), como concentración anual, y cincuenta microgramos por metro cúbico ($50 \mu\text{g}/\text{m}^3$) como concentración de 24 horas.

Se considera sobrepasada la norma primaria de calidad del aire para material particulado fino respirable en los siguientes casos:

Caso 1: Cuando el percentil 98 de los promedios diarios registrados durante un año, sea mayor a $50 [\mu\text{g}/\text{m}^3]$, en cualquier estación monitorea calificada como una estación de monitoreo de material particulado con representatividad poblacional (EMRP).

Caso 2: Cuando el promedio tri-anual de las concentraciones anuales sea mayor a $20 [\mu\text{g}/\text{m}^3]$ en cualquier estación monitorea calificada como EMRP.

Esta misma normativa define los niveles que determinan situaciones de emergencia ambiental para material particulado MP2,5.

Tabla 3. Niveles de emergencia ambiental de MP2,5 en 24 horas. Fuente: Biblioteca del Congreso Nacional, 2017.

Nivel	Concentración 24 horas MP2,5 [$\mu\text{g}/\text{m}^3$]
Alerta	80-100
Preemergencia	110-169
Emergencia	170 o superior

Con respecto al material particulado MP10, el decreto N° 59, en el título 2, artículo segundo establece la norma primaria de calidad del aire para el contaminante material particulado respirable MP10, es ciento cincuenta microgramos por metro cubico normal ($150 [\mu\text{g}/\text{m}^3\text{N}]$) como concentración de 24 horas.

Se considera sobrepasada la norma de calidad del aire para material particulado respirable cuando el percentil 98 de las concentraciones de 24 horas registradas durante un periodo anual en cualquier estación de monitoreo calificada como EMRP sea mayor o igual a $150 [\mu\text{g}/\text{m}^3\text{N}]$

Esta misma normativa define los niveles que determinan situaciones de emergencia ambiental para material particulado MP10.

Tabla 4. Niveles de emergencia ambiental de MP10 en 24 horas. Fuente: Biblioteca del Congreso Nacional, 2017.

Nivel	Concentración 24 horas MP10 [$\mu\text{g}/\text{m}^3$]
Alerta	195-239
Preemergencia	240-329
Emergencia	330 o superior

2.6 Proceso de regadío en CMDIC

El proceso de regadío tiene su foco en la continuidad operacional, controlando los riesgos asociados y energías potenciales generadas por la polución. A través de esto, lograr una mayor productividad y eficiencia del proceso de transporte, sin comprometer los estándares de seguridad y el monitoreo del cumplimiento de los compromisos medioambientales.

A continuación, se describen las actividades del proceso de regadío:

- Recarga en cachimba, los operadores deben estacionarse adecuadamente, de forma horizontal y con el orificio de llenado en la posición de abastecimiento, el llenado dura entre 7 a 8 minutos, esto depende del flujo de agua que se encuentre en la cachimba. Además, hay que considerar el tiempo de maniobra del operador.
- Revisión de difusores, los operadores revisan el estado general del camión al momento de iniciar su turno para descartar y cerciorarse que los difusores estén óptimas condiciones para operar.
- Asegurar el equipo en el abastecimiento, el operador debe asegurar el correcto posicionamiento del camión en la zona de carga, para evitar que se produzcan inconvenientes como pérdida de agua por ubicar incorrectamente el equipo.
- Abastecer equipos perforadores, esta actividad se realiza al inicio del turno para favorecer el uso de los camiones en los horarios donde la polución es crítica, ya que los camiones aljibe son solicitados en toda la operación del rajo. Para realizar esta actividad el operador de la perforadora y del camión aljibe deben estar comunicados para realizar las maniobras de forma segura y correcta.
- Disminuir polución en la frente de carguío, para esto el operador del camión aljibe debe comunicarse con el operador de la pala para poder ingresar a esta zona a realizar el riego, así disminuye la polución en la frente.
- Riego caminos internos del rajo, generalmente en el rajo están operativos entre 4 a 5 camiones regadores los cuales deben encargarse de regar la mayor cantidad de circuitos, pero principalmente las zonas críticas como las rampas, frentes de carguío y lo que el supervisor de servicio y desarrollo considera de mayor importancia.
- Riego caminos exteriores del rajo, estos son mantenidos por el camión más grande de la flota de camiones aljibe, debido a que este equipo presenta desventajas para regar

los circuitos cortos del interior del pit. Cuando el chancador de Rosario se encuentra en mantención, se usa para regar el camino a Ujina.

- Botaderos, a estas zonas se les asigna un camión dependiendo de la cantidad de botaderos operativos y de la fluencia de camiones.
- Planta Enaex, esta actividad se realiza en conjunto con el personal de la empresa externa, ya que los encargados de abastecer sus equipos son los operadores ESED, la compañía solo facilita el agua.

2.6.1 Flota de equipos regadores

La compañía cuenta con 7 camiones regadores, 6 equipos modelo Komatsu HD 785-7 con una capacidad nominal del estanque de 75.700 litros con un rendimiento lineal de cuatro kilómetros. El equipo más grande es un camión de alto tonelaje modificado, modelo Komatsu 830 con una capacidad de almacenamiento de 196.000 litros. La compañía dispone de 28 operadores propios y 17 mantenedores externos.



Figura 4. Camión regador Komatsu HD 785-7. Fuente: CMDIC.

La disposición de los recursos se realiza según requerimiento, los equipos se utilizan para regar constantemente los distintos sectores productivos de la mina, como las fases que se encuentran en producción, caminos a botaderos, chancadores primarios, rampas de acceso, caminos interiores y exteriores al pit, y para abastecer de agua a las perforadoras. Para realizar el riego, existe una previa coordinación entre el jefe de turno y el supervisor de desarrollo y servicios, quienes coordinan la actividad de regadío dependiendo de la distribución de equipos al interior de la mina.

2.6.2 Supresores de polvo utilizados

En el mercado existe una serie de productos que evitan que el material particulado se levante desde el suelo modificando las propiedades físicas de la superficie. Los más utilizados en la industria son productos del tipo bituminosos (emulsiones asfálticas), polímeros sintéticos, sales y cloruros, y agua. De acuerdo con el análisis técnico económico realizado por Herrera (2018), es que se determinó que el mejor supresor desde el punto de vista técnico económico y considerando las pruebas realizadas, la emulsión asfáltica es la que tiene un mejor comportamiento frente a las condiciones particulares de la mina.

Actualmente en la compañía se utiliza como supresor agua y cloruro de magnesio, comercialmente conocido como Bischofita.

2.6.3 Criterios de operación de camión supresor

A continuación, se listan algunos estándares de control del procedimiento de operación del camión regador:

- Se debe dar preferencia al riego de los sectores con mayor afluencia de camiones de extracción como la frente de carguío, botaderos, chancadores primarios, rampas y caminos de transporte.
- Evitar mojar personas, cables eléctricos y maquinarias.
- El regadío en pasacables de piso y aéreos debe evitarse completamente, cortando la salida de agua antes y después de enfrentarlos.
- Las curvas cerradas no deberán ser regadas, como asimismo los últimos 40 metros antes de llegar a una señal de disco pare o ceda el paso y antes de los pasa cables de piso. Si el sector lo requiere, solicitar autorización al supervisor y dar aviso por la frecuencia radial. Este regadío no debe ser en exceso.
- El regadío está restringido por la temperatura, tanto en turnos diurnos como nocturnos. En turno día la temperatura límite para cambiar la manera de regar es de 1°C, mientras que en turno noche es de 4°C.
- En el marco del ciclo de gestión de riesgos, esta actividad debe ser evaluada y autorizada mediante VATS y GRT.

- Durante las estaciones del año, con bajas temperaturas (menos de 4°C), tormentas eléctricas, nevadas y lluvias, se debe resguardar al menos un camión aljibe que se encuentre en reserva, dentro de las instalaciones del Truck Shop, para asegurar disponibilidad en caso de emergencia.

2.6.4 Criterios de aplicación de supresor de polvo

- El operador del camión deberá regirse a las instrucciones de riego definidas en los puntos anteriores y según la programación definidas por el supervisor de Desarrollos y Servicios Mina.
- La solución deberá ser aplicada uniformemente sobre todo el ancho de la superficie. En caso de rampas se debe contar con la autorización del jefe de turno del área, realizándose en forma alternada a través de franjas longitudinales de 40mts según los criterios señalados en el punto anterior.
- Posterior a la aplicación se debe esperar un tiempo para el secado, sin tránsito de equipos por sector, definido por el supervisor.
- Durante bajas temperaturas (menos de 4°C), nevadas o lluvias, se debe suspender la aplicación del supresor de polvo, en todas las áreas.



Figura 5. Imagen referencial de camión regador. Fuente: CMDIC

2.6.5 Consumo de agua del proceso de regadío

El consumo promedio de agua en el período comprendido entre los años 2018 al 2020 es de 24,74 litros por segundo, lo que equivale a un consumo promedio de 57.174.839 litros por mes. Por lo que anualmente se consumen aproximadamente 695.000 metros cúbicos de agua en las actividades del proceso de regadío.

2.7 Compromisos medioambientales

De acuerdo con lo planteado por el proyecto de aumento de capacidad a 110ktpd para el año 2001, se establecen los siguientes acuerdos medioambientales con la entidad reguladora:

- i. Humectación de caminos y rutas de tierra de alto tráfico, principalmente en los sectores de acceso a la mina y botaderos de estéril.
- ii. En faena operan tres camiones aljibe destinados en forma permanente al riego de caminos. El riego solo se interrumpe en periodos de lluvia o cuando se presentan condiciones de congelamiento. En promedio se estima que cada tramo de camino es regado al menos dos veces al día.
- iii. Collahuasi, de acuerdo con sus políticas corporativas, mantendrá un registro interno de las inspecciones en terreno, que estarán disponibles para la revisión de la autoridad en cada una de las subáreas de la faena, orientadas a observar la periodicidad de riego y evaluar su efectividad.
- iv. Se aplica riego de caminos y áreas de circulación de camiones en el rajo y los botaderos por medio de camiones aljibe.
- v. Está prevista la estabilización de los caminos principales del rajo y de los botaderos, mediante el uso de cloruro de magnesio u otro producto similar.
- vi. El agua utilizada para riego proviene de los afloramientos del rajo.
- vii. La cantidad de agua utilizada en esta medida de abatimiento de polvo es de aproximadamente 200 a 300 m³/día.

Respecto al proyecto de optimización a 170ktpd presentado el año 2010, se presentan los siguientes acuerdos relacionados:

- i. Se establecen curvas de iso-concentración de MP10 en el rajo para una explotación de 170ktpd.
- ii. Se establece que las medidas de humectación deben tener un 80% de eficiencia en el abatimiento del polvo.

Conforme a lo establecido por el proyecto Pampa Pabellón del año 2014, sobre los relacionado a la descarga de material estéril en el muro, implementar un procedimiento de humectación con el objetivo de disminuir la emisión de material particulado en esta zona.

Es importante evidenciar que la compañía cumple en cantidad y calidad con los compromisos ambientales antes mencionados. Actualmente se tiene un cumplimiento parcial del primer compromiso, ya que no se cuenta con datos que den cuenta del abatimiento del polvo por concepto de regadío. Para asegurar el 80% de eficiencia en supresión de polvo, primero hay que ser capaces de medir la concentración de polvo que se emite en su totalidad. El compromiso medioambiental sobre el consumo de agua debe ser actualizado con la gerencia de permisos, ya que el consumo actual de agua por la actividad de regadío es mucho mayor.

Capítulo 3. Conceptualización

3.1 Antecedentes de la industria

La mina Mount Arthur Coal se ubica en Nueva Gales del Sur, Australia, su operación se encuentra cercana a grandes ciudades como Muswellbrook, con una población de 12.000 habitantes, esta se ubica tan solo dos kilómetros de la operación minera. El año 2019, la empresa comienza a aplicar un Sistema de Control de Polvo (DCS), el cual ha transformado el proceso de gestión del polvo. Para ello rastrean continuamente los niveles de polvo a través de 11 monitores de polvo y clima en tiempo real. Esto les permite gestionar de manera más eficiente las acciones para controlar el polvo al mismo tiempo que se evita que las emisiones sobrepasen los límites reglamentarios. Además, utilizan un sistema para gestionar la hora óptima para realizar las tronaduras, considerando la información aportada por el sistema de monitoreo. En el tiempo que lleva operando, ha habido una reducción significativa del 69% de las alertas de polvo. Esta iniciativa hizo que la empresa recibiera el premio *Industry Excellence Award* por tener un efecto demostrable en la mejora de la calidad del aire, dicho galardón es otorgado por la *Air Society of Australia and New Zealand 2019*, (BHP, septiembre 2019).

Otro antecedente son las actividades realizadas en el puerto Australiano Port Hedland. El principal método de control utilizado por los exportadores es garantizar que se agregue suficiente humedad al mineral en el sitio de la mina para mantener las emisiones de polvo al mínimo. En este puerto se mueven cerca de 500 millones de toneladas de material, de las cuales 268 millones corresponden a exportaciones de BHP, pese a la gran cantidad de material que circula por este punto durante los últimos 8 años se ha visto una disminución significativa en los eventos de polvo (BHP, mayo 2019). Desde la gerencia creen que esto se explica por la aplicación de estrategias continuas para reducir el polvo, estas van desde la eliminación de actividades que tienen potencial de propiciar niveles altos de polvo hasta aumentar el número de estaciones de lavado de las correas de transporte en todas las instalaciones. Una de las herramientas que más contribuye es el monitoreo en tiempo real de los niveles de polvo y los datos meteorológicos, incluida la dirección del viento y la temperatura. Esto permite determinar rápidamente la fuente de cualquier problema de polvo y así ajustar las operaciones portuarias según sea necesario. Además, desarrollaron una

plataforma web que proporciona el monitoreo en tiempo real a toda la comunidad (Port Hedland Industries Council, 2019).

Un ejemplo más cercano es lo que ocurre en la Mina Spence, su metodología tiene cuatro áreas de enfoque: desempeño de la calidad del aire, mejora de los controles de manejo de polvo, la protección de la fuerza laboral y la participación de las partes. Una de las medidas implementadas es la denominada “Cero polvo”, la cual tiene como objetivo reducir las emisiones de polvo de todas las fuentes, a través del monitoreo dinámico y modelos integrales que informan y ayudan a enfocar los esfuerzos. Además, implementan un modelo regional de emisiones de MP10, el cual predice las condiciones que podrían resultar en niveles dañinos para la salud de los trabajadores, prediciendo variaciones de emisiones de polvo con hasta nueve días de anticipación. Otra estrategia utilizada son las voladuras a gran escala, la cual tiene un impacto positivo en la exposición de la fuerza laboral a las emisiones de MP10 comparada con las tronaduras convencionales, (BHP, septiembre 2020).

3.2 Conceptualización proyecto de optimización

Para Compañía Minera Doña Inés de Collahuasi, la salud y seguridad de sus trabajadores, así como la responsabilidad y respeto por el medio ambiente, comunidades y entorno cercano, son principios valóricos en los que se fundamenta su proceder en el que hacer minero. La visión de la compañía se enmarca en la estrategia global definida por el quinquenio 2015-2019, cuyo foco es el desarrollo sustentable de la compañía. Uno de los principales ejes a reforzar es trabajar haciendo énfasis en la gestión del riesgo y en la excelencia operacional.

El polvo es un agente inherente a las distintas actividades mineras que se realizan en faenas a cielo abierto, según lo establecido por el Dust Control Handbook del 2019, la actividad de transporte y flujo vehicular puede aportar entre el 78% a 97% de las emisiones de polvo. Por lo que este se constituye como un agente de riesgo debido a que disminuye la visibilidad de los operadores de camiones de extracción y es un riesgo para la salud de los trabajadores debido a la fracción respirable de polvo. Además, representa un riesgo ambiental por el alto consumo de recurso hídrico usado para abatir el polvo, y por afectar la calidad del aire por la emisión de material particulado a la atmosfera. Por todo lo anterior, la compañía se enfoca en la gestión y control del material particulado que se genera a partir de las operaciones

mineras, en este caso particular, se busca ser más eficientes en el control y mitigación del polvo producido por el tránsito de camiones de alto tonelaje en caminos no pavimentados.



Figura 6. Imagen referencial de CAEX en tránsito. Fuente: CMDIC

Por lo anterior, se implementa un plan piloto con una empresa de servicios externos, cuyo objetivo es optimizar el proceso de regadío a través del monitoreo en tiempo real de variables como la frecuencia del regadío de las distintas áreas, la eficiencia de la humectación de los sectores productivos, y la concentración de material particulado en conjunto con variables ambientales. Esto permitirá gestionar el proceso con información levantada desde la misma operación, permitiendo asignar la flota de camiones regadores en forma dinámica y según requerimiento y necesidad.

De esta forma se podrán establecer controles claros para dar cumplimiento a los compromisos medioambientales acordados con la autoridad evaluadora.

El servicio externo provee un sistema de medición de calidad de aire, mediante una red fija y una red móvil. Las estaciones fijas se ubican en lugares estratégicos al interior del rajo, como en las distintas fases activas, rutas de acceso al rajo y botaderos, sectores de acopio de material, chancadores, oficinas y casinos. Mediante sensores móviles instalados en camiones de extracción y/o camiones regadores, se obtienen mediciones desde la misma fuente de polvo y desde los distintos lugares por donde transitan los camiones, a la vez, permitirá dar seguimiento a la flota de regadores. Cabe destacar que la información levantada por las estaciones fijas y los sensores móviles corresponden a datos georreferenciados.

Para dar seguimiento en tiempo real de las variables de interés, ambas redes de monitoreo utilizan un sistema 3G para transmitir los datos desde las estaciones a la base de datos del servidor de la compañía.

A partir de la información recolectada por el sistema de monitoreo dinámico, se establece una condición base de material particulado más precisa para las distintas áreas productivas, con esto se construyen curvas de iso-concentración de MP10, lo que permite modelar el comportamiento del polvo en todas las áreas de interés.

Considerando la gran cantidad de datos que se obtendrá desde la operación, mediante algoritmos de aprendizaje de máquina sería factible predecir la condición futura de material particulado. Dicha información serviría para anticiparse a las condiciones desfavorables y optimizar las gestiones de las herramientas de control de abatimiento del polvo. En términos generales los algoritmos pueden procesar de forma rápida y eficaz una gran cantidad de datos, interpretando las relaciones de causalidad entre una y otra variable. Para ello es necesario entrenar los modelos de aprendizaje en base a información histórica disponible, de esta forma se procesan los datos mediante los pasos lógicos de cada algoritmo correlacionando las variables y “aprendiendo” de los datos, para finalmente entregar una interpretación de estos. El output del algoritmo es una alerta preventiva al proceso, que indica una categoría que representa el rango de nivel de polvo que se espera tener en el futuro en un determinado sector de la faena.

El valor agregado del aprendizaje de máquina es que aprende de los eventos pasados y se nutre de los datos que van ocurriendo minuto a minuto, modelando de forma integral el comportamiento de todas las variables críticas.

En términos generales, se logra un sistema de monitoreo de alta resolución espacial en tiempo real, dando una respuesta eficaz y costo-efectiva, que aumenta el conocimiento espaciotemporal con tecnologías de bajo costo. Permitiendo un mayor entendimiento de la dinámica del polvo y una rápida toma de decisiones, teniendo la capacidad de medir la generación de polvo desde su origen. Con base en la gran cantidad de datos levantados desde la operación en tiempo real y según los modelos de aprendizaje de máquina, predecir la condición desfavorable para generar alertas al proceso, que permitan anticiparse a las condiciones desfavorables para controlar el polvo antes de que este se levante del suelo.

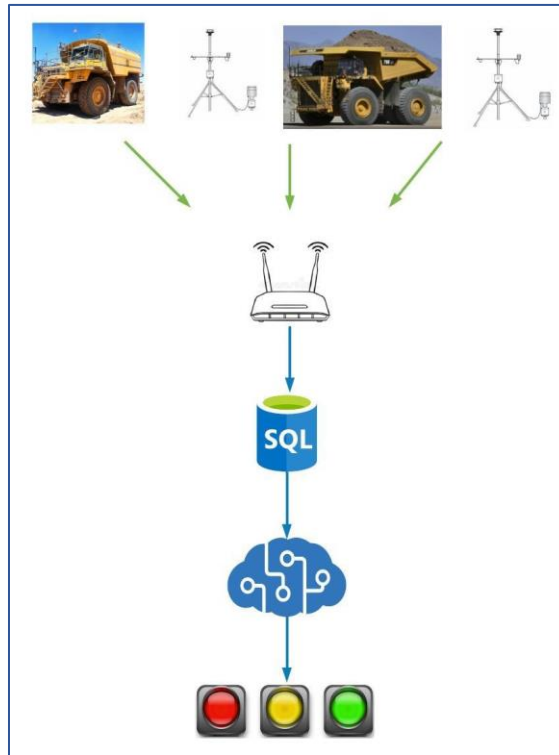


Figura 7. Esquema flujo de información. Fuente: elaboración propia.

3.2 Modelo de trabajo

3.2.1 Descripción modelo de trabajo

Se presenta un esquema gráfico del modelo de trabajo:



Figura 8. Diagrama modelo de trabajo. Fuente: elaboración propia.

Análisis de fuentes de emisión → Conocer el comportamiento del material particulado en las áreas críticas.

Variables significativas → Aquellas que tienen un impacto en el comportamiento del polvo.

Monitoreo de variables → Macrodatos de la operación.

Aprendizaje de máquina → Estadística inferencial de macrodatos.

Análisis de datos de monitoreo → Rangos de favorabilidad del análisis inferencial de las variables críticas.

Alertas de control al proceso → Comunica el estado en que se encuentran las áreas productivas.

3.2.2 Variables a medir

Con los recursos que actualmente se disponen es posible tener registro de las siguientes variables:

1. Temperatura.
2. Humedad relativa.
3. Velocidad del viento.
4. Dirección del viento.
5. Frecuencia de humectación.
6. Tasa de humectación.
7. Concentración de CAEX por fase.
8. Concentración de material particulado (MP10).

3.2.3 Tipos de redes a implementar

- Red fija: Estaciones de monitoreo de calidad de aire ubicadas en lugares estratégicos de la mina: fases activas, rampas de alto tráfico, sectores de casinos, caminos exteriores al rajo. Estas permitirán asegurar el cumplimiento de las curvas de isoconcentración de material particulado MP10.
- Red móvil: Estaciones de monitoreo de calidad de aire ubicadas en camiones de extracción y/o camiones regadores, lo que permite realizar el monitoreo dinámico en todos los sectores por donde circulen los camiones. Permitiendo asegurar el monitoreo y cumplimiento de la eficiencia comprometida y de los compromisos medioambientales pactados.

3.2.4 Variables objetivo del plan piloto

1. Frecuencia de humectación.
2. Tasa de humectación por sector en [ml/m²].
3. Eficiencia de humectación [%].
4. Abatimiento de material particulado MP10 [%].
5. Tiempos de respuesta a condiciones críticas de polvo en suspensión.

Capítulo 4. Marco Teórico

4.1 Estudios previos

Alrededor del mundo se han realizado una serie de estudios que dan cuenta de las aplicaciones de herramientas computacionales, que utilizan algoritmos de Machine Learning para modelar y predecir la calidad del aire. Con el rápido crecimiento de las grandes urbes y los distintos tipos de industrias que cada vez son más diversas y ansiosas por suplir las necesidades del ser humano, es que la contaminación del aire se ha convertido en una gran problemática en los países más industrializados, grandes urbes y en donde sea que realicen actividades potencialmente contaminantes. Los principales agentes nocivos que se encuentran en la atmósfera son el óxido de carbono, compuestos nitrogenados y azufrados, hidrocarburos, metales y material particulado (Ministerio del Medio Ambiente, 2016). Siendo el material particulado uno de los agentes más peligrosos para la salud del ser humano, debido a que es capaz de penetrar en las estructuras más pequeños del sistema respiratorio.

Frente a esta problemática, es primordial conocer la fuente y cantidad de estos contaminantes para gestionar acciones que permitan minimizar los efectos adversos de la polución atmosférica. Una de las alternativas para controlar la contaminación del aire y resguardar a las personas de los impactos a la salud, es implementar un sistema de monitoreo en tiempo real de la calidad del aire. Además, la predicción de calidad del aire se vuelve necesaria para contar con un mayor análisis de la información, y así gestionar decisiones óptimas para el control medio ambiental, (Athira et. al, 2018). En este marco de referencia, los autores aplican tres metodologías de aprendizaje profundo para predecir la condición de material particulado: Recurrent Neural Network (RNN), Long-Short Time Memory (LSTM), y Gated Recurrent Unit (GRU). Los datos utilizados consisten en 6 índices de calidad del aire de 1498 estaciones de monitoreo del Centro Nacional de Monitoreo Ambiental de China, por sus siglas en inglés CNEMC y datos meteorológicos del Sistema Global de Predicción (GFS). Los algoritmos son capaces de establecer relaciones complejas entre datos multivariantes, lo que permite predecir con cierto nivel de confianza lo que podría ocurrir en el futuro con alguna de las variables. Para determinar cuál metodología es mejor se ejecuta una gran cantidad de experimentos para cada modelo, variando hasta 1000 épocas de entrenamiento

variando la tasa de aprendizaje entre [0.01-0.5]. Se observa a partir del estudio que los tres modelos se desempeñaron comparativamente bien en la predicción, donde los resultados obtenidos con la red GRU fueron ligeramente superiores que los obtenidos con RNN y LSTM.

Otra metodología usada para predecir el material particulado es la máquina de vectores de soporte, del inglés Support Vector Machine (SVM). Motivados por el aumento en los niveles de polución del aire en la mayoría de las ciudades de China, Los autores Weizhen et. al. 2014, utilizan la técnica de SVM por su capacidad para modelar el comportamiento de datos no lineales multivariantes. Describen el desarrollo de un modelo de regresión de vector de soporte de sobre relajación sucesiva (SOR-SVR) para la predicción de PM10 y PM2.5, basado en la profundidad óptica de aerosol (AOD) promedio diario y parámetros meteorológicos que se midieron en Beijín durante el año 2011-2012. El modelo predictivo utiliza un kernel Gaussiano y la validación cruzada con K-folds, y búsqueda de cuadrícula, determinan los parámetros que mejoran la capacidad de generalización del modelo. El estudio demuestra que la metodología tiene un alto poder predictivo, el resultado muestra que los valores predichos por el modelo SOR-SVR concuerdan bien con los datos reales y tienen una buena capacidad de generalización para predecir PM10 y PM2.5. Además, la información aportada por el AOD juega un papel importante en la predicción de partículas con el modelo SVR, que deben incluirse en el modelo de predicción. Ya que si solo se consideran los parámetros meteorológicos los resultados no son satisfactorios.

El estudio realizado por Biancofiore et. al, (2017), se muestra un análisis que utiliza como entradas variables meteorológicas y la concentración de MP10, para predecir la condición promedio diaria de material particulado, anticipando de uno a tres días en el futuro. Para ello comparan el desempeño de tres modelos: un modelo de regresión lineal múltiple y redes neuronales con y sin arquitectura recursiva. Utilizan los datos recopilados por una estación de monitoreo ubicada en la costa del mar Adriático, al sur de Pescara, Italia. Las simulaciones realizadas muestran que la red neuronal con arquitectura recursiva tiene un mejor comportamiento comparado con los otros modelos. Los resultados de los pronósticos se comparan con los límites de la calidad del aire para la protección de la salud para probar su desempeño como herramienta operativa. Además, realizan otras simulaciones incluyendo

como datos de entrada la concentración de monóxido de carbono (CO), el cual tiene un impacto positivo en el comportamiento del modelo y en la precisión del pronóstico. La información adicional capacita a los modelos para poder interpolar la predicción de MP2.5, incluso cuando este no se incluye como variable de entrada.

En la universidad de Birmingham, Suleiman et. al, (2018), presentan una nueva metodología para evaluar la efectividad de los modelos de aprendizaje de máquina, con el objetivo de ser usados para reducir la concentración de MP10 y MP2.5 en caminos no pavimentados, para ello aplican modelos de redes neuronales artificiales, árbol de regresión impulsado, del inglés Boosted Regression Trees (BRT) y máquina de vectores de soporte (SVM). Para el análisis se considera información de tráfico de vehículos por los caminos monitoreados, variables meteorológicas y los niveles de polución, la información proviene de diecinueve estaciones de monitoreo del Air Quality Monitoring (AQM), en la ciudad de Londres. A partir de las métricas utilizadas concluyen que el ANN y BRT tienen un mejor comportamiento para predecir el MP10 que la máquina de vectores de soporte. Sobre el pronóstico de MP2.5, no es posible anteponer un modelo sobre otro, ya que las métricas utilizadas muestran un desempeño similar entre las metodologías. El estudio señala que al considerar variables como la concentración de óxido nitroso la predicción mejora debido a la correlación que existe entre estas variables.

En el estudio realizado por Zhao et. al. (2019), consideran que la predicción precisa de la calidad del aire tiene una gran importancia práctica para ayudar a controlar los niveles de polución en muchas ciudades de China. Proponen un modelo basado en datos, denominado red neuronal de memoria corto plazo y completamente conectada, por sus siglas en inglés LSTM-FC. El modelo predictivo consta de dos componentes, el uso de un simulador temporal basado en LSTM y el uso de una combinación espacial basada en una red neuronal para capturar las dependencias espaciales entre la contaminación por material particulado de los datos de la estación central y las estaciones vecinas. Evaluando el modelo en un conjunto de datos con información proveniente de 36 estaciones de monitoreo de la calidad del aire de la ciudad de Beijing, y comparando con el desempeño de una red neuronal artificial (ANN) y un modelo clásico de memoria corto-largo plazo en el mismo conjunto de datos, se muestra que el modelo LSTM-FC tiene un mejor rendimiento predictivo.

En el estudio realizado por Xayasouk et. al. (2020), se desarrollan modelos para predecir la concentración de material particulado usando memoria corto largo plazo (LSTM) y métodos de codificación automática profunda (DAE), comparando los resultados en términos de la raíz cuadrática del error medio (RMSE). Se aplican los modelos en datos obtenidos de 25 estaciones de Seoul, Corea del Sur, entre el 1 de enero 2015 al 31 de diciembre de 2018. La predicción se realiza para los diez días siguientes a la fecha del intervalo, configurando el modelo con una ratio de aprendizaje óptimo de 0.01 para 100 épocas de entrenamiento y un tamaño de lote de 32 para LSTM y de 64 para los modelos DAE. Los modelos propuestos predicen efectivamente la concentración de material particulado fino, donde el modelo LSTM presenta un mejor comportamiento conforme a las métricas usadas. Según los autores, con este modelo predictivo sería posible aportar información confiable para mejorar la toma de decisiones del usuario.

Para los autores Nath et. al. (2021), los avances en materia de predicción de contaminación del aire se han convertido en una cuestión clave para mejorar la toma de decisiones que ayuden a mitigar los efectos a largo plazo de la polución. Hasta el momento solo se han propuesto métodos estadísticos de aprendizaje profundo para predecir el comportamiento de los contaminantes atmosféricos, pero rara vez se han utilizado para pronosticar la tendencia de la contaminación futura a largo plazo. En este estudio se presenta la comparación entre los modelos estadísticos: Holt-Winters, AR, SARIMA, Prophet. Y distintas variaciones de la red neuronal recurrente de memoria corto-largo plazo LSTM, con el objetivo de predecir la condición a largo plazo del MP10 y MP2.5. Para el estudio se utiliza la información histórica recolectada de las estaciones de monitoreo de la ciudad de Calcuta, India. Las simulaciones realizadas muestran que, dentro de los métodos estadísticos, Holt-Winter presenta en general los mejores resultados en termino de las métricas RMSE y MAE. En el caso del aprendizaje profundo, la red neuronal LSTM convolucional y apilado, dan los mejores resultados de predicción de MP10 y MP2.5 con respecto a los parámetros de comparación utilizados. Para los autores el aporte del estudio es muy relevante y se propone como una herramienta que ayuda a las organizaciones responsables, de formular políticas que permitan implementar nuevas medidas y regulaciones para bajar los niveles de contaminación en sus ciudades y hacer más saludable el medio donde se desarrolla la población.

Los antecedentes planteados dan cuenta que los algoritmos de LSTM y SVM son metodologías que tienen buen comportamiento cuando se busca predecir la polución del aire, en específico del material particulado.

Tabla 5. Tabla comparativa revisión bibliográfica.

Autor	Metodología ML	Resultados
Athira (2018)	Recurrent Neural Network	Los mejores resultados de la predicción se obtienen con Gated Recurrent Unit
	Long short Time Memory	
	Gated Recurrent Unit	
Weizhen (2014)	Describen una metodología de SVM, modelo de regresión de vector de soporte SPR-SVR.	Demuestran que su metodología tiene un alto poder predictivo para la polución del aire.
Biancofore (2017)	Modelo de regresión lineal múltiple	Las tres metodologías son capaces de realizar buenas predicciones, destacando las redes neuronales recursivas.
	Redes neuronales	
	Redes neuronales recursivas	
Suleiman (2018)	Redes neuronales	Para la predicción de MP10 tienen un mejor comportamiento la red neuronal y BRT, y para MP2.5 no se establece ninguna conclusión ya que los 3 métodos tienen un comportamiento similar.
	Boosted Regression Trees (BRT)	
	Support Vector Machine	
Zhao (2019)	Proponen una red neuronal de memoria de corto plazo y completamente conectada, del inglés LSTM-FC.	Al comparar con un ANN y un LSTM clásico el LSTM-FC tiene un mejor poder de predicción.
Xayasouk (2020)	Long short Time Memory (LSTM)	Ambos modelos predicen efectivamente la condición de polvo fino, pero LSTM tiene un mejor comportamiento en función de las métricas usadas.
	Codificación automática profunda	
Nath (2021)	Holt-Winters	Las simulaciones realizadas muestran que los mejores resultados se obtienen con Holt-Winters.
	SARIMA	
	Prophet	
	Long Short Time Memory	

4.2 Series de tiempo

Una serie de tiempo es una lista de fechas que puede contener o no información sobre la hora, donde a cada paso de tiempo se le asocia una observación sobre los valores que toma una o múltiples variables en diferentes momentos del tiempo. Las series de tiempo son un modo

estructurado de representar datos, visualmente es una curva que evoluciona a través del tiempo, (Ríos G, 2008).

Los datos se pueden comportar de diferentes formas a través del tiempo, presentando algún tipo de tendencia, un ciclo, presentar una forma aleatoria, variaciones estacionales. Los modelos de series de tiempo tienen un enfoque netamente predictivo y en ellos los pronósticos se elaboran solo con base en el comportamiento pasado de la variable de interés.

Se distinguen dos tipos de estructuras de series de tiempo, **modelos deterministas**, son métodos de extrapolación sencillos en los que no se hace referencia a las fuentes o naturaleza de la aleatoriedad subyacente de la serie, esta simplicidad va acompañada de una menor precisión. Un ejemplo de estos modelos son los de media móvil, donde se calcula el pronóstico de la variable a partir de un promedio de los “n” elementos anteriores. Por otro lado, se tienen los **modelos estocásticos**, en términos sencillos se asume que la serie se extrae de un grupo de variables aleatorias con cierta distribución conjunta difícil de determinar, por lo que se construyen modelos aproximados que sean útiles para realizar algún pronóstico.

4.3 Tipos de aprendizaje automático

Los modelos de aprendizaje de máquina se diferencian en la forma en que el sistema aprende de los datos, distinguiendo los siguientes métodos:

- Aprendizaje supervisado

El aprendizaje supervisado es una técnica para deducir una función a partir de datos de entrenamiento. Esta es controlada por un agente externo que determina la respuesta que debería generar la red a partir de entradas etiquetadas, en otras palabras, le decimos al modelo que es lo que queremos que aprenda. En caso de que estas no coincidan con la salida esperada, se deben modificar los pesos de las conexiones para que la salida se aproxime a la salida objetivo. A su vez, dependiendo del tipo de etiqueta, se distinguen dos tipos de modelos:

- Modelos de clasificación, producen como salida una etiqueta discreta, es decir, una etiqueta dentro de un conjunto finito de etiquetas posibles. Clasificación binaria y multiclase.
- Modelos de regresión, producen como salida un valor real.

- Aprendizaje no supervisado:

El aprendizaje no supervisado no requiere la influencia externa para ajustar los pesos de las conexiones entre sus neuronas, o sea trabaja con datos no etiquetados. Se usan principalmente en tareas donde es necesario analizar los datos para extraer nuevo conocimiento o agrupar entidades por afinidad. Aplicaciones para reducir la dimensionalidad o simplificar conjuntos de datos. En el caso de agrupar por afinidad, el algoritmo define una métrica de similitud o distancia que le sirva para comparar los datos entre sí. Para reducir dimensionalidad los algoritmos utilizan técnicas matemáticas y estadísticas para disminuir la dimensión del conjunto de datos a cambio de perder un poco de información.

- Aprendizaje de refuerzo:

El aprendizaje por refuerzo se centra en los procesos de aprendizajes reglamentados, en los que se proporcionan algoritmos de aprendizaje automáticos con un conjunto de acciones, parámetros y valores finales. El sistema enseña a la máquina a través de procesos de ensayo y error. Se formula generalmente como un proceso de decisión de Markov.

4.4 Redes neuronales

Las redes neuronales artificiales, por sus siglas en inglés ANN son modelos de regresión no lineales que en términos simples se asemejan al funcionamiento del sistema nervioso, donde las unidades básicas son las neuronas que se organizan en capas. Es un modelo simplificado que emula el modo en que el cerebro humano procesa la información, esto quiere decir que simultáneamente procesa múltiples unidades de procesamiento interconectadas que representan una versión rudimentaria de las neuronales, donde cada una de ellas es responsable de un pequeño calculo simple (IBM Corporation, 2021).

En las redes se distinguen tres partes, una capa inicial con campos que representan las variables de entradas (inputs). Las capas ocultas, las que no tienen conexión directa con el entorno, proporcionan grados de libertad a la red neuronal gracias a los cuales es capaz de representar más fehacientemente determinadas características de los datos que intenta modelar. La capa de salida se compone de neuronas que proporcionan la respuesta de la red neuronal (output). En términos generales, una red consiste en un gran número de unidades

simples de proceso, que actúan en paralelo, están agrupadas en capas y están conectadas mediante vínculos ponderados. Cada neurona recibe inputs desde otras neuronas y genera un resultado que depende sólo de la información localmente disponible, esta servirá de input para otras neuronas. De acuerdo con lo anterior, véase la figura N°9:

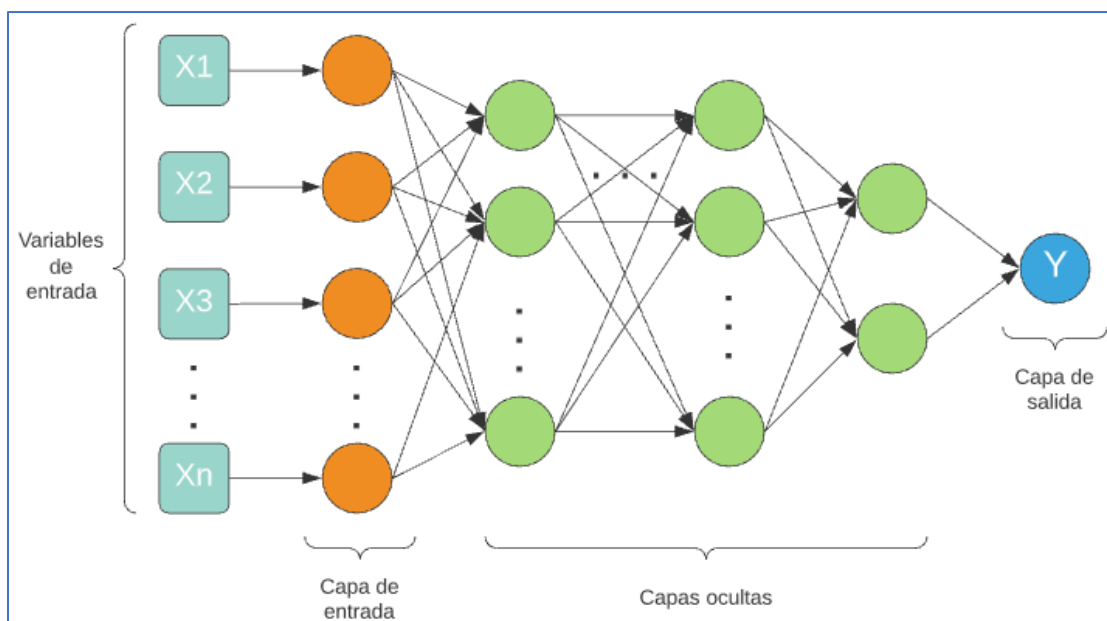


Figura 9. Estructura generalizada de una red neuronal. Fuente: elaboración propia.

Cada círculo representa una neurona que está conectada a las subsiguientes neuronas mediante enlaces, representadas por funciones matemáticas de activación y salida. Estos enlaces tienen un peso atribuido que multiplica el valor de salida de la neurona, otorgando más o menos valor a la activación de las siguientes neuronas. En general, la red va jugando con estos pesos ponderados de manera automática y sin ser manualmente controlada por un ser humano, de esta manera la máquina obtiene la mejor combinación posible, modificando los pesos de las neuronas para minimizar la función de pérdida. Sin embargo, la principal desventaja de estas redes es que sufren de “amnesia”, ya que para generar una salida solo consideran la entrada actual y no entradas anteriores.

4.5 Redes neuronales recurrentes

Las redes neuronales recurrentes (RNN) permiten conexiones de las neuronas consigo mismas y con neuronas de capas inferiores, dotando a las mismas de una dinámica de comportamiento que no es posible ser replicada con las redes ANN. Las redes recurrentes

son una familia de las redes neuronales artificiales que se especializa en el procesamiento de datos secuenciales, por esto son útiles en problemas donde existe dependencia temporal. En las redes neuronales artificiales las funciones de activación sólo actúan en una dirección, hacia adelante, desde la capa de entrada hasta la capa de salida, en otras palabras, estas no recuerdan valores previos, Valenzuela (2018). En una red RNN se incluyen conexiones que apuntan hacia atrás, o sea que hay una señal que va y vuelve, produciendo una retroalimentación entre las neuronas dentro de la capa. Una RNN se puede ver como un sistema dinámico de la forma:

$$y^t, s^t = g(x^t, s^{(t-1)}) \quad (1)$$

Donde $x^t = (x_1^t, x_2^t, \dots, x_n^t)^T$ es el vector de entrada de la secuencia en el instante de tiempo “t”, y el término $y^t = (y_1^t, y_2^t, \dots, y_o^t)^T$ corresponde al vector de salida en el instante “t”, por otra parte, el término s^t es el estado interno del modelo y $g()$ es la función de transición, la cual realiza un mapeo desde el instante anterior al actual. Este sistema dinámico es representado por la ilustración 7:

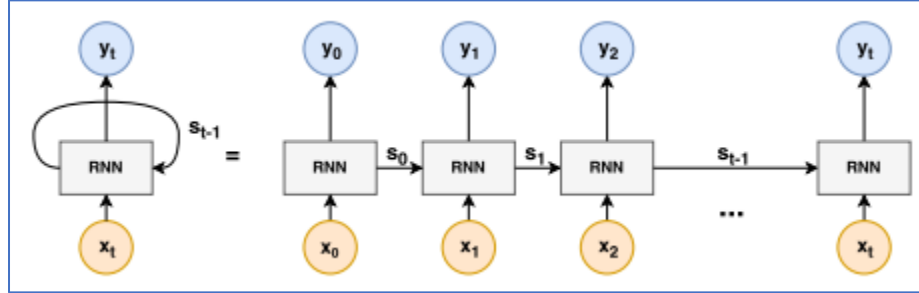


Figura 10. Representación esquemática de red neuronal recurrente.

Los nodos de la capa de entrada representan a los valores del vector de entrada x^t . A su vez, los nodos de las otras capas toman como valor de entrada los valores de las capas provenientes de la capa anterior y entregan un valor de salida que sigue la forma de la siguiente combinación lineal:

$$a_{q,d} = f \left(w_0 + \sum_{j=1}^m w_{qj} a_{j,d-1} \right) \quad (2)$$

En esta expresión $a_{q,d}$ es la salida del nodo q en la capa d , w_0 representa el sesgo, w_{qj} es el peso entre el j –ésimo nodo de la capa anterior y el nodo q de la capa d , y $f(\cdot)$ es una función no lineal denominada como función de activación.

La función de activación otorga la capacidad de encontrar relaciones no lineales entre las variables de entrada. Las funciones de activación más utilizadas son:

- Tangente hiperbólica

$$\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \quad (3)$$

- Función sigmoideal

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (4)$$

- Rectified unified unit

$$ReLU(z) = \max(0, z) \quad (5)$$

4.6 Long-Short Time Memory (LSTM)

Memoria de corto-largo plazo del inglés Long Short Time Memory, es un modelo propuesto por (Hochreiter, 1997). Estas son un tipo de red neuronal recurrente capaz de aprender la dependencia del orden en los problemas de predicción de secuencias. Las RNN no son capaces de aprender en presencia de retrasos de tiempo superiores a 5-10 pasos de tiempo discretos entre los eventos de entrada y las señales objetivo. Por su parte LSTM puede aprender a superar retrasos de tiempo mínimos de 1000 pasos de tiempo discretos mediante la aplicación de un flujo de error constante a través de “carruseles de error constante” (CEC) dentro de unidades especiales llamadas celdas, según lo establecido por Gers, A. et. al (2000).

El funcionamiento de las redes LSTM es relativamente similar a las RNN, usan una capa de entrada, una capa oculta y una capa de salida, sin embargo, la capa oculta se encuentra

completamente auto conectada e interconectada con las otras celdas, a su vez, esta contiene en su interior celdas de memoria y unidades de puerta de entrada y salida. La arquitectura interna de cada celda de memoria garantiza un flujo de errores constante dentro de su carrusel de errores constante CEC, esto representa la base para salvar información de retrasos de tiempo muy largos. Dos unidades de puerta aprenden a abrir y cerrar el acceso al flujo de error dentro del CEC de cada celda de memoria. La compuerta de entrada multiplicativa brinda protección a la CEC contra perturbaciones por entradas que no aportan mucha información, así mismo, la puerta de salida multiplicativa protege a otras unidades de la perturbación por contenidos de memoria actualmente irrelevantes.

El esquema de funcionamiento de la red LSTM se muestra en la ilustración N°9, donde σ es una función sigmoidea, como muestra la ecuación (4). La red contiene una puerta de entrada, una puerta de salida y una de olvido, donde la operación interactiva entre estas tres puertas hace que LSTM tenga la capacidad suficiente para resolver el problema de las dependencias a largo plazo que las RNN generales no pueden resolver, Huang et. al, (2018). Otro problema común de las redes neuronales profundas se llama desaparición del gradiente, es decir, la velocidad de aprendizaje de las capas ocultas anteriores es más lenta que en las capas ocultas más profundas. Este fenómeno puede incluso conducir a una disminución de la tasa de precisión a medida que aumentan las capas ocultas. Sin embargo, la estructura de la celda de memoria puede resolver eficazmente el problema de desaparición del gradiente en la propagación hacia atrás y puede aprender secuencias de entrada con pasos de tiempo muy largos. Por lo tanto, LSTM se usa comúnmente para resolver aplicaciones relacionadas con problemas de series de tiempo. La derivación de la fórmula específica de LSTM se ilustra en las siguientes ecuaciones:

$$\bar{Z}^t = W_z X^t + R_z y^{t-1} + B_z \quad (6)$$

$$Z^t = \tanh(\bar{Z}^{(t-)}) \quad (7)$$

$$\bar{I}^t = W_I \lambda^t + R_I y^{t-1} + p_i \odot C^{t-1} + B_I \quad (8)$$

$$I^t = \text{sigmoidea}(\bar{I}^t) \quad (9)$$

$$\bar{F}^t = W_F \lambda^t + R_F y^{(t-1)} + p_f \odot C^{t-1} + B_F \quad (10)$$

$$F^t = \text{sigmoidea}(\bar{F}^t) \quad (11)$$

$$C^t = Z^t \odot y0^t + C^{t-1} \odot f^t \quad (12)$$

$$\bar{O}^t = W_0 \lambda^t + R_0 y^{(t-1)} + p_0 \odot C^{t-1} + B_0 \quad (13)$$

$$O^t = \text{sigmoidea}(\bar{O}^t) \quad (14)$$

$$y^t = \tanh(C^t) \odot O^t \quad (15)$$

Donde W_z, W_i, W_f y W_0 son pesos de entrada; R_z, R_i, R_f y R_0 son pesos recurrentes, p_i, p_f y p_0 son pesos de mirilla; B_z, B_i, B_f y B_0 son ponderadores de sesgo; Z^t es la puerta de entrada al bloque; f^t corresponde a la puerta de olvido; C^t es la celda; O^t es la puerta de salida; y^t es la salida del bloque; y $\odot$ representa el producto punto.

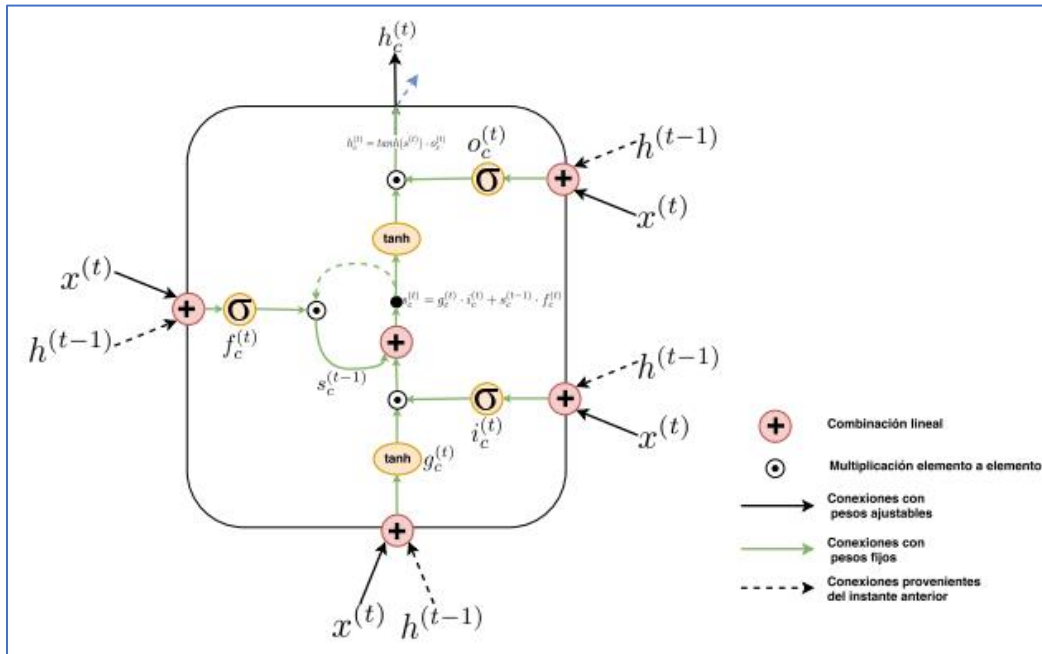


Figura 11. Celda de memoria del modelo LSTM

4.7 Support Vector Machine (SVM)

La máquina de vectores de soporte, del inglés support vector machine (SVM), constituye un método basado en aprendizaje supervisado para la resolución de problemas de clasificación, regresión y detección de valores atípicos. El algoritmo se basa en la teoría planteada por Vapnik, (1995). En términos simples, el objetivo del algoritmo es encontrar un hiperplano en un espacio N-dimensional, que clasifique claramente el conjunto de datos. En este sentido, dado un conjunto de datos de entrenamiento etiquetados, el algoritmo genera un hiperplano óptimo que clasifica los nuevos ejemplos en dos espacios dimensionales.

Partiendo de un conjunto de puntos en un plano que pertenecen a clases distintas, existe un numero infinito de rectas que podrían satisfacer el problema. El objetivo es encontrar la recta que tenga el margen máximo, vale decir, la mayor distancia media de los conjuntos de puntos con la recta. Maximizar esta distancia proporciona mayor confianza al momento de clasificar nuevos valores.

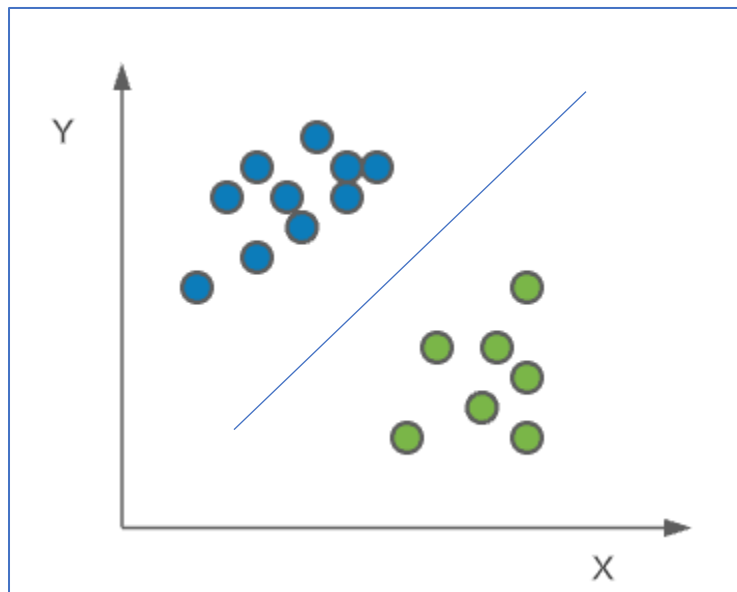


Figura 12. Escenario sencillo de un conjunto de datos. Fuente: elaboración propia.

En la mayoría de los casos, las variables a clasificar no van a ser linealmente separables. Si, por ejemplo, nos encontramos en un espacio unidimensional, para los siguientes puntos no existe un único hiperplano (en este caso un punto) que separe los puntos azules de los verdes:

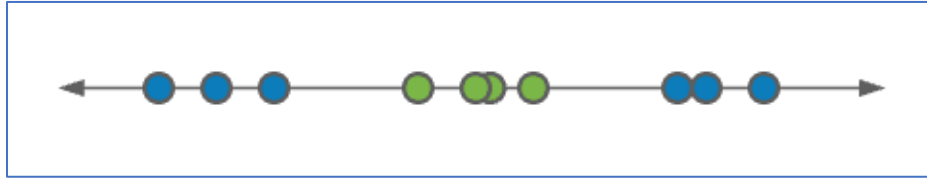


Figura 13. Puntos en espacio unidimensional. Fuente: elaboración propia.

Pero ¿qué sucedería si aplicamos una función a estos datos para llevarlos a una dimensión superior?, los puntos pasarían a un espacio bidimensional, de forma que la coordenada X sería la misma pero la coordenada Y sería igual al valor de la función evaluada en X, de manera que en este nuevo espacio sería posible encontrar una línea que separa los conjuntos de datos.

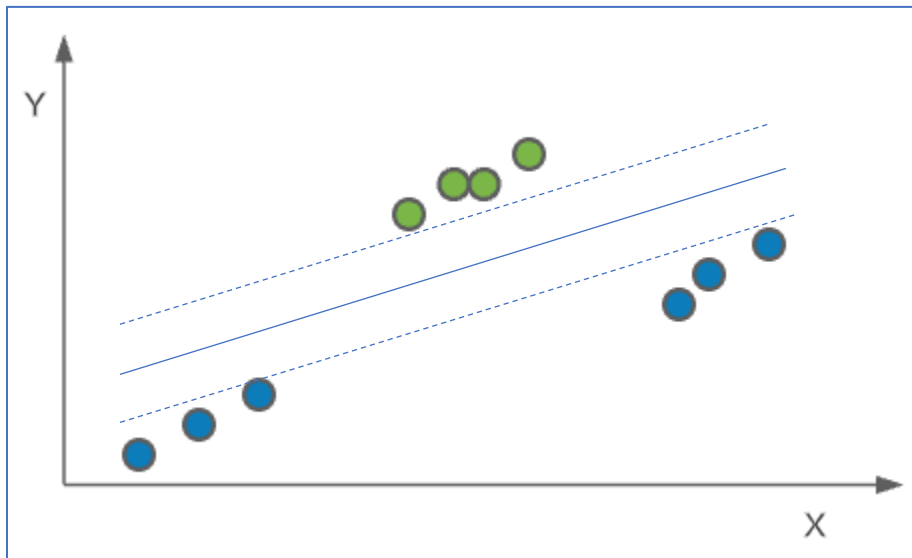


Figura 14. Hiperplano en dos dimensiones. Fuente: elaboración propia.

En el nuevo espacio, como se muestra en la imagen anterior, sí es posible crear una recta que separe ambos grupos de puntos. El algoritmo de SVM busca la recta que maximiza la distancia de los grupos de puntos con la recta. La posición del hiperplano de máximo margen, distancia a la línea continua o ϵ , queda determinada por los puntos más cercanos al hiperplano o los vectores de soporte.

Dependiendo del número de entradas P , se tiene un espacio dimensional de P dimensiones, por lo que el hiperplano se define como un subespacio plano y afín de dimensiones $P-1$. Por lo que en un espacio unidimensional el hiperplano es un punto, en un espacio de dos

dimensiones el hiperplano es una línea, y en un espacio de tres dimensiones el hiperplano es un plano. Pero ¿qué sucede si el conjunto de datos tiene más de 3 variables?, intuitivamente no es posible representar un espacio de más de 3 dimensiones, sin embargo, las matemáticas que soportan el método nos permiten trabajar en espacios n-dimensionales.

En otras palabras, el algoritmo traslada el conjunto de datos a una dimensión superior, en este nuevo espacio el método busca un vector de soporte de clasificación que separa la información en este nuevo ambiente. Para ello se aplica una función Kernel, la cual lleva a los datos a un espacio superior n-dimensional, para sistemáticamente encontrar el mejor vector de soporte de clasificación. Los vectores de soporte son puntos de datos que se encuentran cercanos al hiperplano e influyen en la posición y orientación de este. La función general de regresión de SVM viene dada por:

$$f(x) = W_i * \phi_i(x) + b \quad (16)$$

Donde W_i corresponde a un factor ponderador del vector, b corresponde al sesgo y ϕ representa la función de transformación no lineal, el espacio n-dimensional depende del número de variables de entrada.

Los vectores correspondientes al conjunto de entrenamiento son mapeados a un espacio de mayor dimensionalidad por una cierta función Kernel (ϕ). Este último viene determinado por una matriz cuadrada cuya dimensión es igual al número de variables que se analizan, el contenido de la matriz es el resultado de aplicar el kernel en los datos de entrenamiento. De acuerdo con el entorno de ejecución utilizado, las funciones núcleo más utilizadas son:

- Kernel Lineal

$$K(x_i, x_j) = x_i^T \odot x_j \quad (17)$$

En la ecuación, x_i y x_j representan los valores de la matriz de coeficientes del conjunto de entrenamiento, x_i^T representa la traspuesta de x_i . El símbolo $\odot$ representa el producto punto entre las matrices.

- Kernel Polinomial

$$K(x_i, x_j) = (\gamma x_i^T \odot x_j + r)^d \quad (18)$$

Cuando se emplea $d = 1$ y $r = 0$, el resultado es el mismo que el de un kernel lineal, si $d > 1$, se generan límites de decisión no lineales, aumentando la no linealidad a medida que d crece. El parámetro γ viene dado por el valor gamma de el vector de soporte de clasificación.

- Kernel RBF (Radial Basis Function)

$$K(x_i, x_j) = e^{-\gamma \|x_i x_j\|^2} \quad (19)$$

En esta función, la expresión $\|x_i x_j\|$ representa una matriz cuadrada con la distancia euclidiana entre cada par de puntos de x_i .

- Kernel Sigmoidal

$$K(x_i, x_j) = \tanh(\gamma x_i^T \odot x_j + r) \quad (20)$$

Donde los parámetros γ y r vienen dados por gamma y coef0 del SVC.

4.8 Medidas de rendimiento

Para evaluar los modelos se utilizan métricas que evalúan el comportamiento de los algoritmos en los datos de entrenamiento y prueba comparado con la predicción realizada en cada conjunto. Las más utilizadas se describen a continuación:

- Mean Squared Error (MSE)

$$MSE = \frac{1}{n} \left(\sum_{i=1}^n (\hat{y}_i - y_i)^2 \right) \quad (21)$$

Donde $\hat{y}_i$ representa el valor actual y y_i es el valor que el modelo predice. Para cada punto, calcula la diferencia entre la predicción y el valor objetivo y luego promedia los valores. Cuanto mayor sea este valor, peor es el modelo. Nunca es negativo, y los valores cercanos a cero indican un buen modelo.

- Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}} \quad (22)$$

Esta expresión equivale a la raíz cuadrada del MSE

- Mean absolute error (MAE)

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (23)$$

De acuerdo con lo anterior, se utilizan las expresiones antes mencionadas para evaluar el desempeño de los modelos propuestos.

Capítulo 5. Metodología y Datos

En este capítulo se describe la metodología utilizada, se presentan los conjuntos de datos empleados y se explica la construcción de los modelos de aprendizaje automático seleccionados en base a la revisión bibliográfica.

5.1 Conjunto de datos

Los conjuntos de datos utilizados provienen de tres estaciones de monitoreo medio ambientales ubicadas en las cercanías de la mina: Estaciones 460, Estación Llama y Estación Puma. Dichas estaciones monitorean la concentración de polvo en $[\mu\text{g}/\text{m}^3]$, temperatura en $[\text{°C}]$, presión atmosférica en $[\text{Pa}]$, humedad relativa del ambiente en porcentaje, velocidad del viento en $[\text{m}/\text{s}]$ y la dirección del viento en grados sexagesimales. Los datos facilitados son del 13 de agosto del 2019 al 28 de septiembre del mismo año, cada estación reporta 4370 datos en dicho período, registrando datos cada 15 minutos.

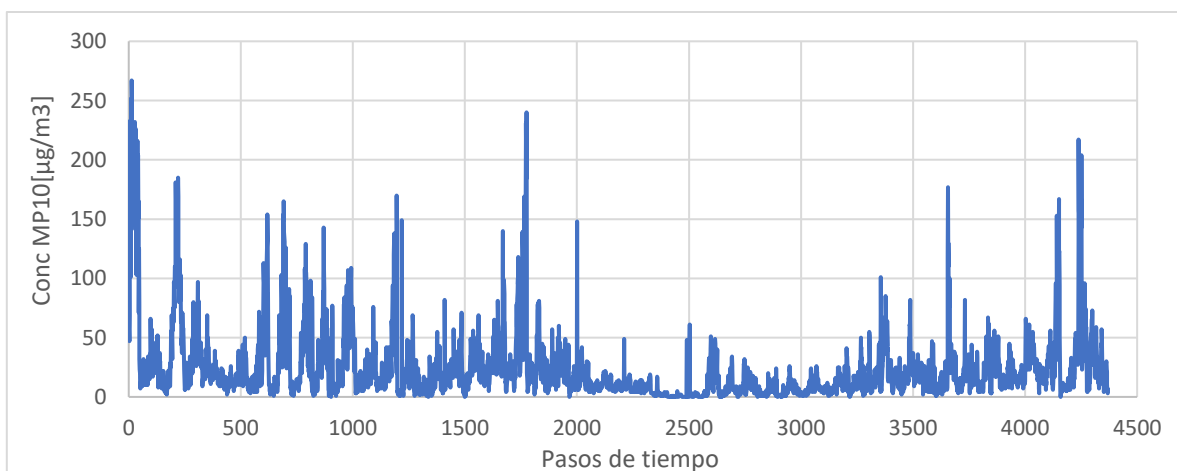


Figura 15. Estación 460 - Conc MP10 [UG/M3]. Fuente: elaboración propia.

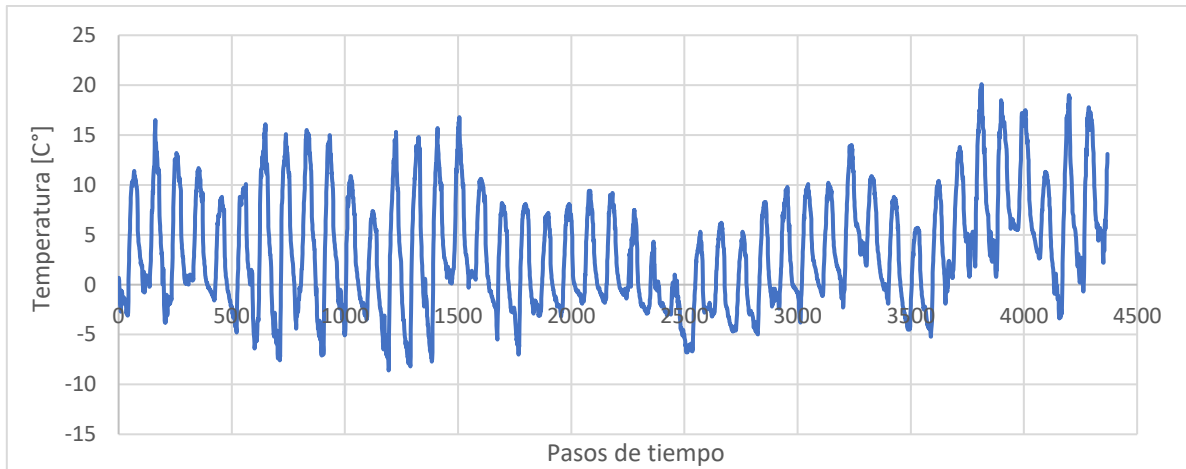


Figura 16. Estación 460 - Temperatura [°C]. Fuente: elaboración propia.

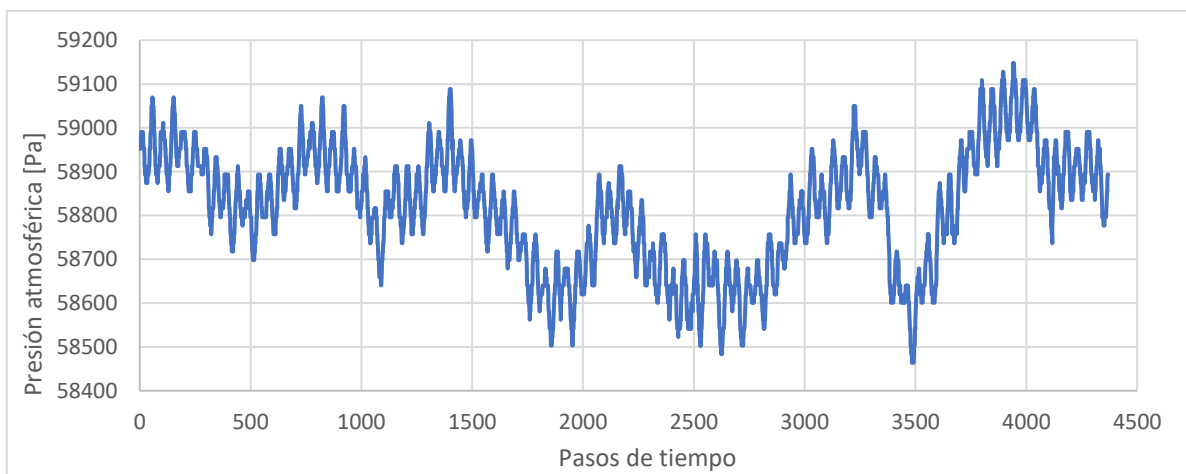


Figura 17. Estación 460 - Presión atmosférica [PA]. Fuente: elaboración propia.

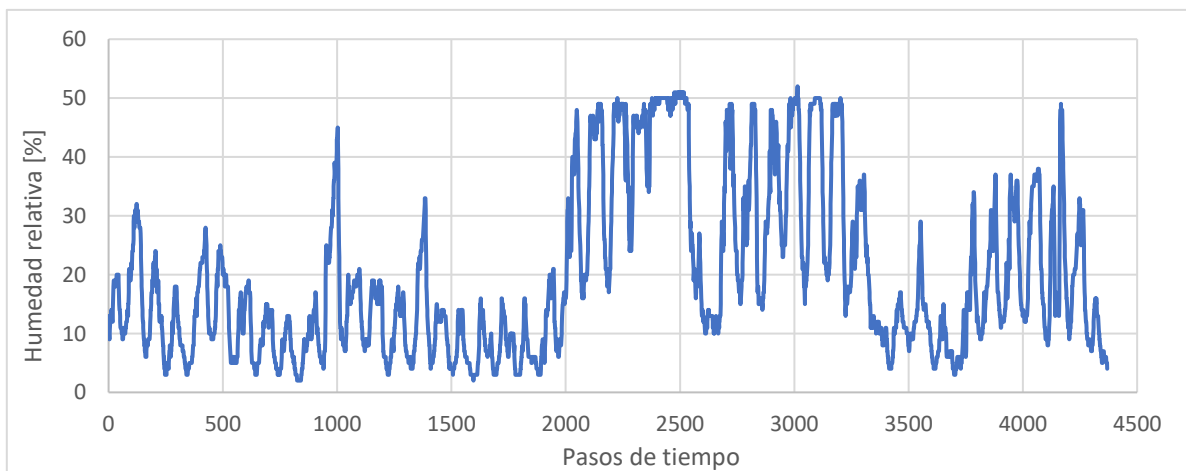


Figura 18. Estación 460 - Humedad relativa [%]. Fuente: elaboración propia.

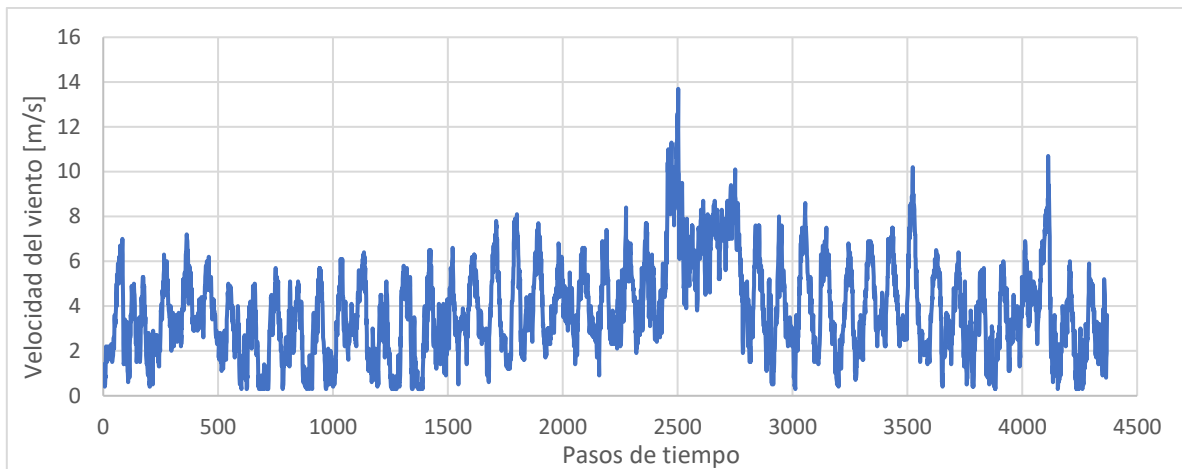


Figura 19. Estación 460 - Velocidad del viento [m/s]. Fuente: elaboración propia.

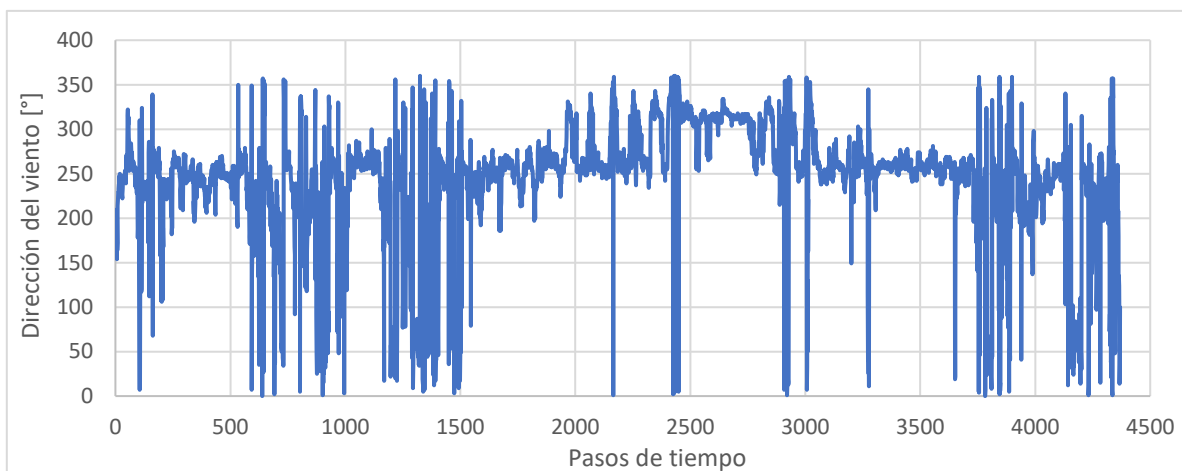


Figura 20. Estación 460 - Dirección del viento [°]. Fuente: elaboración propia.

Desde la figura 15, se observa que entre las mediciones de los pasos de tiempo 2000 a 3500 bajos niveles de MP10. Esto guarda relación con la figura 18 donde se muestra un alto porcentaje de humedad relativa para estas mismas mediciones. Además, en la figura 19, se observa un alza de velocidad del viento durante estas mediciones. Por tanto, la baja en la concentración de material particulado se debe en parte a la alta humedad relativa y a la velocidad del viento.

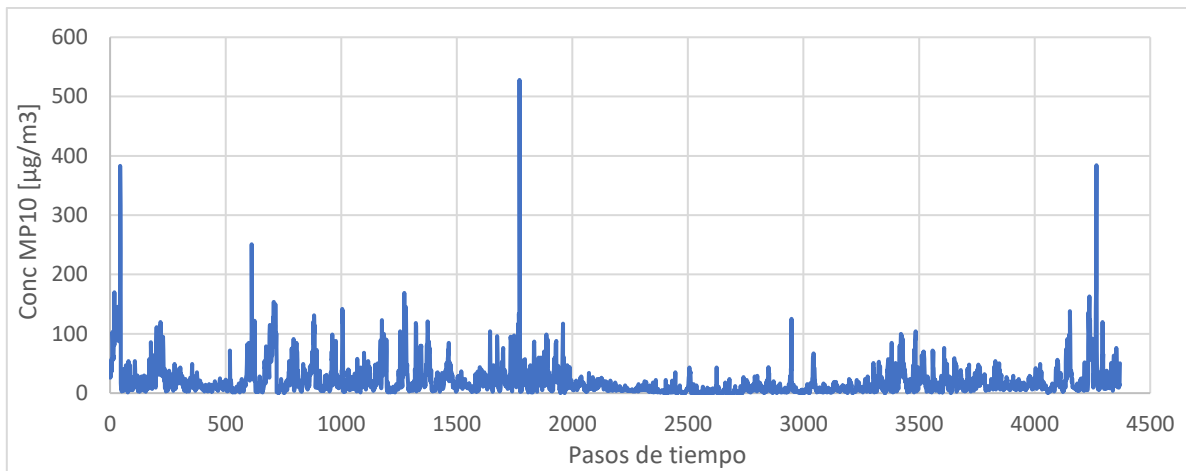


Figura 21. Estación Llama - Conc MP10 [UG/M3]. Fuente: elaboración propia.

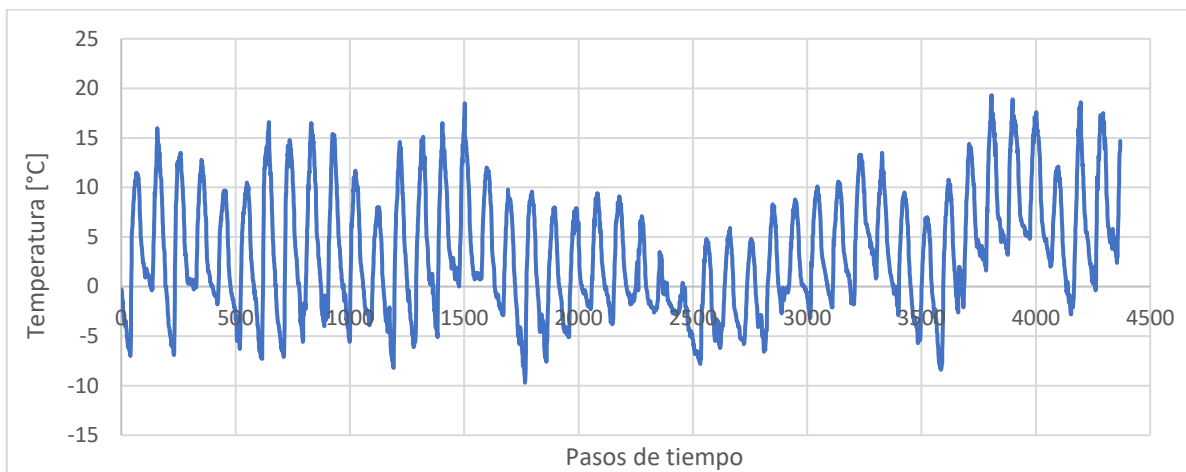


Figura 22. Estación Llama - Temperatura [$^{\circ}\text{C}$]. Fuente: elaboración propia.

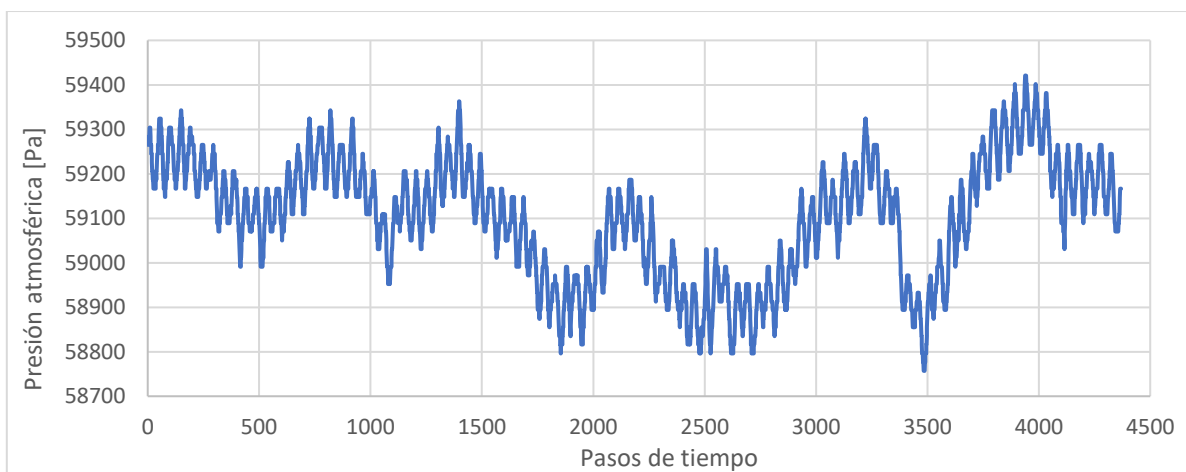


Figura 23. Estación Llama - Presión atmosférica [Pa]. Fuente: elaboración propia.

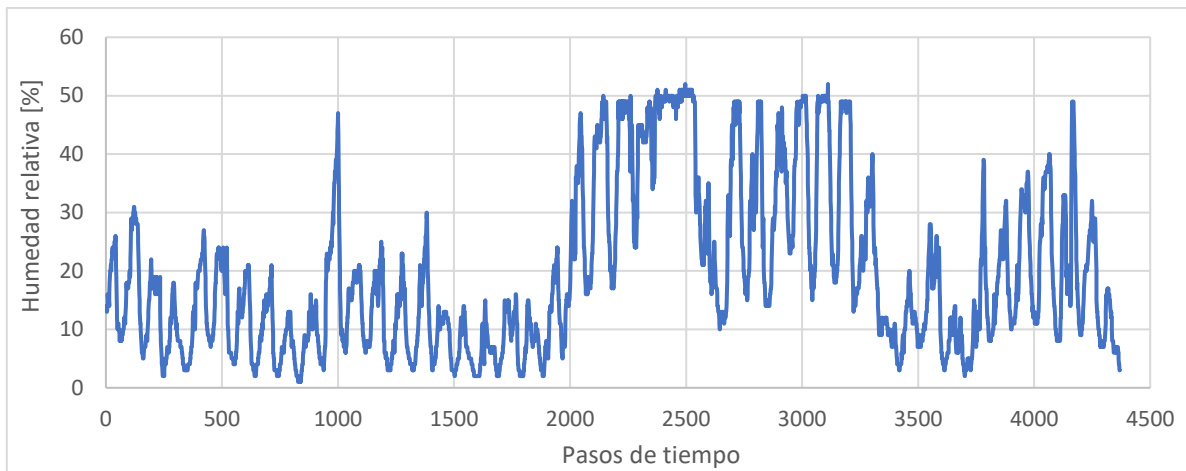


Figura 24. Estación Llama - Humedad relativa [%]. Fuente: elaboración propia.

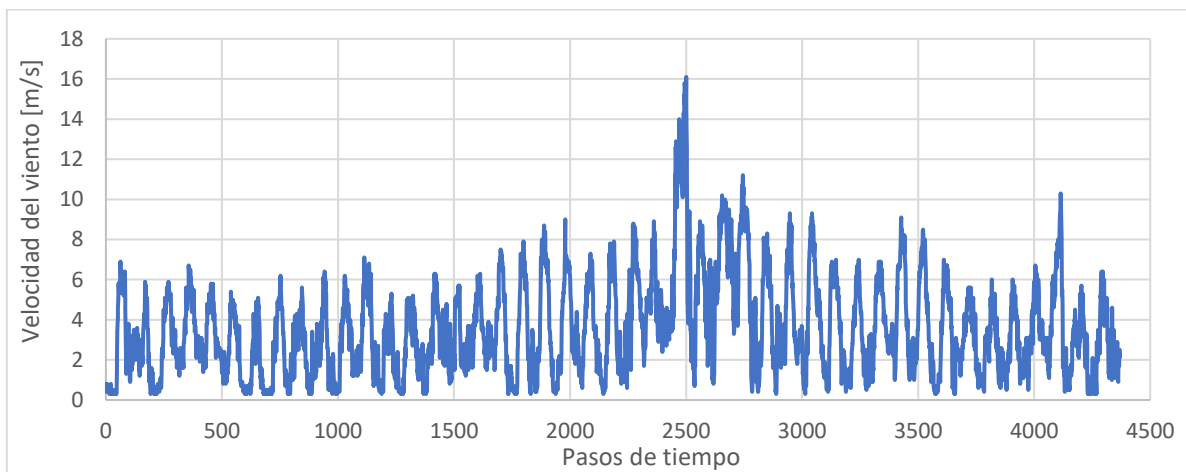


Figura 25. Estación Llama - Velocidad del viento [M/s]. Fuente elaboración propia.

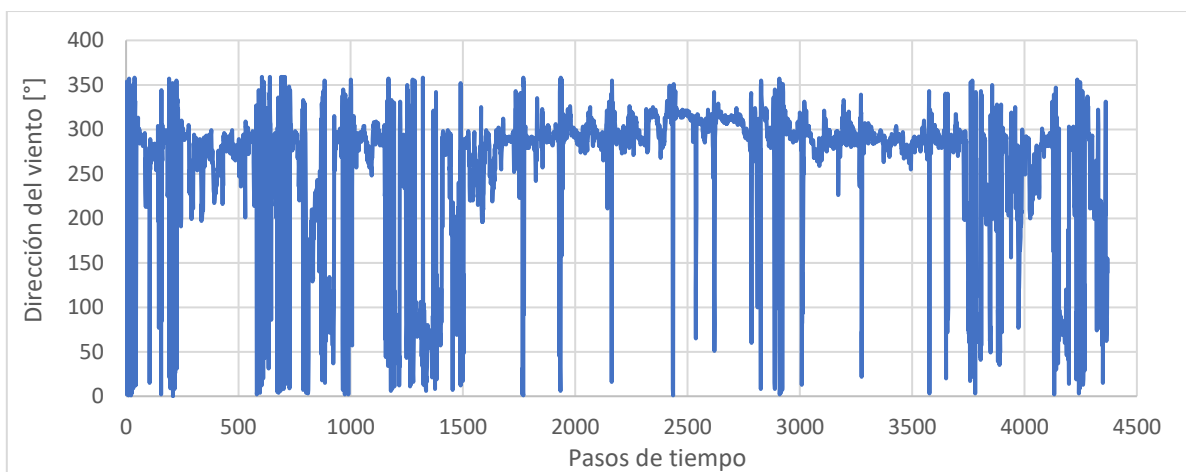


Figura 26. Estación Llama - Dirección del Viento [°]. Fuente: elaboración propia.

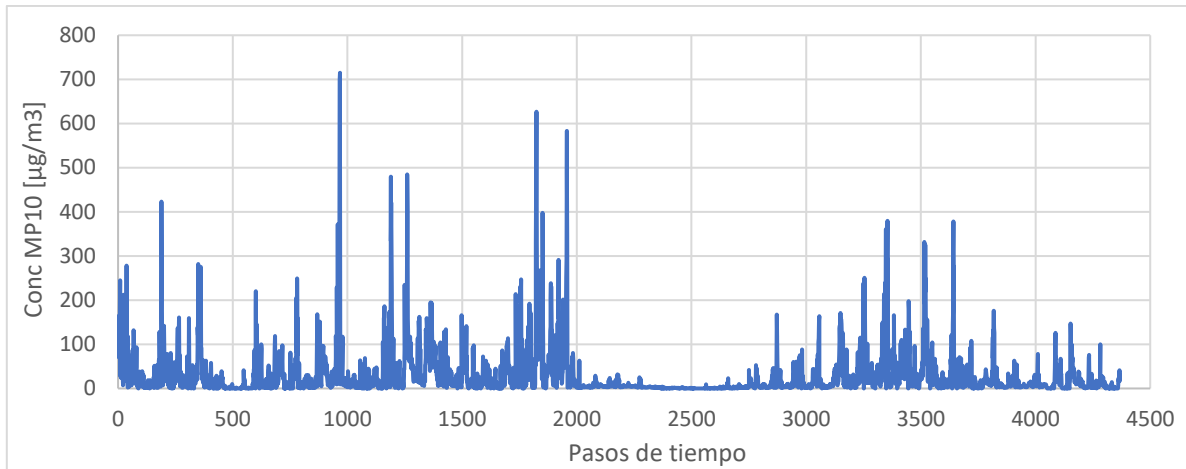


Figura 27. Estación Puma - Conc MP10 [UG/M3]. Fuente: elaboración propia.

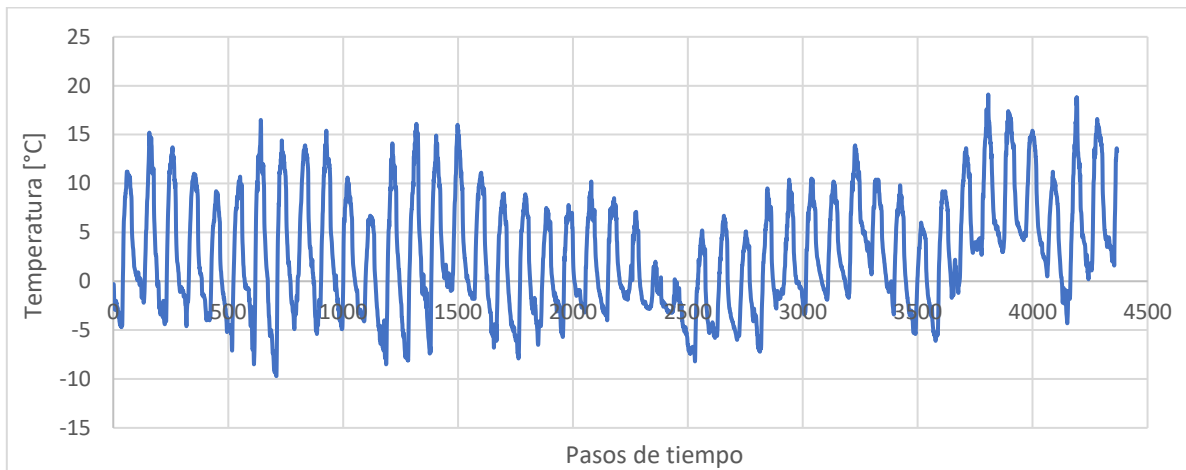


Figura 28. Estación Puma - Temperatura [°C]. Fuente: elaboración propia.

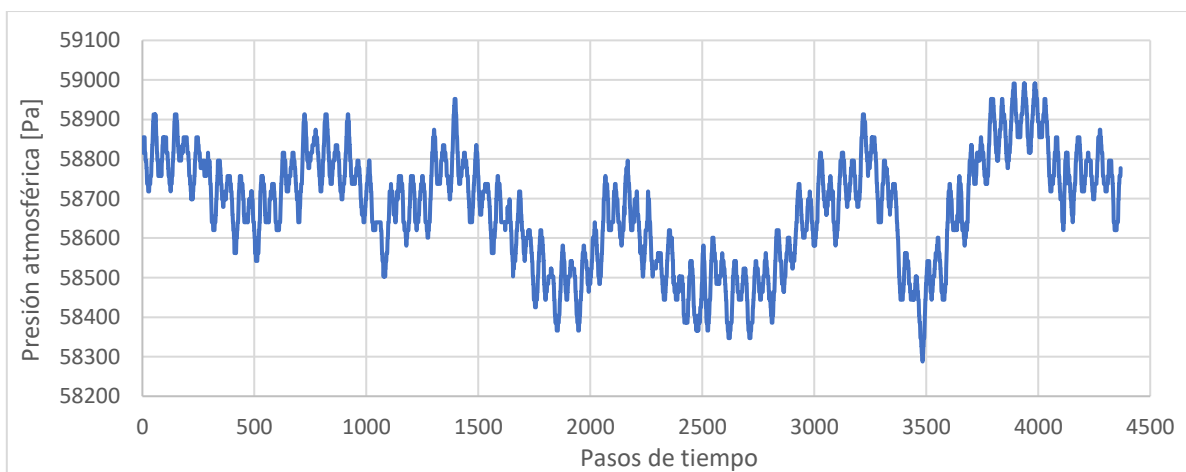


Figura 29. Estación Puma - Presión atmosférica [Pa]. Fuente: elaboración propia.

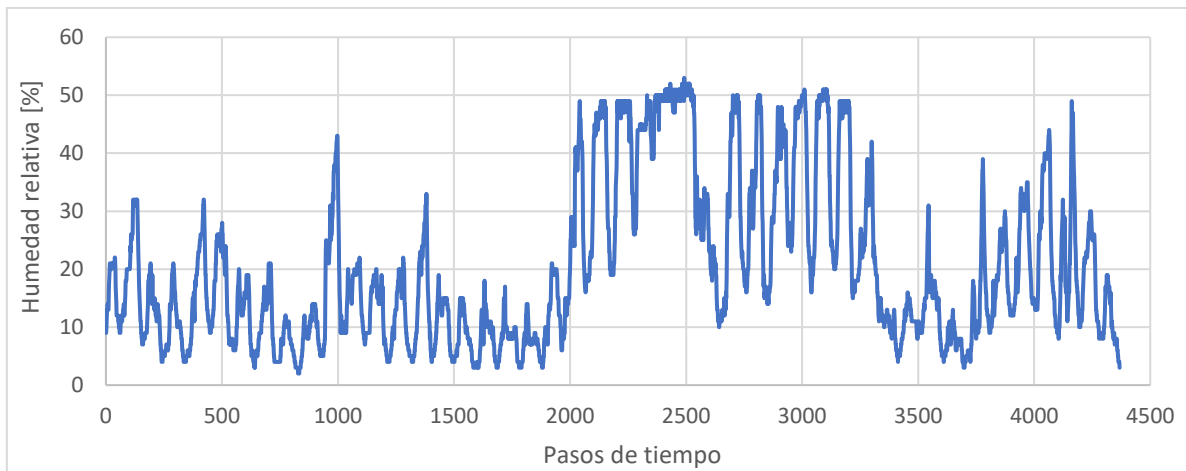


Figura 30. Estación Puma - Humedad relativa [%]. Fuente: elaboración propia.

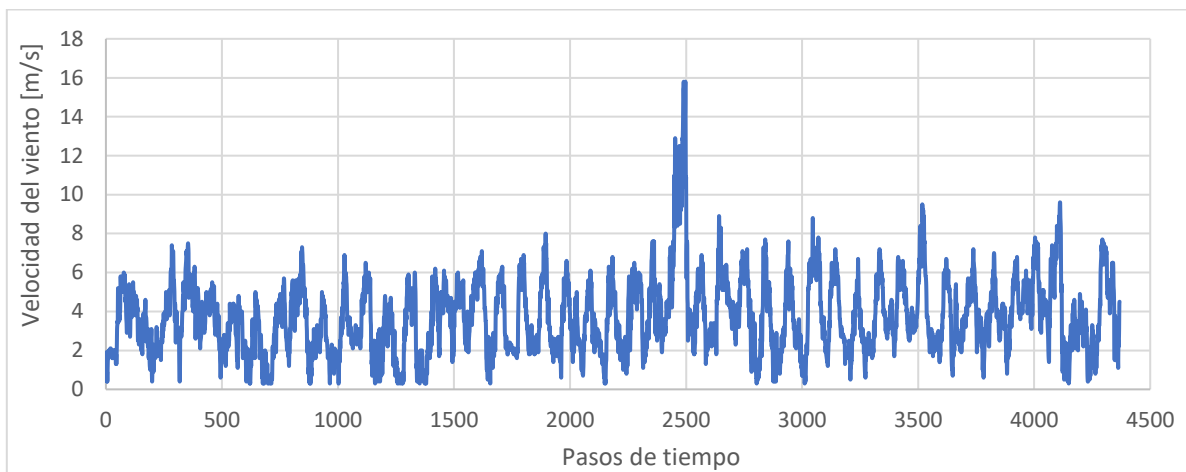


Figura 31. Estación Puma - Velocidad del viento [M/s]. Fuente: elaboración propia.

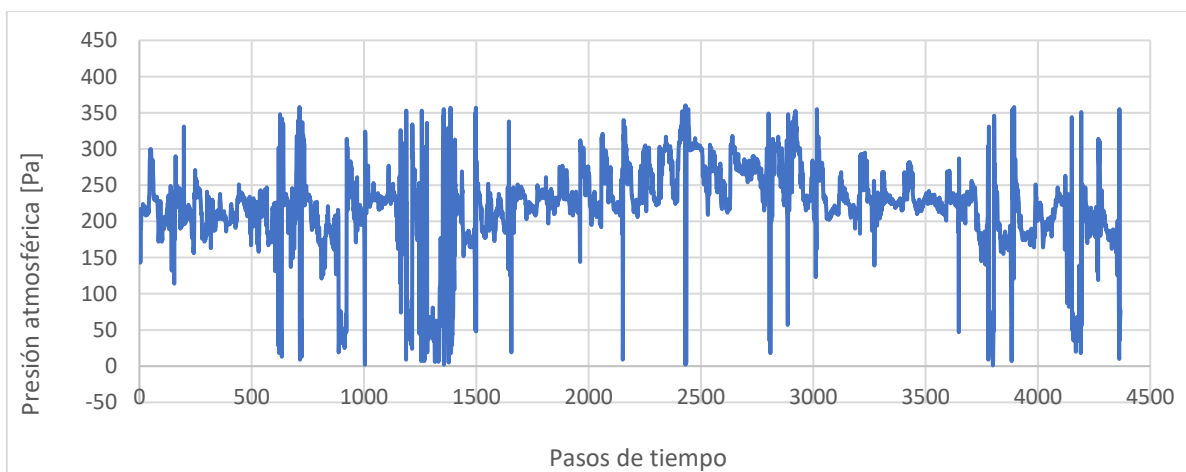


Figura 32. Estación Puma - Dirección del viento [°]. Fuente: elaboración propia.

En las ilustraciones antes presentadas, se observa el comportamiento en el tiempo del material particulado, temperatura, presión atmosférica, humedad relativa, velocidad del viento y la dirección del viento. Observándose un comportamiento acorde a las condiciones climáticas altiplánicas de este período. La toma de datos se realiza durante el periodo de invierno continental, caracterizado por fuertes vientos, bajas temperaturas, nevadas intensas, viento blanco, y congelamiento de caminos cuando se presentan malas condiciones climatológicas.

5.2 Análisis de datos.

A continuación, se presentan las estadísticas globales (medias, varianzas, cuantiles, etc.) de los conjuntos de datos utilizados. Donde **Conc** [$\mu\text{g}/\text{m}^3$], representa la concentración de material particulado MP10 en microgramos por metro cúbico. **T** [$^{\circ}\text{C}$], representa la temperatura en grados Celsius. **BP** [**Pa**], representa la presión barométrica en Pascales. **Rhi** [%], es la humedad relativa del ambiente en porcentaje. **WS** [**m/s**], es la velocidad del viento en metros por segundo. **WD** [$^{\circ}$], es la dirección del viento en grados sexagesimales.

Tabla 6. Estadísticas globales Estación 460.

Estación 460						
	Conc [$\mu\text{g}/\text{m}^3$]	T [$^{\circ}\text{C}$]	BP [Pa]	Rhi [%]	WS [m/s]	WD [$^{\circ}$]
Nº muestras	4369	4369	4369	4369	4369	4369
Media	23,76	3,32	58816	19,81	3,79	239
Des vest	30,42	5,72	136	14,18	2,06	70
Mínimo	0	-8,60	58464	2	0,30	0
25%	7	-1,20	58718	9	2,30	230
50%	15	2,40	58835	15	3,50	255
75%	28	7,50	58913	28	5,10	268
Máximo	267	20,10	59148	52	13,70	360

Tabla 7. Estadísticas globales Estación Llama.

Estación Llama						
	Conc [$\mu\text{g}/\text{m}^3$]	T [$^{\circ}\text{C}$]	BP [Pa]	Rhi [%]	WS [m/s]	WD [$^{\circ}$]
Nº muestras	4369	4369	4369	4369	4369	4369
Media	20,11	3,32	59100	19,74	3,58	254
Des vest	27,74	5,95	135	14,25	2,45	83
Mínimo	0	-9,70	58757	1	0,30	0
25%	6	-1,30	58991	8	1,70	250
50%	12	2,50	59109	16	3,10	288
75%	24	7,70	59206	27	5,10	298
Máximo	528	19,30	59421	52	16,10	359

Tabla 8. Estadísticas globales Estación Puma.

Estación Puma						
	Conc [$\mu\text{g}/\text{m}^3$]	T [$^{\circ}\text{C}$]	BP [Pa]	Rhi [%]	WS [m/s]	WD [$^{\circ}$]
Nº muestras	4369	4369	4369	4369	4369	4369
Media	28,03	2,77	58671	20,44	3,72	217
Des vest	54,1	5,86	141	14,08	1,97	60
Mínimo	0	-9,70	58288	2	0,30	1
25%	3	-1,90	58562	9	2,30	200
50%	9	1,90	58698	16	3,50	224
75%	30	7,10	58777	28	5,00	243
Máximo	699	19,10	58991	53	15,80	360

Para verificar la correlación entre las variables, y ver la codependencia que existe entre cada una de ellas, se utiliza una matriz de correlación. Donde los valores más cercanos a uno y menos uno indican una relación más fuerte entre los datos. Cabe mencionar que la correlación entre dos variables no implica por sí misma ninguna relación de causalidad. Además, no se considera la dirección del viento ya que estos valores varían entre 0° y 360° , por lo que posteriormente para el algoritmo esta variable se transforma a una categórica.

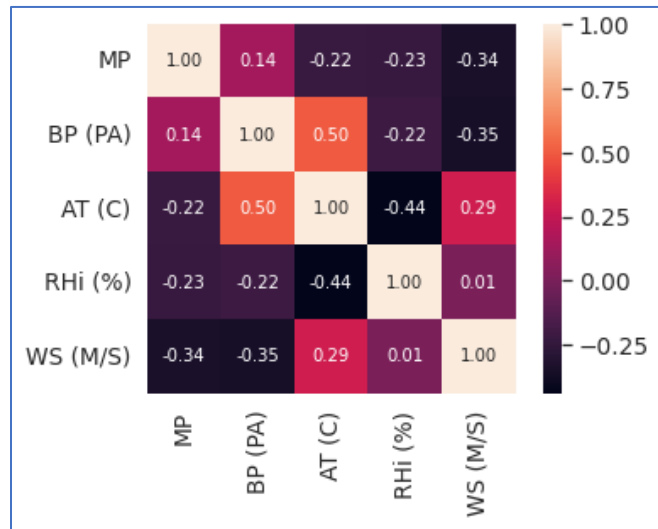


Figura 33. Matriz correlación datos estación 460.

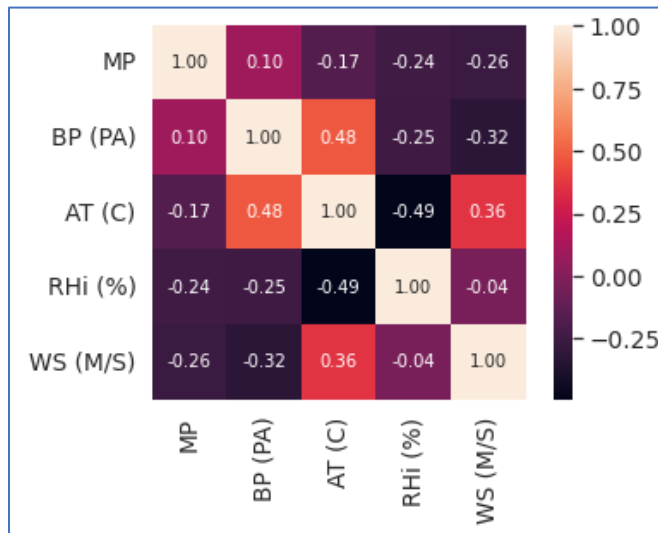


Figura 34. Matriz correlación datos estación Llama.

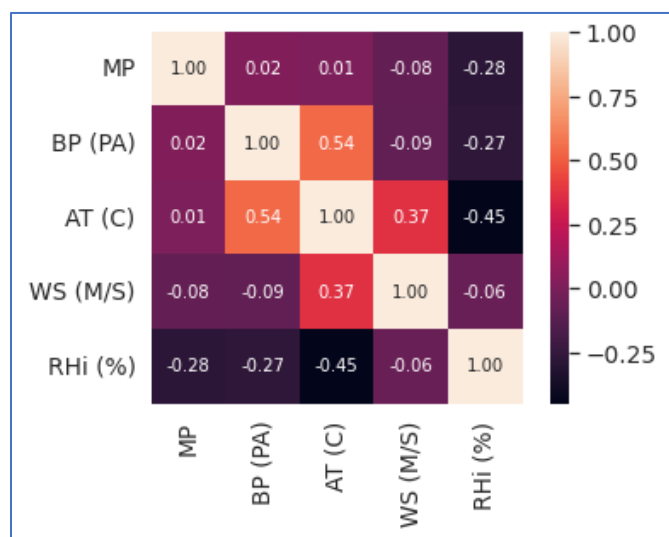


Figura 35. Matriz correlación datos estación Puma.

Conforme a las estadísticas globales y las matrices de correlación, se desprende que la concentración de material particulado tiene una mayor relación con la variación de velocidad del viento y con la humedad relativa del ambiente. Donde a mayor humedad relativa se tiene una menor concentración de MP10 y a mayor velocidad del viento menor nivel de polución.

5.3 Implementación de las metodologías

Para la predicción de la condición futura de material particulado se usan las metodologías descritas en los apartados 4.6 y 4.7 del capítulo 4. Dada la cantidad de datos disponibles, se separa del conjunto original de datos el último día de muestras. Estos datos se utilizarán para probar la capacidad de predicción del modelo. Para entrenar y validar el modelo se utilizan los datos comprendidos entre el 13 de agosto hasta el 27 de septiembre, de estos datos se utiliza el 70% para entrenar y el 30% restante para validar dichos modelos.

Para verificar la veracidad de los modelos, no solo en termino de las métricas, y comprobar la capacidad para predecir datos en él tiempo se usan las mediciones del 28 de septiembre para realizar la predicción y comparar los resultados obtenidos de las predicciones con los datos reales de material particulado.

Como ambos modelos corresponden a métodos de aprendizaje supervisado, se utiliza la variable de material particulado como variable objetivo, donde los modelos a su manera

interpretan la relación de las variables y predicen el MP10 considerando la temporalidad de las variaciones de cada dato de entrada.

Sobre la estructura general de los códigos elaborados para realizar la predicción de la serie temporal, se describen los siguientes pasos lógicos de los códigos en general:

1. Importar librerías de códigos al entorno de ejecución.
2. Importar conjunto de datos en formato csv.
3. Preparación de los datos
 - i. Análisis componente principales.
 - ii. Eliminar columnas que no aportan información.
 - iii. Cambiar formato de dato de la columna de fechas.
 - iv. Eliminar valores nulos y datos anómalos.
 - v. Unificar formato de datos tipo flotantes, enteros y objetos.
4. Graficar datos.
5. Preprocesamiento de datos
 - i. Codificación a variable categórica (MP10 y dirección del viento).
 - ii. Estandarización de variables de entrada.
 - iii. Convertir los datos a una serie supervisada (Caso LSTM).
 - iv. Dividir el conjunto de datos en entrenamiento y prueba.
6. Aplicación de algoritmo SVM y LSTM.
7. Validación del modelo.
8. Output de la predicción.

La primera función definida importa al entorno de ejecución (Anexos), todas las librerías relevantes que permiten procesar matrices de datos (Pandas), comandos que transforman el formato de fecha estructurado en día-mes-año-hora a un tipo de formato que el sistema puede interpretar de forma sencilla (Datetime). Se utiliza la librería de Matplotlib para plotear los datos y configurar gráficas interactivas. Se importa la librería de Sklearn para poder preprocesar los datos, codificando en variables categóricas, escalar y unificar las variables. Finalmente, desde la biblioteca de Google Colab, importa la función “Drive” que permite acceder a los archivos de una cuenta particular para extraer los archivos de texto que contienen los datos.

En la preparación de datos se utilizan las librerías antes mencionadas llamando a la función que importa los códigos y que a su vez abre los datos, en esta parte se eliminan las columnas del conjunto original que no aportan información. Se sigue con la transformación de la columna con la información de la fecha a un formato de datos que el algoritmo es capaz de leer, interpretar y procesar.

Se eliminan los valores nulos y los datos anómalos que pidieran representar un problema en los modelos y la predicción. Luego de este paso, con el comando “plt” de Matplotlib se grafican los datos a utilizar.

En la etapa de preprocesamiento, la columna con información sobre la dirección del viento se encuentra en grados sexagesimales, estos datos se convierten a una estructura de datos tipo objeto o “string”, transformando en función de los puntos cardinales según la siguiente tabla:

Tabla 9. Grado a punto cardinal

Punto cardinal	Rango en grados
N	0 - 11.25
NNE	11.25 - 33.75
NE	33.75 - 56.25
ENE	56.25 - 78.75
E	78.75 - 101.25
ESE	101.25 - 123.75
SE	123.75 - 146.25
SSE	146.25 - 168.75
S	168.75 - 191.25
SSW	191.25 - 213.75
SW	213.75 - 236.25
WSW	236.25 - 258.75
W	258.75 - 281.25
WNW	281.25 - 303.75
NW	303.75 - 326.25
NNW	326.25 - 348.75
N	348.75 - 360

Primero es necesario realizar esta transformación para posteriormente codificar los valores, ya que para el modelo es más sencillo interpretar la relación entre las variables numéricas que tienen un mayor impacto en la predicción MP10 con la variable categórica de dirección

del viento. Con esta última variable sería necesario abordar una metodología más compleja que considera vectores direccionales y relaciones de distancia entre las estaciones de monitoreo, lo cual no es abordado en este estudio.

De igual forma, los datos de concentración de material particulado tienen un amplio rango de variación que va desde el 0 hasta valores superiores a 700 [$\mu\text{g}/\text{m}^3$], por lo que predecir un valor preciso en este rango es complicado. Considerando los datos y las estadísticas globales, se utilizan los siguientes valores categóricos para los rangos establecidos de material particulado:

Tabla 10. Valor categórico para rangos de MP10.

Valor categórico	Rango MP10 [mg/m^3]
1	0-25
2	25-50
3	50-75
4	75-100
5	100-150
6	150-1000

Con el comando LabelEncoder de la biblioteca Sklearn, se transforman los datos en formato “string” de la dirección del viento a valores numéricos, así el modelo puede procesar e interpretar los datos, y finalmente relacionarlos con la variable de entrada original. Con esto, todos los datos de entrada se encuentran en un formato amigable con el modelo, pero se encuentran en distintos rangos de escala. Para ello se utiliza el estimador “MinMaxScaler” de Sklearn para escalar todas las variables en un mismo rango, configurando el escalador con una media cero y varianza unitaria.

Para la metodología LSTM, es necesario definir una función que transforma los datos a una estructura de serie supervisada, esto quiere decir que para predecir el paso de tiempo actual “ t ” se considera la medición de MP10 y las variables meteorológicas del paso de tiempo anterior. De esta forma, para predecir el tiempo actual, la serie supervisada considera la información de los pasos de tiempo “ $t - n$ ”. Por lo para anticiparse a una mayor cantidad de pasos de tiempo en el futuro se puede considerar más información. Este paso es el más

importante para este método, dado el uso de la retro propagación a través de los datos ordenados en series temporales.

Con los datos estandarizados, se procede a separar los datos en conjuntos de entrenamiento y prueba, considerando el 70% de datos para el entrenamiento y el restante para el testeo. En esta etapa los datos se separan en 4 matrices: `train_x` y `test_x` contienen los datos de entrenamiento y prueba respectivamente, considerando las columnas de material particulado y variables meteorológicas. Las matrices denominadas como `train_y` y `test_y` contienen los datos de entrenamiento y prueba respectivamente pero solo considerando la columna de material particulado. Además, como los datos vienen ordenados siguiendo la temporalidad en que fueron tomados, se separan considerando un orden cronológico. De otro modo se perdería la relación entre la variable y el tiempo en que estas ocurren.

Para el modelo LSTM, se define una función que importa las librerías de Keras y TensorFlow, estas contienen los comandos que permiten aplicar este método de machine learning. Se importa el comando “LSTM” donde se configura el número de nodos que debe tener la capa, se indican las matrices de entrenamiento, y se configura un regularizador que evita el sobreajuste del modelo. El código solicita la cantidad de lags a considerar para transformar los datos a una estructura de serie supervisada. En el modelo se utiliza como función de pérdida el “mse” o mean squared error, como función optimizadora “Adam” y las métricas para evaluar “mse”, “mae”, “rmse” y “accuracy”. Luego se dan las instrucciones para que el modelo evalúe su desempeño en el conjunto de prueba en función de las métricas antes mencionadas. Una vez que el modelo se encuentra entrenado y evaluado en los conjuntos, se realiza la predicción donde el código pide al usuario la cantidad de pasos en el tiempo que quiere anticipar.

En el caso del modelo SVM, se sigue el mismo procedimiento hasta la definición del método de aprendizaje automático, desde la librería Sklearn se importa el comando “svm” y “SVR”. Los códigos entrenan 4 modelos distintos consideran los 4 kernels disponibles en las librerías: “rbf”, “poly”, “linear” y “sigmoidal”. Estas funciones mapean los datos que se encuentran en una dimensión inferior a una superior para ajustar un hiperplano que modele el comportamiento entre las variables. Los cuatro modelos se ajustan con un parámetro C igual a 10. Posteriormente se entrenan los modelos con la función “fit” en los conjuntos

respectivos. Para validar los modelos y obtener las métricas de desempeño, primero se debe utilizar el comando “TimeSeriesSplit”, el cual es una variación del clásico método k-folds. Los datos de series de tiempo se caracterizan por la correlación entre las observaciones y el tiempo en que fueron realizadas, por lo que este comando toma los conjuntos de entrenamiento sucesivos en superconjuntos de los anteriores, vale decir, si inicialmente se considera la partición k como conjunto de entrenamiento y el conjunto $k+1$ como conjunto de prueba, en la siguiente validación se considera el conjunto $k+1$ como conjunto de entrenamiento y $k+2$ como set de prueba. De esta forma se sigue hasta validar todo el conjunto de datos. Finalmente, se aplica el comando “predict” a los cuatro modelos de SVM, donde el código pide al usuario la cantidad de pasos en el tiempo que quiere anticipar.

Los códigos de ambos modelos se encuentran en la sección de anexos.

5.4 Entorno de ejecución

Se utiliza el entorno Google Colaboratory, también llamado Colab, el cual permite ejecutar y programar en Python sin la necesidad de tener un equipo poderoso para procesar grandes conjuntos de datos, requiere una cuenta en Google y conexión a internet para tener acceso a esta herramienta. Dentro de sus ventajas destacan el acceso gratuito a GPUs, no requiere ninguna configuración, permite compartir contenido fácilmente, aprovechar toda la potencia de las bibliotecas más populares de Python para analizar, procesar y visualizar datos, y permite desarrollar y entrenar redes neuronales.

Capítulo 6. Experimentos y resultados

6.1 LSTM

Los hiper – parámetros relevantes del modelo LSTM son la cantidad de *memory cells* en las capas ocultas y las funciones de activación de los nodos. La arquitectura configurada considera 3 capas LSTM con 64, 32 y 16 nodos cada una, las funciones de activación que se utilizaron son: Tangente hiperbólica (tanh), Sigmoidal y ReLU.

Otro parámetro que se analiza en los experimentos es la cantidad de pasos de tiempo anteriores que se consideran para predecir el tiempo presente, estos pasos corresponden a la cantidad de instantes que el gradiente se propaga en el tiempo. Además, se consideran dos tamaños de lotes distintos (Batch), donde este número es la cantidad de muestras que se procesan antes de que se actualice el modelo. Los resultados obtenidos de la validación del modelo, en término de las medidas de rendimiento expuestas en el punto 4.8 se muestran en las siguientes tablas:

Tabla 11. Medidas de rendimiento LSTM datos 460.

Estación 460						
	Batch 16			Batch 32		
Pasos	1	10	38	1	10	38
MSE	0,0286	0,0313	0,0312	0,0281	0,0281	0,0286
RMSE	0,1691	0,1769	0,1766	0,1676	0,1676	0,1691
MAE	0,087	0,0917	0,0946	0,1128	0,1149	0,1149
Precisión	72,3%	72,3%	72,3%	73,2%	72,8%	72,8%

Tabla 12. Medida de rendimiento LSTM datos Llama.

Estación Llama						
	Batch 16			Batch 32		
Pasos	1	10	38	1	10	38
MSE	0,0247	0,0249	0,0246	0,0232	0,0232	0,023
RMSE	0,1572	0,1578	0,1568	0,1523	0,1523	0,1517
MAE	0,0754	0,0749	0,0732	0,0966	0,1	0,0973
Precisión	78,7%	78,7%	79,6%	78,8%	78,7%	79,4%

Tabla 13. Medidas de rendimiento LSTM datos Puma.

Estación Puma						
	Batch 16			Batch 32		
Pasos	1	10	38	1	10	38
MSE	0,063	0,0635	0,0652	0,0585	0,0589	0,0589
RMSE	0,2510	0,2520	0,2553	0,2419	0,2427	0,2427
MAE	0,1387	0,1441	0,1454	0,1709	0,17	0,1731
Precisión	70,8%	70,9%	70,2%	70,8%	70,9%	70,2%

La abreviación MSE se refiere a “Mean squared error”, RMSE es “Root mean squared error” y MAE es la abreviación de “Mean absolute error”.

Conforme a los resultados de las métricas globales obtenidas, no muestran resultados tan concluyentes en base a si el modelo mejora cuando se utiliza más información para anticipar el tiempo presente. Considerando la cantidad de datos por lotes o batch size, se ve una disminución del error cuadrático medio y del error absoluto cuando aumenta el batch size. Dentro de las sucesivas iteraciones se verifican las gráficas obtenidas al comparar entre los datos reales y la predicción realizada. Haciendo una comparación visual, se tiene que para los datos de la estación 460, el mejor resultado se obtiene con un batch size de 32 y un paso de tiempo anterior para realizar la predicción. Para los datos de la estación Llama, la mejor gráfica se obtiene con un batch size de 16 y un 38 pasos de tiempo anteriores para el anticipar el paso presente. Considerando los datos de la estación Puma, el resultado gráfico no es tan concluyente ya que, para diversas iteraciones, utilizando ambos batch size y tiempo de lag entre 10 y 38, el resultado gráfico es similar. A continuación, se presentan las gráficas obtenidas al predecir los datos del 28 de septiembre luego de entrenar y validar los modelos.

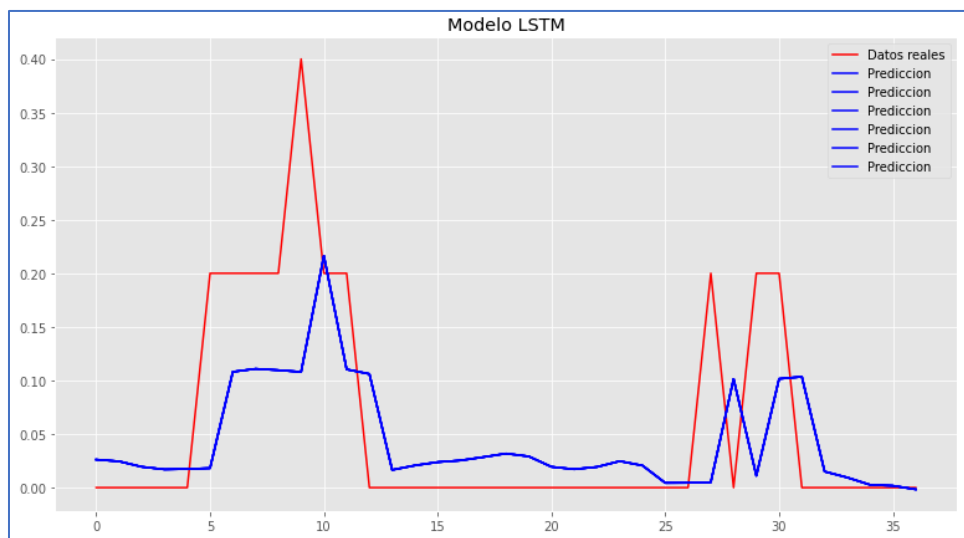


Figura 36. Predicción LSTM datos 460.

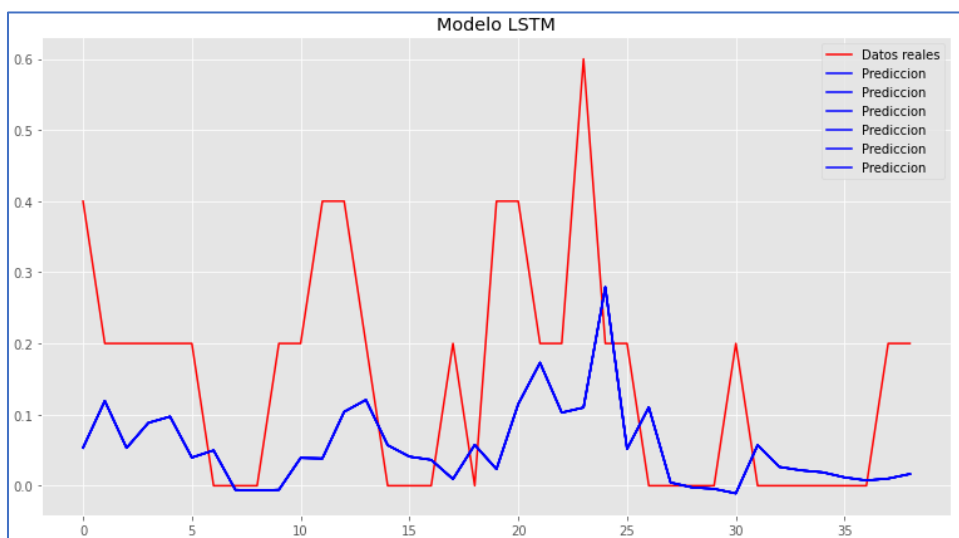


Figura 37. Predicción LSTM datos Llama.

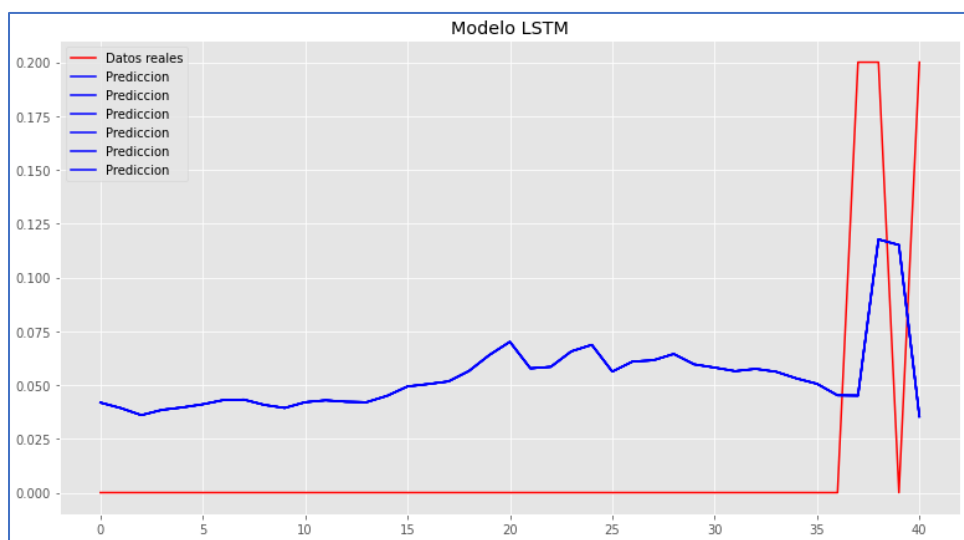


Figura 38. Predicción LSTM datos Puma.

6.2 SVM

Conforme a la metodología de SVM, se consideran cuatro modelos con los kernel RBF, polinomial, lineal y sigmoideal. En este caso el único hiper – parámetro que se podría configurar sería el parámetro “C”, el cual señala la cantidad de muestras que se permite que sobrepasen los márgenes a una distancia épsilon del hiperplano, tal como muestra la ilustración n°15. En este caso de estudio se utiliza el parámetro que viene por defecto con el modelo. Los resultados obtenidos en término de las medidas de rendimiento expuestas en el punto 4.8 se muestran en las siguientes tablas:

Tabla 14. Medidas de rendimiento SVM datos 460.

Estación 460				
Kernel	RBF	Polinomial	Lineal	Sigmoideal
MSE	0,009	0,038	0,008	56290
RMSE	0,092	0,173	0,090	214
MAE	0,078	0,111	0,088	161

Tabla 15. Medidas de rendimiento SVM datos Llama.

Estación Llama				
Kernel	RBF	Polinomial	Lineal	Sigmoidal
MSE	0,010	0,015	0,008	70397
RMSE	0,095	0,116	0,089	236
MAE	0,083	0,085	0,087	177

Tabla 16. Medidas de rendimiento SVM datos Puma.

Estación Puma				
Kernel	RBF	Polinomial	Lineal	Sigmoidal
MSE	0,010	0,037	0,008	22290
RMSE	0,100	0,166	0,090	140
MAE	0,080	0,093	0,087	99

Conforme a las métricas de desempeño, el modelo con una kernel radial presenta un mejor desempeño. De igual forma, al comparar la curva de valores reales con las predicciones realizadas de los modelos se verifica que la curva que se ajusta de mejor forma a los datos reales es la curva del modelo RBF. Sobre el modelo sigmoidal no se muestra en las gráficas ya que los resultados que entrega se encuentran muy por fuera de los valores de la curva de datos reales, lo cual guarda relación con las métricas de desempeño obtenidas en los tres casos. A continuación, se presentan las gráficas obtenidas al predecir los datos del 28 de septiembre luego de entrenar y validar los modelos.

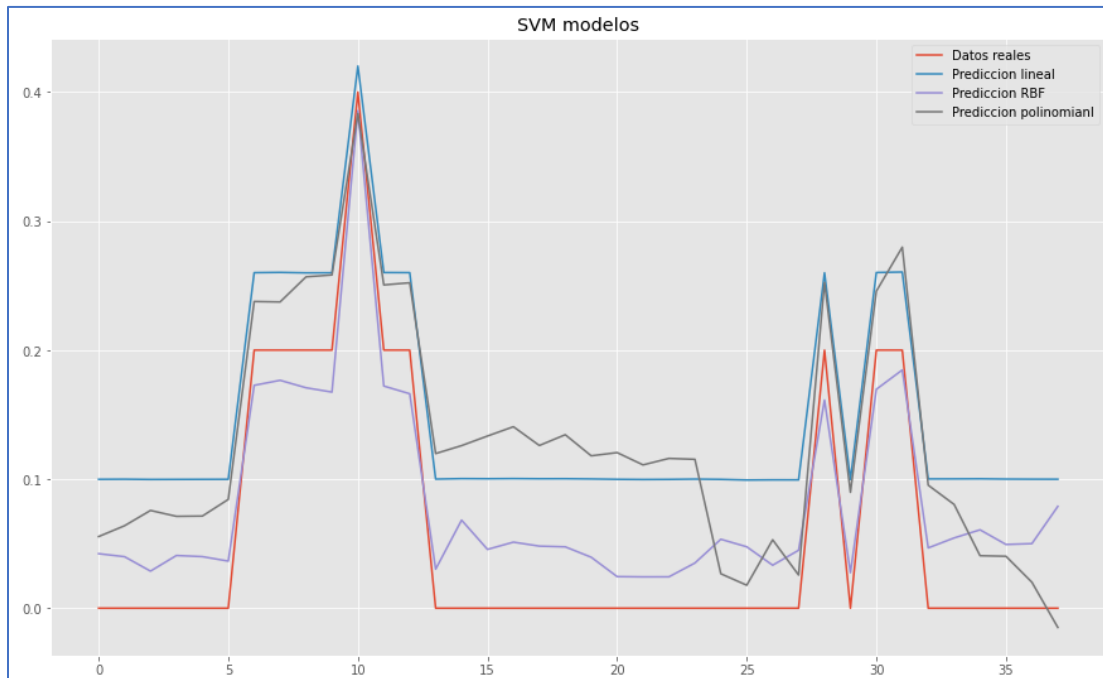


Figura 39. Predicción SVM datos 460.

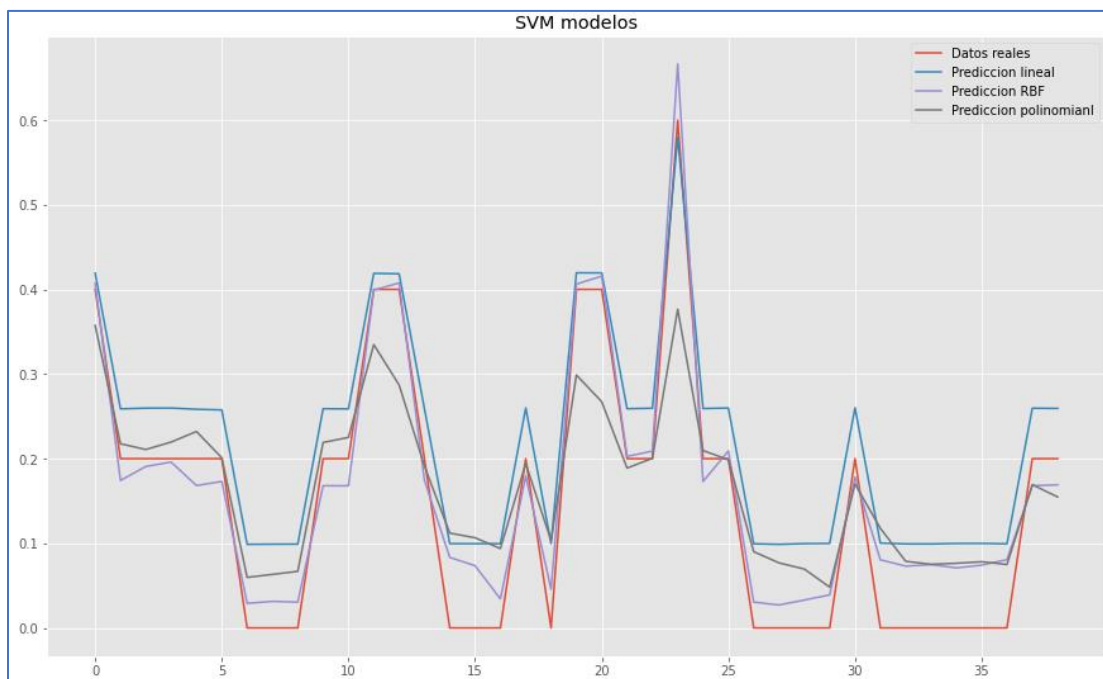


Figura 40. Predicción SVM datos Llama.

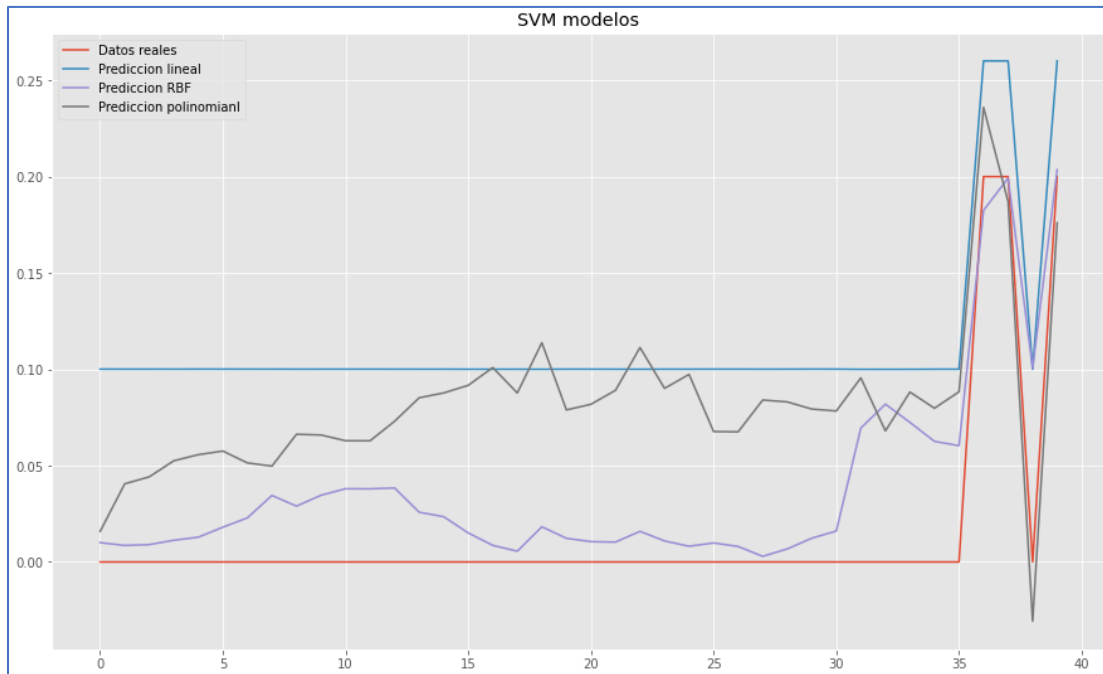


Figura 41. Predicción SVM datos Puma.

Verificando los valores que entregan los modelos, se observa que en los tres gráficos la curva del modelo radial o RBF es la que más se aproxima a la curva de datos reales. Estas curvas representan los valores escalados y normalizados de los resultados de la predicción y de los datos reales, por lo que al volver este valor a la variable categórica real se aprecia que el modelo RBF acierta en todos los datos reales del día que se predice.

Capítulo 7. Conclusiones y recomendaciones

En este trabajo de memoria, se intenta predecir la condición de material particulado en base a datos meteorológicos de tres estaciones de monitoreo. De acuerdo con la metodología de LSTM no se obtiene un buen resultado ya que las métricas de desempeño no entregan resultados concluyentes en base a los hiper parámetros que se iteran para buscar una solución más aproximada al problema real. Por otra parte, al verificar las gráficas obtenidas al predecir los valores reales, con los datos de las estaciones 460 y Puma se obtiene una curva de predicción que se asemeja a los datos reales, pero en el caso de la estación Llama donde los datos a predecir son bastante cambiantes en el tiempo, la curva obtenida no presenta ninguna correlación.

Sin embargo, sobre la metodología *support vector machine*, las métricas de desempeño muestran que el kernel radial o RBF tiene el mejor desempeño entre los cuatro modelos propuestos, obteniendo los valores más cercanos a cero para las medidas rendimiento utilizadas. Cabe mencionar que los modelos lineales y polinomiales tienen un desempeño levemente menor que RBF. Además, comparando las gráficas obtenidas la curva que mejor se ajusta a los datos reales es la curva de predicción radial.

En síntesis, este trabajo demuestra que por medio de estas metodologías es posible predecir la condición de material particulado en base a datos históricos de concentración de polvo y datos meteorológicos, esta herramienta es aplicable a la problemática minera y puede ser usada para anticiparse a las condiciones desfavorables de polución para la operación, permitiendo realizar una mejor gestión de los recursos disponibles en la compañía para abatir el polvo.

A modo de recomendación se propone lo siguiente:

- Se recomienda continuar con el estudio para aplicar los algoritmos de machine learning a la operación real, ya que se presenta como una herramienta de vanguardia para ayudar a optimizar los distintos procesos mineros.
- Profundizar en LSTM y SVM, ya que existen metodologías que permiten optimizar los hiper parámetros del modelo en base al comportamiento de los datos, lo que aumenta el poder predictivo de los modelos.

- Se recomienda utilizar más datos para entrenar los modelos.
- Integrar más de un modelo para predecir la condición de material particulado. Ya que, desde los antecedentes, los estudios muestran que modelos híbridos de machine learning tienen un mejor comportamiento cuando se intenta predecir el material particulado.
- Hay que considerar que, una vez instalado el sistema de monitoreo, este entregara variables que no se consideran en el modelo planteado, sin embargo, es posible hacer ajustes en los códigos para que consideren estas nuevas variables de entrada.

Referencias bibliográficas

1. Athira, V., Geetha, P., Vinayakumar, R., & Soman, K. P. (2018). Deepairnet: Applying recurrent networks for air quality prediction. *Procedia computer science*, 132, 1394-1403.
2. BHP. (2020, septiembre 18). Como la gestión estratégica del polvo está mejorando nuestras emisiones atmosféricas. Recuperado de: <https://www.bhp.com/sustainability/case-studies/2020/09/dust-managements/>
3. BHP. (2019, mayo 29). Control de polvo en Port-Hedland. Recuperado de: <https://www.bhp.com/sustainability/community/community-news/2019/05/dust-control-in-port-hedland/>
4. BHP. (2019, septiembre 26). Mt Arthur Coal Annual Review FY19.
5. Biancofiore, F., Busilacchio, M., Verdecchia, M., Tomassetti, B., Aruffo, E., Bianco, S. & Di Carlo, P. (2017). Recursive neural network model for analysis and forecast of PM10 and PM2.5. *Atmospheric Pollution Research*, 8(4), 652-659.
6. Cecala AB, O'Brien AD, Schall J, Colinet JF, Franta RJ, Schultz MJ, Haas EJ, Robinson J, Patts J, Holen BM, Stein R, Weber J, Strebel M, Wilson L, and Ellis M. (2019). Dust control handbook for industrial minerals mining and processing. Second edition. Pittsburgh PA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health, DHHS (NIOSH) Publication No. 2019-124, RI 9701.
7. Gers, A; Schmidhuber, J; Cummins, F. (2000). Aprender a olvidar: Predicción continua con LSTM. *Computación neuronal*.
8. Hochreiter, S; Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735-1780, 1997.
9. Huang, C.-J.; Kuo, P.-H. Un modelo CNN-LSTM profundo para la previsión de partículas (PM_{2,5}) en ciudades inteligentes. *Sensores* **2018**, 18, 2220. Recuperado de: <https://doi.org/10.3390/s18072220>
10. IBM, Corporation. (2021). Recuperado de: <https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=networks-neural-model>
11. Ley 19.300, Decreto N°12. Biblioteca del Congreso Nacional de Chile. Santiago de Chile. Publicada el 09 de mayo de 2011.

12. Ley 19.300, Decreto N°59. Biblioteca del Congreso Nacional de Chile. Santiago de Chile. Publicada el 17 de octubre de 2015.
13. Ministerio del medio ambiente. (2016). Guía de calidad del aire y educación medioambiental. Recuperado de: <https://mma.gob.cl/wp-content/uploads/2018/08/Guia-para-Docentes-Sobre-Calidad-del-Aire-003.pdf>
14. Nath, P., Saha, P., Middya, AI y Roy, S. (2021). Pronóstico de contaminación de series de tiempo a largo plazo utilizando métodos estadísticos y de aprendizaje profundo. *Computación neuronal y aplicaciones* , 1-20.
15. Port Hedland Industries Council. (2019). PHIC- Real time data reporting. Recuperado de: <http://www.phicmonitoring.com.au/index.html>
16. Ríos, G., & Hurtado, C. (2008). Series de tiempo. *Universidad de Chile. Facultad de Ciencias Fisicas y Matematicas*, 52.
17. Suleiman, A; Tight, Miles; Quinn, Andrew. (2018). Applying machine learning methods in managing urban concentrations of traffic-related particulate matter (PM10 and PM2.5). *Atmospheric Pollution Research*. Recuperado de: <https://doi.org/10.1016/j.apr.2018.07.001>
18. Vapnik, Vladimir. (1995). *The Nature of Statistical Learning Theory*, Springer, New York.
19. Weizhen, H., Zhengqiang, L., Yuhuan, Z., Hua, X., Ying, Z., Kaitao, L., Yan, M. (2014, marzo). Uso de la regresión de vectores de soporte para predecir PM10 y PM2. 5. En *la serie de conferencias IOP: ciencias de la tierra y el medio ambiente* (Vol. 17, No. 1, p. 012268). Publicación de IOP.
20. Xayasouk, T., Lee, H., & Lee, G. (2020). Air pollution prediction using long short-term memory (LSTM) and deep autoencoder (DAE) models. *Sustainability*, 12(6), 2570. Recuperado de: <https://www.mdpi.com/2071-1050/12/6/2570>
21. Xing, Y.-F.; Xu, Y.-H.; Shi, M.-H.; Lian, Y.-X. The impact of PM2.5 on the human respiratory system. *J. Thorac. Dis.* 2016, 8, E69–E74. Recuperado de: <https://www.sciencedirect.com/science/article/abs/pii/S0045653518324639>
22. Zhao, J., Deng, F., Cai, Y. y Chen, J. (2019). Memoria a corto plazo: Red neuronal completamente conectada (LSTM-FC) para PM2. 5 predicción de

concentración. *Chemosphere* , 220 , 486-492. Recuperado de:
<https://www.sciencedirect.com/science/article/abs/pii/S0045653518324639>

Anexos

1. Código LSTM

def abrir_datos():

Primero se importan todas las librerías importantes

Luego se ejecutan dos líneas que permiten acceder a las carpetas de drive y crear una variable que contenga los datos

import pandas as pd

from pandas import read_csv

from datetime import datetime

from math import sqrt

from matplotlib import pyplot as plt

from sklearn.metrics import mean_squared_error

from pandas import DataFrame

from pandas import concat

from sklearn import preprocessing

from sklearn.preprocessing import LabelEncoder

from sklearn.preprocessing import MinMaxScaler

from google.colab import drive

drive.mount('/content/gdrive')

datos = ('/content/gdrive/MyDrive/Colab Notebooks/Casino Llamacsv.csv')

return datos

##PREPARACION DE DATOS

def preparación_de_data():

import pandas as pd

from pandas import read_csv

from matplotlib import pyplot as plt

a = abrir_datos()

dataset = read_csv(a, delimiter=';', decimal=',')

dataset.drop(['Flow (l/m)', 'RHx (%)', 'BV (V)', 'Alarm'], axis=1, inplace=True) ## probe agregando la variable RHx y empeoro la predicción

```

## cambiamos nombre de columnas y el formato de dato
dataset['Time'] = pd.to_datetime(dataset['Time'])
dataset = dataset.astype({"Time": int})
dataset = dataset.rename(columns={"Time": "fecha"})
dataset = dataset.rename(columns={"WD (Deg)": "WD"})
dataset.index.name = 'fecha'
fechas = dataset['fecha']

##ahora cambiamos la variable WD a una categorica
grados = [0,11.25,33.75,56.25, 78.75,101.25, 123.75,146.25, 168.75,
          191.25, 213.75, 236.25, 258.75,281.25, 303.75, 326.25, 348.75, 360]
cardinales = ['N','NNE','NE','ENE','E','ESE','SE','SSE','S',
              'SSW','SW','WSW','W','WNW','NW','NNW','N1']
dataset['WindDir'] = pd.cut(x=dataset['WD'],bins=grados ,
                           labels=cardinales ).replace({'N1':'N'})

## ahora cambiamos la variable de MP a una categorica
aqi = [-1,25,50,75,100, 150, 1100]
calidad = ['1','2','3','4','5','6']
dataset['MP'] = pd.cut(x=dataset['Conc (UG/M3)'],bins= aqi , labels=calidad )

## ahora verificamos los datos nulos
dataset.isnull().values.any()
dataset[dataset['MP'].isnull()]
dataset = dataset.dropna(axis=0, subset=['MP'])

## esto se hace para que luego las variables se conviertan todas a float, antes tuve
problemas de formato

## definimos el estilo de los graficos y el tamaño
## eliminamos la columna WD
dataset.drop(['WD','Conc (UG/M3)'], axis=1, inplace=True)
dataset = dataset.astype({"WindDir": str}#, "MP":int})
plt.style.use('ggplot')

```

```

plt.rcParams['figure.figsize']=[12,6]
return dataset
print(dataset.head(1000))

## Ahora creamos funcion para visualizar los datos

def graficar_datos():
    from matplotlib import pyplot as plt
    a = preparación_de_data()
    values = a.values
    ## especificamos las columnas a graficar
    grupo = [1,2,3,4]
    i = 1
    ## ploteamos cada columna
    plt.figure()
    for x in grupo:
        plt.subplot(len(grupo), 1, i)
        plt.plot(values[:,x])
        plt.title(a.columns[x], y=0.5, loc='right')
        i += 1
    plt.show()

## se convierten las series a aprendizaje supervisado
## esta funcion toma una serie de tiempo multivariante y la enmarca como un conjunto de
datos de aprendizaje supervisado

def serie_supervisada(dataset, n_in =1 , n_out= 1, dropnan=True):
    from pandas import DataFrame
    from pandas import concat
    from sklearn import preprocessing
    from sklearn.preprocessing import LabelEncoder

```

```

from sklearn.preprocessing import MinMaxScaler

n_vars = 1 if type(dataset) is list else dataset.shape[1]
df = DataFrame(dataset)
columnas , nombres = list(), list()
## secuencia de entrada (t-n, ... ,t-1)
for i in range(n_in, 0 , -1):
    columnas.append(df.shift(i))
    nombres += [('var%d(t-%d)' % (j+1, i)) for j in range(n_vars)]
    ## secuencia de forecast (t, t+1,...´t+n)
for i in range(0, n_out):
    columnas.append(df.shift(-i))
    if i== 0:
        nombres += [('var%d(t)' % (j+1)) for j in range(n_vars)]
    else:
        nombres += [('var%d(t+%d)' % (j+1, i)) for j in range(n_vars)]
## ahora se junta todo con concatenar
junto = concat(columnas, axis=1)
junto.columns = nombres
if dropnan:
    junto.dropna(inplace=True) ## importante eliminar los Nan
return junto

def pre_procesamiento():
    from sklearn import preprocessing
    from sklearn.preprocessing import LabelEncoder
    from sklearn.preprocessing import MinMaxScaler
    a = preparación_de_data()
    valores = a.values
    valores = valores[:,1:7]

```

```

valores = valores[:, [-1, 0, 1, 2, 3, 4]]
codificacion = LabelEncoder()
codificacion2 = LabelEncoder()

valores[:, -1] = codificacion.fit_transform(valores[:, -1])
valores = valores.astype(float)
escalador = MinMaxScaler(feature_range=(0, 1))
escaladas = escalador.fit_transform(valores) ## no escalo ni mp ni wd
look_back = int(input("Introduce look_back:"))
valores_reframed = serie_supervisada(escaladas, look_back, 1)
valores_reframed.drop(valores_reframed.columns[[7, 8, 9, 10, 11]], axis=1, inplace=True)
valores = valores_reframed.values
return valores, escalador, codificacion2

```

```
def modelo():
```

```

    import numpy as np
    import pandas as pd
    from keras.models import Sequential
    from keras.layers import Dense
    from keras.layers import LSTM
    from keras.layers import GRU, Dropout
    from keras.metrics import MeanSquaredError, MeanAbsoluteError
    from keras.regularizers import L2
    import tensorflow as tf
    from numpy import concatenate
    from matplotlib import pyplot as plt

```

```
valores2, escalador1, codificacion1 = pre_procesamiento()
```

```
horas_de_entrenamiento = 3000
```

```

entrenamiento = valores2[:,horas_de_entrenamiento,:]
prueba = valores2[horas_de_entrenamiento,:,:]

train_x , train_y = entrenamiento[:,0:6], entrenamiento[:,6]
test_x, test_y = prueba[:,0:6], prueba[:,6]
## ahora cambiamos la forma de los datos a una 3D
train_x = train_x.reshape((train_x.shape[0], 1 , train_x.shape[1]))
test_x = test_x.reshape((test_x.shape[0], 1 , test_x.shape[1]))
print(train_y)

## MODELO LSTM
from keras.models import Sequential
from keras.layers import Dense
from keras.layers import LSTM
from keras.layers import GRU , Dropout
from keras.metrics import MeanSquaredError, MeanAbsoluteError
from keras.regularizers import L2
import tensorflow as tf
from numpy import concatenate
from matplotlib import pyplot as plt
from sklearn.model_selection import TimeSeriesSplit
from sklearn.model_selection import cross_val_score, cross_val_predict
## ahora diseñamos la red
modelo = Sequential()
modelo.add(LSTM(64, input_shape=(train_x.shape[1], train_x.shape[2]) ,
                return_sequences= True, kernel_regularizer=L2(0.01),
                recurrent_regularizer=L2(0.01), bias_regularizer=L2(0.01)))
modelo.add(Dropout(0.3))
modelo.add(LSTM(32, return_sequences=True))
modelo.add(Dropout(0.3))

```

```

modelo.add(LSTM(16, return_sequences=True))
modelo.add(Dropout(0.3))
modelo.add(Dense(10))
modelo.add(Dense(1))
modelo.compile(loss='mse', optimizer='adam', metrics=['mean_squared_error','accuracy'])
## optimizer= tf.keras.optimizers.RMSprop(0.0001), SGD, adam,
sparse_categorical_crossentropy
## loss= mse, mae

epocas = int(input("Escribe 23, corresponde al número optimo a entrenar:"))
historia = modelo.fit(train_x, train_y, epochs=epocas, batch_size=16,
                      validation_data=(test_x,test_y), verbose=2, shuffle=False)
## ahora ploteamos historia
plt.plot(historia.history['loss'], label='train')
plt.plot(historia.history['val_loss'], label='test')
plt.legend()
plt.show()
evaluando_el_modelo = modelo.evaluate(test_x,test_y)
## ahora hacemos la prediccion
n_futuro = int(input("Ingresa el paso de tiempo en el futuro:"))
pronostico = modelo.predict(test_x[-n_futuro:])
print(pronostico) ##
## ahora realizamos una transformación inversa para volver a escalar al rango original
## dado que usamos 5 variables para transformar, la inversa espera las mismas
dimensiones
valores = valores2[:,0:6] ## esto se hace debido a que valores2 es una matriz con 12
columnas
forecast_copias = np.repeat(pronostico, valores.shape[1], axis=-1)
v,b,n = forecast_copias.shape
forecast_copiass = forecast_copias.reshape((v,b*n))
prediccion_futuro = escalador1.inverse_transform(forecast_copiass)

```

```

prediccion_final = prediccion_futuro[:,0]

plt.style.use('ggplot')
plt.rcParams['figure.figsize']=[13,7]
plt.plot(test_y[-n_futuro:], label='Datos reales',color='red')
plt.plot(forecast_copiass, label='Prediccion', color = 'blue')
plt.title('Modelo LSTM')
plt.legend()
plt.show()
return prediccion_final

```

2. Código SVM

```

import pandas as pd
import sklearn
import numpy as np
from pandas import read_csv
from datetime import datetime
from math import sqrt
from matplotlib import pyplot as plt
from sklearn.metrics import mean_squared_error
from google.colab import drive
## ahora cargamos los datos
drive.mount('/content/gdrive')
datos = ('/content/gdrive/MyDrive/Colab Notebooks/460_V2.csv')
dataset = read_csv(datos, delimiter=';', decimal=',')
dataset.drop(['Flow (l/m)', 'RHx (%)', 'BV (V)', 'Alarm'], axis=1, inplace=True) ## probe
agregando la variable RHx y empeoro la prediccion
dataset.head(1000)

dataset['Time'] = pd.to_datetime(dataset['Time'])

```

```

dataset = dataset.astype({"Time": int})
dataset = dataset.rename(columns={"Time": "fecha"})
dataset = dataset.rename(columns={"WD (Deg)": "WD"})
dataset.index.name = 'fecha'
dataset.dtypes

grados = [0,11.25,33.75,56.25, 78.75,101.25, 123.75,146.25, 168.75, 191.25, 213.75,
236.25, 258.75,281.25, 303.75, 326.25, 348.75, 360]

cardinales =
['N','NNE','NE','ENE','E','ESE','SE','SSE','S','SSW','SW','WSW','W','WNW','NW','NNW','N
1']

dataset['WindDir'] = pd.cut(x=dataset['WD'],bins=grados , labels=cardinales )
aqi = [-1,25,50,75,100,150,1100]
calidad = ['1','2','3','4','5','6']
dataset['MP'] = pd.cut(x=dataset['Conc (UG/M3)'],bins= aqi , labels=calidad )
## eliminar valores nulo
dataset.isnull().values.any()
dataset[dataset['Conc (UG/M3)'].isnull()]
dataset = dataset.dropna(axis=0, subset=['Conc (UG/M3)'])
dataset.shape
## eliminamos la columna WD
dataset.drop(['WD','Conc (UG/M3)'], axis=1, inplace=True)
##,'Conc (UG/M3)'
dataset = dataset.astype({"WindDir": str}) ## esto se hace para que luego las variables se
conviertan todas a float, antes tuve problemas de formato
##,"MP":int
## definimos el estilo de los graficos y el tamaño
plt.style.use('ggplot')
plt.rcParams['figure.figsize']=[15,9]
dataset.info()

```

```

from pandas import DataFrame
from pandas import concat
from sklearn import preprocessing
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import MinMaxScaler
df = DataFrame(dataset)
valores = df.values
valores
## codificar valores categóricos
codificacion = LabelEncoder()
##valores[:,5] = codificacion.fit_transform(valores[:,5])
valores[:,2] = codificacion.fit_transform(valores[:,2])
valores = valores[:,1:8]
valores = valores[:,[-1,0,1,2,3,4]]
valores = valores.astype(float)
print(valores)

## normalizar entradas
escalador = MinMaxScaler(feature_range=(0,1))
escaladas = escalador.fit_transform(valores)

##SEPARAR EN DATOS DE ENTRENAMIENTO Y PRUEBA
horas_de_entrenamiento = 3000
entrenamiento = escaladas[:horas_de_entrenamiento,:]
prueba = escaladas[horas_de_entrenamiento:,:]
## ahora dividimos en inputs y output
train_x , train_y = entrenamiento[:,0:], entrenamiento[:,0]
test_x, test_y = prueba[:,0:], prueba[:,0]
## ahora cambiamos la forma de los datos a una 3D
train_x = train_x.reshape((train_x.shape[0], train_x.shape[1]))

```

```

test_x = test_x.reshape((test_x.shape[0] , test_x.shape[1]))
print(train_x.shape , train_y.shape , test_x.shape , test_y.shape)

## MODELOS SVM
from sklearn import svm
from sklearn.svm import SVR

modelo_rbf= svm.SVR(kernel='rbf',C=10)
modelo_polinomial = svm.SVR(kernel='poly', C=10)
modelo_lineal = svm.SVR(kernel='linear', C=10)
modelo_sigmoide = svm.SVR(kernel='sigmoid', C=10)

modelo_rbf.fit(train_x, train_y)
modelo_polinomial.fit(train_x , train_y)
modelo_lineal.fit(train_x , train_y)
modelo_sigmoide.fit(train_x , train_y)

n_futuro = int(input("Pasos de tiempo a predecir:"))
prediccion_rbf = modelo_rbf.predict(test_x[-n_futuro:])
prediccion_poly = modelo_polinomial.predict(test_x[-n_futuro:])
prediccion_lineal = modelo_lineal.predict(test_x[-n_futuro:])
prediccion_sigmoide = modelo_sigmoide.predict(test_x[-n_futuro:])

prediccion_rbf = prediccion_rbf.reshape((-1,1))
print(prediccion_rbf)
forecast_copias = np.repeat(prediccion_rbf, escaladas.shape[1], axis=-1)
print(forecast_copias)
prediccion_futuro = escalador.inverse_transform(forecast_copias)
##print(prediccion_futuro)##
prediccion_final = prediccion_futuro[:,0]

```

```
##prediccion_final = codificacion1.inverse_transform(prediccion_final) ## AQUI  
INTENTO DEVOLVER LA CODIFICACIÓN DE LAS VARIABLES CATEGÓRICAS
```

```
print(prediccion_final)
```

```
## ahora ploteamos el modelo con kernel polinomial
```

```
plt.style.use('ggplot')
```

```
plt.rcParams['figure.figsize']=[15,9]
```

```
plt.plot(test_y[-n_futuro:], label='Datos reales')
```

```
plt.plot(prediccion_lineal, label='Prediccion lineal')
```

```
plt.title('SVM lineal')
```

```
plt.legend()
```

```
plt.show()
```

```
## ahora ploteamos el modelo con kernel RBF
```

```
plt.style.use('ggplot')
```

```
plt.rcParams['figure.figsize']=[13,7]
```

```
plt.plot(test_y[-n_futuro:], label='Datos reales')
```

```
plt.plot(prediccion_rbf, label='Prediccion RBF')
```

```
plt.title('SVM RBF')
```

```
plt.legend()
```

```
plt.show()
```