Logotipo

Descripción generada automáticamente con confianza media

**Predicción de la mortalidad de salmónidos en centro de aguamar a través de un modelo de machine learning**

Proyecto para optar al Grado de

MAGISTER EN DATA SCIENCE

**Profesor Guía : Mg. Danilo Gómez Correa**

**Estudiantes : Juan Pablo Carrasco Lillo**

**Richard Sandoval Aguilar**

**Puerto Montt, Chile**

**2023**

© Juan Pablo Carrasco Lillo

© Richard Sandoval Aguilar

Queda prohibida la reproducción parcial o total de esta obra en cualquier forma, medio o procedimiento sin permiso por escrito del autor.

**HOJA DE CALIFICACIÓN**

En ………., el ……. de ………………. de ……………..., los abajo firmantes dejan constancia que el o los estudiantes xxxxx del Magíster en xxxx ha aprobado la tesis/proyecto para optar al Grado de Magister xxxxx con una nota de…..

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Nombre y firma profesor evaluador)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Nombre y firma profesor evaluador)

# DEDICATORIA

# AGRADECIMIENTOS

.

# RESUMEN

# ABSTRACT

# TABLA DE CONTENIDOS

[DEDICATORIA 4](#_Toc140757840)

[AGRADECIMIENTOS 5](#_Toc140757841)

[RESUMEN 6](#_Toc140757842)

[ABSTRACT 7](#_Toc140757843)

[TABLA DE CONTENIDOS 8](#_Toc140757844)

[INDICE DE GRÁFICOS 11](#_Toc140757845)

[ÍNDICE DE TABLAS 11](#_Toc140757846)

[ÍNDICE DE ILUSTRACIONES 12](#_Toc140757847)

[CAPÍTULO I: ANTECEDENTES DEL PROBLEMA 13](#_Toc140757848)

[1. Antecedentes de la organización. 13](#_Toc140757849)

[2. Formulación del problema 17](#_Toc140757850)

[3. Alcances 20](#_Toc140757851)

[4. Delimitaciones 20](#_Toc140757852)

[4.1 Delimitación Geográfica 20](#_Toc140757853)

[4.2 Delimitación Poblacional 20](#_Toc140757854)

[4.3 Delimitación de Tiempo 20](#_Toc140757855)

[5. Limitaciones 21](#_Toc140757856)

[6. Objetivos 21](#_Toc140757857)

[6.1 Objetivo general 21](#_Toc140757858)

[6.2 Objetivos específicos 21](#_Toc140757859)

[CAPÍTULO II: MARCO TEÓRICO 22](#_Toc140757860)

[1. Estado del Arte 22](#_Toc140757861)

[2. Análisis de supervivencia 24](#_Toc140757862)

[3. Ciencia de datos 25](#_Toc140757863)

[3.1 Big data 27](#_Toc140757864)

[3.2 Data Mining 29](#_Toc140757865)

[3.2 Data Visualization 30](#_Toc140757866)

[4. Machine Learning 31](#_Toc140757867)

[4.1 Aprendizaje supervisado 33](#_Toc140757868)

[4.2 Aprendizaje no supervisado 34](#_Toc140757869)

[4.3 Aprendizaje por refuerzo 35](#_Toc140757870)

[5. Redes Neuronales 36](#_Toc140757871)

[6. Métricas de precisión 38](#_Toc140757872)

[6.1 Error absoluto medio (MAE) 39](#_Toc140757873)

[6.2 Error cuadrático medio (MSE) 40](#_Toc140757874)

[6.3 Raíz del Error cuadrático medio (RMSE) 40](#_Toc140757875)

[6.4 Coeficiente de determinación (R2) 40](#_Toc140757876)

[CAPÍTULO III: METODOLOGÍA 42](#_Toc140757877)

[1. Comprensión del negocio 43](#_Toc140757878)

[2. Comprensión de los datos 43](#_Toc140757879)

[3. Preparación de los datos 44](#_Toc140757880)

[4. Modelamiento 44](#_Toc140757881)

[5. Evaluación 45](#_Toc140757882)

[6. Despliegue 45](#_Toc140757883)

[CAPÍTULO IV: ANÁLISIS DE RESULTADOS 46](#_Toc140757884)

[1. Comprensión del negocio 46](#_Toc140757885)

[2. Comprendiendo los datos 47](#_Toc140757886)

[2.1. Recolección de datos 47](#_Toc140757887)

[2.2. Descripción de los datos: 47](#_Toc140757888)

[2.3. Exploración de los datos 49](#_Toc140757889)

[2.4. Verificación de calidad de los datos 52](#_Toc140757890)

[3. Preparación de los datos 52](#_Toc140757891)

[4. Modelamiento 55](#_Toc140757892)

[4.1 Selección de técnicas de modelaje 55](#_Toc140757893)

[4.1.1 Regresión Lineal 55](#_Toc140757894)

[4.1.2 Random Forest 56](#_Toc140757895)

[4.1.3 XG Boost 57](#_Toc140757896)

[4.1.4 Red neuronal recurrente LSTM 59](#_Toc140757897)

[4.2 Generación diseño de prueba 62](#_Toc140757898)

[4.3 Construcción de los modelos 62](#_Toc140757899)

[4.4 Evaluación de los modelos 63](#_Toc140757900)

[5. Evaluación 64](#_Toc140757901)

[CAPÍTULO V: CONCLUSIONES Y RECOMENDACIONES 65](#_Toc140757902)

[BIBLIOGRAFÍA 66](#_Toc140757903)

[ANEXOS 70](#_Toc140757904)

# INDICE DE GRÁFICOS

[Gráfico 1. Análisis de mortalidad por tipo de mortalidad 46](#_Toc140702083)

[Gráfico 2. Distribución de variables numéricas 48](#_Toc140702084)

[Gráfico 3. Distribución de variables categóricas 48](#_Toc140702085)

[Gráfico 5. Distribución de porcentajes de mortalidad por rangos de días 49](#_Toc140702086)

[Gráfico 6. Distribución de peso promedio por rango de días 50](#_Toc140702087)

[Gráfico 7. Dispersión entre porcentaje de mortalidad y el peso 50](#_Toc140702088)

[Gráfico 8. Mortalidad infecciosa y total 51](#_Toc140702089)

[Gráfico 9. Escala de correlaciones de variables del modelo 51](#_Toc140702090)

[Gráfico 10. Escala de correlaciones sin variables altamente correlacionadas 53](#_Toc140702091)

[Gráfico 11. Análisis días de vida vs mortalidad registrada 54](#_Toc140702092)

[Gráfico 12. Distribución de temperaturas y oxigeno por estación 54](#_Toc140702093)

[Gráfico 13. Comparativa de Métricas de precisión por modelo 65](#_Toc140702094)

# ÍNDICE DE TABLAS

Tabla 1. Estructura societaria de la compañía 12

Tabla 2. Regiones de Chile con operaciones de la empresa 13

Tabla 3. Reglas de decisión sobre coeficiente de determinación r2 41

Tabla 4. Variables del modelo 47

Tabla 5. Tamaño del set de entrenamiento y prueba 62

# ÍNDICE DE ILUSTRACIONES

[Ilustración 1. Dotación de colaboradores por área de negocio (2022) 16](#_Toc140701479)

[Ilustración 2. Ciclo de vida de los datos 27](#_Toc140701480)

[Ilustración 3. Tipos de problemas a resolver desde el Machine Learning 32](#_Toc140701481)

[Ilustración 4. Algoritmos de machine learning según tipo de problema 32](#_Toc140701482)

[Ilustración 5. Diagrama de aprendizaje supervisado 34](#_Toc140701483)

[Ilustración 6. Diagrama de aprendizaje no supervisado 35](#_Toc140701484)

[Ilustración 7. Diagrama de aprendizaje por refuerzo 36](#_Toc140701485)

[Ilustración 8. Red neuronal 37](#_Toc140701486)

[Ilustración 9. Ejemplos de funciones de activación 38](#_Toc140701487)

[Ilustración 10. Bondad de ajuste en diagrama 41](#_Toc140701488)

[Ilustración 11. Ciclo de la metodología CRISP-DM 43](#_Toc140701489)

[Ilustración 12. Técnica del Random Forest 56](#_Toc140701490)

[Ilustración 13. Representación del algoritmo XGBoost 58](#_Toc140701491)

[Ilustración 14. Arquitectura de una celda de memoria 60](#_Toc140701492)

[Ilustración 15. Arquitectura de RNN Long short-term memory 61](#_Toc140701493)

[Ilustración 16. Algoritmo Regresión Lineal 62](#_Toc140701494)

[Ilustración 17. Algoritmo Random Forest 62](#_Toc140701495)

[Ilustración 18. Algoritmo XGBoost 63](#_Toc140701496)

[Ilustración 19. Algoritmo Red Neuronal Recurrente LSTM 63](#_Toc140701497)

[Ilustración 20. R2 comparado por modelo 63](#_Toc140701498)

[Ilustración 21. Algoritmo Random Forest con hiperparámetros optimizados 63](#_Toc140701499)

[Ilustración 22. Algoritmo XGBoost con hiperparámetros optimizados 64](#_Toc140701500)

[Ilustración 23. Algoritmo red neuronal recurrente LSTM con hiperparámetros optimizados 64](#_Toc140701501)

# CAPÍTULO I: ANTECEDENTES DEL PROBLEMA

## Antecedentes de la organización.

La Memoria Anual de Empresas AquaChile S.A (2017) señala que la empresa nace en 1998 de la fusión entre Salmones Pacífico Sur S.A. (constituida el año 1979 en la ciudad de Coyhaique), y AquaChile S.A. (constituida el año 1988 en la ciudad de Santiago).

En sus inicios, cada una de estas empresas se orientó a distintas etapas del proceso de cultivo del Salmón. Salmones Pacífico Sur S.A. estaba enfocada en el cultivo en agua de mar, mientras que AquaChile S.A. estaba concentrada en la etapa de agua dulce, con la producción de ovas, alevines y smolts.

Las dos compañías eran totalmente complementarias, con fuertes posiciones estratégicas dentro de sus líneas de negocio, con una organización bastante desarrollada, un extenso conocimiento del negocio y accionistas conscientes de la oportunidad única de lograr en conjunto un crecimiento más rentable y sólido.

Desde ese momento empresas AquaChile S.A. se ha consolidado como la empresa salmonicultura más grande de Chile y el mundo. Actualmente es considerada la segunda más grande del mundo después de la fusión de las empresas noruegas SalMar y NRS (Soltveit, 2022).

En junio de 2018, la empresa Agrosuper anuncia la compra de los activos del área Salmones de la empresa Friosur S.A. perteneciendo al grupo AquaChile. Durante agosto del mismo año, Agrosuper anuncia la compra de las Empresas AquaChile.

Un conjunto de letras blancas en un fondo blanco

Descripción generada automáticamente con confianza baja

Tabla 1. Estructura societaria de la compañía

Fuente: Reporte integrado AquaChile S.A (2022)

Actualmente la empresa tiene operaciones en cuatro regiones de Chile, las cuales se presentan en la siguiente tabla. Además, la empresa cuenta con oficinas comerciales en 9 países; Estados Unidos, Brasil, España, Francia, Suiza, Alemania, Kazajistán, Japón y China.

|  |  |
| --- | --- |
| **Región** | **Operaciones** |
| **Metropolitana** | 1 tienda comercial |
| **Araucanía** | 4 pisciculturas  1 centro reproductores |
| **Los Lagos** | 1 oficina central  1 tienda comercial  1 planta de alimentos  5 plantas de procesos  6 pisciculturas  26 centros de cultivo  1 centro de innovación  1 laboratorio |
| **Magallanes** | 1 tienda comercial  1 planta de procesos  1 piscicultura  11 centros de cultivo |

Tabla 2. Regiones de Chile con operaciones de la empresa

Fuente: Reporte integrado AquaChile S.A (2022)

La empresa ha definido un modelo de negocios en el cual se ha integrado verticalmente en su totalidad, destacando todo el proceso de crianza del salmón, hasta su comercialización.

La compañía ha definido una serie de objetivos estratégicos tales como mantener el modelo de negocio, a través del crecimiento de la producción y procesamiento local e internacional, y fortalecer la red de distribución y el potencial de marcas a través de la comercialización de productos en todos los mercados, garantizando los más altos estándares de bienestar animal, visibilizando y construyendo relaciones de largo plazo con los grupos de interés, y buscando la adaptación a las operaciones del futuro.

La Compañía ha definido solo un segmento de operación (salmón), el cual se definió en base a su estructura organizacional y a la naturaleza del negocio. Este incorpora la producción, procesamiento, distribución y comercialización de productos en base a salmón, de las especies atlántico y del pacífico tanto en el mercado nacional como el de exportación. Lo anterior se realiza bajo las marcas Aqua y Verlasso.

Este proceso de negocios se encuentra integrado verticalmente, desde la producción de ovas hasta la comercialización de sus productos en el mercado nacional e internacional, a través de la extensa cadena de distribución con que cuenta AquaChile.

En relación con el cultivo del salmón, la compañía mantiene su propio programa de reproducción, autoabasteciéndose de ovas y smolts, lo que le permite contar con un amplio margen de seguridad en esta etapa clave de la cadena productiva. Posterior a esto, se inicia el proceso de engorda, utilizándose sólo concesiones y centros de engorda propios. Para el procesamiento, además de un acopio en tierra, contamos con plantas de avanzada tecnología, donde se procesa el 100% de la producción, elaborando productos de distintas características, dependiendo del mercado y/o cliente objetivo a satisfacer. Lo anterior nos permite obtener una trazabilidad completa y estándares de calidad y bioseguridad superiores a la industria.

En este contexto, el estatus sanitario de todos los procesos, la excelencia ambiental de su entorno, la bioseguridad de sus operaciones y la inocuidad de sus productos, son intensamente controlados en todas las etapas. Para esto, la Compañía cuenta con múltiples certificaciones: ISO 9001 y HACCP 14 (calidad de sus productos), ISO 14001 (gestión ambiental), OHSAS 18001 (estándar de seguridad ocupacional), IFS (International Food Standard) nivel v5, BRC (British Retail Consortium) v5, Global GAP y BAP (buenas prácticas en la agricultura y acuicultura) y certificación Kosher. Todas estas certificaciones nos permiten llegar a los mercados más exigentes del mundo, como es el caso del mercado norteamericano y el asiático.

En materia de sustentabilidad, se trabaja en base a los pilares ambiental, social y económico, los cuales son un factor determinante para nuestro quehacer diario. Para ello, hemos desarrollado un modelo de producción que incluye una gestión ambiental responsable, a través de la incorporación de tecnología y una permanente innovación en cada uno de nuestros procesos. De esta manera, apuntamos a mejorar continuamente, optimizando cada etapa de nuestra cadena de valor y haciendo un uso eficiente de los recursos naturales. Para lo anterior, contamos con un Comité de Sustentabilidad que sesiona una vez al mes con el objetivo de dar seguimiento y velar por el cumplimiento de los indicadores sociales y medioambientales.

A diciembre de 2022, la compañía declara en su Informe de Gestión Integrado (AquaChile, 2022) que cuenta con 5.827 colaboradores, donde el 93,7% son chilenos y el 39% son mujeres.

En la ilustración, se presenta la dotación de colaboradores por área de negocio.

Imagen de la pantalla de un celular con texto e imagen

Descripción generada automáticamente con confianza media

Ilustración 1. Dotación de colaboradores por área de negocio (2022)

Fuente: Reporte Integrado AquaChile 2022

Como dato relevante, la compañía tiene el 75,4% de su dotación contratada prestando servicios en la región de Los Lagos, que es el área de influencia del estudio de este proyecto.

Misión

Una producción acuícola saludable que alimente a esta y las generaciones futuras, respetando y valorando el medio ambiente y a nuestra gente de manera sostenible.

Visión

Ser el mejor productor de salmón del mundo

Valores corporativos

* Ponemos alma y pasión en el trabajo
* Compartimos y multiplicamos nuestras capacidades
* Tenemos un claro sentido único
* Nuestra palabra tiene el valor de un contrato
* Sencillez y austeridad
* Espíritu emprendedor
* Responsabilidad y disciplina
* Espíritu de compañerismo y cooperación en nuestra empresa
* Confiamos en nuestro equipo y su compromiso
* Creemos en el liderazgo a través del ejemplo

## Formulación del problema

La industria de la salmonicultura es uno de los sectores más importantes en la producción de alimentos de Chile y se posiciona como una industria relevante en cuanto a la provisión de proteínas a nivel mundial. Sin embargo, como en cualquier actividad agrícola o pecuaria, enfrenta desafíos relacionados con la salud y el bienestar de los animales que se crían, en particular los salmones, por ello uno de los aspectos cruciales en la salmonicultura es el control y la predicción del porcentaje de mortalidad en las poblaciones de salmones.

La predicción del porcentaje de mortalidad se ha convertido en una herramienta fundamental, ya que le permite a la empresa tomar decisiones informadas y proactivas para prevenir y manejar posibles brotes de enfermedades y otros factores que podrían causar una alta tasa de mortalidad en los salmones. La capacidad de anticipar y predecir estos eventos es esencial para garantizar la sostenibilidad y rentabilidad de la industria.

Existen varios factores que pueden influir en la tasa de mortalidad de los salmones, como enfermedades infecciosas, cambios ambientales, calidad del agua, nutrición deficiente y estrés. La detección temprana de estos factores y la predicción de su impacto en la población de peces pueden ayudar a implementar medidas preventivas y de control de manera oportuna, minimizando así las pérdidas económicas y ambientales.

El monitoreo regular de los indicadores de salud y el seguimiento de las condiciones ambientales permiten a los acuicultores implementar medidas de manejo adecuadas, como ajustar la alimentación, mejorar la calidad del agua o administrar tratamientos veterinarios específicos. Esto ayuda a minimizar el sufrimiento de los peces y mantener su salud óptima, promoviendo así una producción sostenible y ética.

En la etapa de producción de las especies salmónidas se presentan eventos de mermas de la población, como lo son la mortalidad por diferentes motivos, manejos que realiza el acuicultor, ataque de predadores y enfermedades. De estas últimas las que ocupan y preocupan mayormente al departamento de salud y producción de la compañía son las de carácter infeccioso. Estas deben ser tratadas oportunamente, idealmente anticipar las acciones, evitando su propagación y el daño económico / ambiental que pudiese ocurrir.

Para ello es necesario proyectar o calcular la esperanza de vida que tienen estos peces, ya sea en número o en porcentaje esperado a un momento determinado, esto permitirá por un lado determinar la perdida futura en cuanto a unidades, o evaluar dependiendo de la etapa de vida o cumplimiento del peso objetivo de los peces involucrados, tomar la decisión de ser enviados a proceso de planta como cosecha.

La información analizada específicamente es cuando los peces se encuentran en la etapa de engorda, agua mar, dado que en esta etapa impacta significativamente en el costo total de producción.

Otra arista para considerar en estos días son las multas por sobreproducción, con la ayuda de modelos de modelos de predicción de porcentaje de mortalidad se puede estimar de forma óptima las poblaciones a fin de ciclo productivo.

Sin embargo, en términos prácticos los modelos conocidos para estimar la supervivencia de individuos, por ejemplo, los análisis de supervivencia, que se basan en que los individuos están vivos o no en un determinado momento, no permiten su aplicación directa cuando se trata con poblaciones, donde alguna porción de la población podría no sobrevivir, pero sigue existiendo una cantidad determinada de individuos vivos. En este escenario, es importante comprender que lo que se busca establecer es un modelo que permita anticipar oportunamente la mortalidad de un centro de cultivo, o en una jaula específica, y desde ahí poder conocer las variables productivas y/o ambientales que puedan explicar esa mortalidad.

En resumen, el problema identificado se busca resolver a través de la predicción del porcentaje de mortalidad. Para la industria es de vital importancia, ya que permite tomar decisiones estratégicas basadas en datos y anticiparse a posibles problemas que podrían afectar la salud y el bienestar de los salmones. Así mismo, permitirá tomar decisiones productivas relevantes en cuanto a tiempos de cosecha, programación de transportes, alimentación, planificación de planta de procesos. En cuanto a la sustentabilidad, permitirá pronosticar la cantidad de biomasa en el centro, que está directamente relacionado con el cumplimiento de la normativa medioambiental. Y desde la perspectiva comercial, permitirá a la empresa disponer de información respecto a que tipo de cosecha se tendrá, lo cual es un insumo para el cumplimiento de los forecast comerciales. Finalmente, al prevenir la mortalidad y mejorar las condiciones de crianza, se promueve la sostenibilidad económica y ambiental de la industria, así como el cumplimiento de estándares éticos y de bienestar animal.

## Alcances

El proyecto contempla diseñar y testear la metodología CRISP-DM para pronosticar la mortalidad de salmónidos en centros de cultivo de aguamar de la empresa AquaChile S.A. El tipo de alcance a utilizar será del tipo exploratorio y correlacionales, ya que la investigación se realizará desde una perspectiva innovadora que nos permitirá generar nuevos modelos de estimación predictiva a partir de una serie de datos obtenidos de los sistemas de la compañía.

Para ellos, se utilizará lenguaje de programación Python, donde se realizarán pruebas de algoritmos de machine learning y una arquitectura de redes neuronales, que permita finalmente comparar la potencia y eficiencia de cada una, eligiendo pues el mejor indicador entre las técnicas desarrolladas.

## Delimitaciones

### 4.1 Delimitación Geográfica

Los datos con lo que se trabajará en el proyecto representan la dinámica productiva del barrio 34, con una superficie de 812km2 ubicado en la región de Aysén, donde actualmente la empresa tiene 10 centros de cultivo.

### 4.2 Delimitación Poblacional

La población objetivo del proyecto corresponde a la alta gerencia de la compañía en cuanto al proceso de toma de decisiones estratégicas. Desde una perspectiva más de terreno, se considera relevante a los jefes de centro, veterinarios, jefes de planta de proceso que están vinculados directamente al proceso productivo estudiado.

### 4.3 Delimitación de Tiempo

Se utilizarán los datos diarios obtenidos de los sistemas de la compañía, desde 1/1/2020 hasta el 1/5/2023, esto representa dos ciclos de producción. La duración del proyecto contempla una duración de 6 meses desde la obtención de la información.

## Limitaciones

Los datos se procesan con equipos computacionales personales, sin considerar una inversión en recursos tecnológicos para el desarrollo de los modelos.

Restricción de exponer datos reales de la compañía debido a alta relevancia en el desarrollo competitivo de la empresa.

## Objetivos

### 6.1 Objetivo general

Predecir el porcentaje de mortalidad de salmónidos en centro de aguamar a través de un modelo de machine learning

### 6.2 Objetivos específicos

O.E.1 Evaluar y determinar la cantidad de datos necesarios para la formulación del modelo.

O.E.2 Determinar y medir las causas de mortalidad y supervivencia en el salmónidos en etapa de crianza.

O.E.3 Implementar el modelo predictivo a través de algoritmos de Machine Learning.

O.E.4 Testear y entrenar el modelo con datos reales obtenidos desde sistema de producción de la empresa.

O.E.5 Evaluar métricas de rendimiento de los modelos de machine learning implementados.

# CAPÍTULO II: MARCO TEÓRICO

## Estado del Arte

Los modelos de predicción son herramientas estadísticas utilizadas para predecir la relación entre una variable dependiente y una o múltiples variables independientes. A través de estos modelos se busca predecir, en base a una relación estadísticamente significativa, si la o las variables explicativas, normalmente denota con una , son capaces de explicar el comportamiento de la variable explicada, denotada como

Conceptualmente, la regresión es la estimación de la media condicional de respecto de , lo que simbólicamente se representa de la siguiente manera:

(1)

Donde denota una función de la variable explicativa .

A la ecuación (1) se le denomina función de esperanza condicional o función de regresión poblacional.

Si bien es cierto actualmente es uno de los modelos predictivos más utilizados y de él han derivado una gran cantidad de modelos, la historia de los modelos de regresión no es nueva.

La historia de los modelos de predicción se remonta a las antiguas civilizaciones orientales, donde ya los babilonios contaban con un método matemático para determinar la variación de la duración del día durante un año solar. El caso más emblemático es el de la *Tablilla de Venus Ammisaduqa*, actualmente en el museo británico de Londres, que es un texto del siglo VII a.C. donde se encuentran registradas observaciones astronómicas del planeta Venus, incluidas ciertas “predicciones” respecto de su órbita. Si bien es cierto, no hay acuerdo sobre la datación exacta de la tablilla, escrita en lenguaje cuneiforme, los científicos si concuerdan en su contenido.

Durante la edad antigua y media, los modelos predictivos eran más especulativos que científicos, sin embargo, fue hasta la edad Moderna donde se expandió a todos los campos del conocimiento, que se tomó el real interés por realizar predicciones más precisas. Uno de los primeros modelos de predicción conocidos en la historia moderna fue el modelo Malthusiano, diseñado por el economista británico Thomas Malthus (1766-1834) quién en su publicación *Ensayo sobre el principio de la población (1798)* postula que la población tiene la capacidad de doblarse cada 25 años, creciendo de periodo en periodo, en una progresión geométrica, sin embargo, para que esto fuera posible era necesario que existiera una gran cantidad de alimentos. Es decir, la población podría doblarse en un cuarto de siglo siempre y cuando se cumpla la condición que exista abundancia de alimentos, que, de acuerdo con su estudio, evoluciona en una progresión aritmética. El crecimiento poblacional dependía de la oferta de alimentos.

Durante el siglo XIX no hubo grandes aportes a los modelos de regresión, desde una perspectiva matemática, sin embargo, el modelo de Darwin a través del cual estudió y predijo la evolución de las especies se basa en la selección natural y la competencia por los recursos, del cual desprendió la conclusión que la evolución de las especies dependía de ciertas variables observables y que con el pasar de los años, incluso se podrían controlar.

Quizá el aporte más significativo en esta materia, y que abrió el campo de estudio de la regresión y la correlación fue el que realizó el matemático, antropólogo, genetista, estadístico y fisiólogo británico Francis Galton (1822-1911) quien en su obra *Natural Inheritance* (1889) planteó que, a pesar de la tendencia de los padres de estatura alta a procrear hijos altos y los padres de estatura baja, hijos bajos, la estatura promedio de los niños de padres de una estatura determinada tendía a desplazarse, o “regresar”, a la estatura promedio de la población total. *La Ley de la regresión universal* o *“regresión a la mediocridad”* como la llamó, establecía que la estatura de los hijos de padres inusualmente altos o bajos tiende a dirigirse a la estatura promedio de la población.

## Análisis de supervivencia

Los análisis de supervivencia, que corresponde a un modelo de regresión, realiza un análisis del tiempo de seguimiento de cada unidad o individuo observado hasta que ocurre un fenómeno predefinido y que es, normalmente, la razón del estudio (muerte por cáncer de colon, recaída después de un tratamiento contra la heroína, no pago de un crédito bancario, etc.), es decir, realiza una predicción de tiempos de duración, que es la variable de interés, de una determinada situación, que termina con la ocurrencia de un evento.

Durante el siglo XVIII los avances en esta disciplina estuvieron marcados por los estudios sobre la inoculación de la viruela de Daniel Bernoulli (1766) que dieron la base a la teoría de los riesgos competitivos, y su presentación de la primera doble tabla de vida decreciente. Estos estudios condujeron a establecer cuantas muertes habría habido en una población de estudio, si se aplican un conjunto determinado de tasas de muerte o bajo la presencia de un conjunto de diversas causas de muerte.

A través del siglo XIX fueron múltiples los avances metodológicos realizados, además de los modelos paramétricos de análisis de supervivencia recién mencionados, se agregaron las primeras manipulaciones gráficas simultaneas del tiempo calendario y la edad a través de los Diagramas de Lexis (1875).

Los primeros años del siglo XX están caracterizados por un escepticismo de los estadísticos de la época sobre el valor real de estas tablas de vida en la investigación estadística. De acuerdo con Matthews (2006), los estadísticos Greenwood (1922) y Westergaard (1925) expresan sus dudas respecto a la efectividad de las tablas de vida, argumentando que no había referencia sobre la variabilidad muestral o error estándar. Estas publicaciones generan gran interés en la estadística aplicada a las problemáticas médicas de la época, y dan pie a numerosos trabajos como los de Boag (1949) que incorporó el enfoque de máxima verosimilitud, el cálculo de la tasa de sobrevida de pacientes con cáncer de Berkson y Gage (1950), los estimadores de los tiempos de muertes y censura en grandes intervalos irregulares de tiempo de Harris et al. (1950), los riesgos competitivos estudiados por Fijar y Neyman (1951), el uso de la distribución exponencial por Littell (1952) y el estudio de máxima utilización de método de la tabla de vida en el análisis de la supervivencia por Cutler y Ederer (1958).

Sin embargo, los máximos aportes en esta disciplina fueron realizados por el matemático Edward Kaplan (1920 – 2006) y el estadístico Paul Meier (1924 – 2011) que, en su publicación de 1958 *“Nonparametric Estimation from Incomplete Observations”* incluyeron una metodología para tratar los datos censurados, lo que era hasta ese momento un problema al momento de realizar los análisis de supervivencia. El estimador de la función de supervivencia o estimador Límite-producto, se calcula como el producto de probabilidades condicionadas. Así mismo, en 1972 el estadístico británico David Cox (1924 – 2022) complementa el trabajo de Kaplan y Meier a través de su modelo de regresión semi paramétrico para la función de riesgo, denominada modelo de riesgos proporcionales de Cox (1972). Este supone que existe un conjunto de covariables que influyen sobre el comportamiento del tiempo de ocurrencia de los eventos.

Este marco referencial, nos permite tener una idea general de la historia y trascendencia que los análisis de regresión han tenido a lo largo de los años y como se han ido incorporando aportes que permiten mejorar la eficiencia de estos modelos y, por ende, obtener conclusiones más significativas en cuanto a la predicción y/o estimación de los tiempos trascurridos hasta el evento de estudio.

## Ciencia de datos

Los orígenes de esta disciplina se remontan a 1962 cuando el estadístico estadounidense John W. Tukey (1915- 2000), entre otros aportes, creador de los diagramas de caja o *bloxplot*, discutía el futuro de la estadística matemática como ciencia empírica. En su publicación "El futuro del análisis de datos" escribe: (…)*“durante mucho tiempo pensé que era un estadístico, interesado en inferencias de lo particular a lo general. Pero a medida que he visto evolucionar la estadística matemática, he tenido motivos para preguntarme y dudar. He llegado a sentir que mi interés central está en el análisis de datos. El análisis de datos, y las partes de la estadística que se adhieren a él, deben adoptar las características de la ciencia más que las de las matemáticas, el análisis de datos es intrínsecamente una ciencia empírica. ¿Cuán vital e importante es el auge del ordenador electrónico de programa almacenado? En muchos casos la respuesta puede sorprender a muchos siendo 'importante pero no vital', aunque en otros no cabe duda de que el ordenador ha sido 'vital'".*

Luego en 1974 el científico y pionero de la informática Peter Naur (1928 – 2016) puso énfasis en los datos y acuñó el término “Ciencia de Datos” en su publicación *Concise Survey of Computer Methods*, definiéndola como *"La ciencia de tratar con datos, una vez que se han establecido, mientras que la relación de los datos con lo que representan se delega en otros campos y ciencia”.* Es la primera publicación conocida donde se habla abiertamente de la ciencia de datos. Antes, en 1968, Naur ya había hecho referencias a la ciencia de datos, a través de un plan de curso en el Congreso de la IFIP (International Federation for Information Processing) llamado: *“”Datalogy”, the science of data and of data processes and its place in education”.*

A partir de 1977 el término fue integrado en varias asociaciones y conferencias de ámbito estadístico y computacional. Ese mismo año se crea la Asociación Internacional de Computación Estadística (IASC) como Sección del ISI. *"La misión del IASC es vincular la metodología estadística tradicional, la tecnología informática moderna y los conocimientos de los expertos en la materia para convertir los datos en información y conocimiento."*

De acuerdo con la definición de Amazon AWS, *La ciencia de datos es el estudio de datos con el fin de extraer información significativa para empresas. Es un enfoque multidisciplinario que combina principios y prácticas del campo de las matemáticas, la estadística, la inteligencia artificial y la ingeniería de computación para analizar grandes cantidades de datos. Este análisis permite que los científicos de datos planteen y respondan a preguntas como “qué pasó”, “por qué pasó”, “qué pasará” y “qué se puede hacer con los resultados”.*

En términos simples, “*la ciencia de datos es la disciplina que convierte los datos en conocimiento útil”,* es decir, domina el espectro completo del ciclo de vida de los datos *(Ver: ilustración 1)*. Para ello se sirve de las matemáticas y estadísticas, Machine Learning y la computación.

*Diagrama

Descripción generada automáticamente*

Ilustración 2. Ciclo de vida de los datos

Fuente: R for Data Science (2017)

De esta definición se desprende que la ciencia de datos se compone de tres grandes áreas; el *big data*, el *data mining* y la *data visualization*.

### 3.1 Big data

En cuanto al *big data*, García et al. (2018) comenta “*Muchas son las definiciones que entidades y organizaciones han dado para el término big data, pero todas ellas se pueden resumir en el conjunto de datos cuyo tamaño supera considerablemente la capacidad de captura, almacenado, gestión y análisis del software convencional de bases de datos”.*

La empresa Gartner (2001) define el *big data* como “(…) *datos que contienen una mayor variedad y que se presentan en volúmenes crecientes y a una velocidad superior”*. De esta definición se desprenden las tres características fundamentales del big data; Volumen, Variedad, Velocidad. Que son denominadas como “las 3 V”. A continuación, una breve descripción de cada una:

* Volumen
  + La cantidad de datos importa. Con Big Data, tendrá que procesar grandes volúmenes de datos no estructurados de baja densidad. Puede tratarse de datos de valor desconocido o equipo con sensores. En algunos casos esto puede suponer decenas de terabytes de datos o incluso cientos de petabytes.
* Velocidad
  + La velocidad es el ritmo al que se reciben los datos y (posiblemente) al que se aplica alguna acción. La mayor velocidad de los datos normalmente se transmite directamente a la memoria, en vez de escribirse en un disco. Algunos productos inteligentes habilitados para Internet funcionan en tiempo real o prácticamente en tiempo real y requieren una evaluación y actuación en tiempo real.
* Variedad
  + La variedad hace referencia a los diversos tipos de datos disponibles. Los tipos de datos convencionales son estructurados y pueden organizarse claramente en una base de datos relacional. Con el auge del big data, los datos se presentan en nuevos tipos de datos no estructurados. Los tipos de datos no estructurados y semiestructurados, como el texto, audio o vídeo, requieren un preprocesamiento adicional para poder obtener significado y habilitar los metadatos.

Desde esa definición, empresas como IBM y Oracle han incorporado otras V, siendo las más relevantes:

* Veracidad
  + Tanto en entornos de datos masivos como en entornos tradicionales, resulta crítico que los datos que estamos manejando sean veraces pues, de otra forma, se perdería toda la credibilidad en el resultado.
* Valor
  + El resultado del análisis debe aportar valor a la empresa o la organización.

### 3.2 Data Mining

Un segundo concepto asociado a la definición de la Ciencia de Datos es el Data Mining, o minería de datos.

Según Microsoft (2022), la minería de datos es el proceso de detectar la información procesable de los conjuntos de grandes de datos. Utiliza el análisis matemático para deducir los patrones y tendencias que existen en los datos. Normalmente, estos patrones no se pueden detectar mediante la exploración tradicional de los datos porque las relaciones son demasiado complejas o porque hay demasiado datos.

Estos patrones y tendencias se pueden recopilar y definir como un modelo de minería de datos. Los modelos de minería de datos se pueden aplicar en escenarios como los siguientes:

* Pronóstico
  + Cálculo de las ventas y predicción de las cargas del servidor o del tiempo de inactividad del servidor.
* Riesgo y probabilidad
  + Elección de los mejores clientes para la distribución de correo directo, determinación del punto de equilibrio probable para los escenarios de riesgo, y asignación de probabilidades a diagnósticos y otros resultados.
* Recomendaciones
  + Determinación de los productos que se pueden vender juntos y generación de recomendaciones.
* Búsqueda de secuencias
  + Análisis de los artículos que los clientes han introducido en el carrito de la compra y predicción de posibles eventos.
* Agrupación
  + Distribución de clientes o eventos en grupos de elementos relacionados, y análisis y predicción de afinidades.

### 3.2 Data Visualization

Finalmente, la visualización de datos, o Data Visualization, se hace fundamental cuando la generación y el almacenamiento de grandes volúmenes de información hacen que el mismo pase desapercibido y muchas veces se pierde la oportunidad de encontrar valor en ella.

La visualización de datos es el proceso de representación de datos, en formato gráfico, de una manera clara y eficaz. El periodista y diseñador, David McCandless en una conferencia para TED Talks en 2010 presenta su libro *“The beauty of data visualization”* donde expone que la problemática sobre la sobrecarga de información o exceso de datos que se sufre hoy en día es acarreada por el uso de la tecnología web. Afirma que existe una solución fácil que radica en la visualización de la información. Por medio de ésta se pueden ver patrones y conexiones, de modo que la información tiene más sentido y permite centrarse únicamente en lo que es importante.

Por su parte, los autores del libro “*Data Visualization for dummies”* (2014) Yuk y Diamond presentan que la visualización de datos es el estudio de cómo representar los datos usando un enfoque visual o artístico en lugar del método tradicional de reportes. Estos autores consideran que ser útil, usable y deseable son características recomendables de toda visualización, y a ellas les agregan:

* ser visualmente atractivas,
* escalables,
* proveer al usuario de información correcta y,
* ser accesibles en cualquier momento.

Finalmente, podemos señalar que la ciencia de datos nos permite expandir nuestras capacidades de cálculo, lógicas, relacionales, de abstracción y visualización en cualquier área del conocimiento. Esta relación entre los grandes volúmenes de datos disponibles y almacenados, sean estructurados o no, con la capacidad de entenderlos y generar conclusiones, a través de algoritmos procesados a la velocidad de la luz, junto a la capacidad de generar representaciones gráficas y visualizaciones que permitan entender la totalidad y las partes, de los datos procesados, consigue generar un nivel de comprensión de nuestro entorno y sus componentes, que sin duda, permiten generar mejores soluciones a problemas más complejos.

## Machine Learning

El estudio de algoritmos de computación que mejoran automáticamente su rendimiento gracias a la experiencia. Se dice que un programa informático aprende sobre un conjunto de tareas, gracias a la experiencia y usando una medida de rendimiento, si su desempeño en estas tareas mejora con la experiencia. Así define Tom Mitchell (1986) en su libro *Machine Learning: A Guide to Current Research* a esta disciplina que nace en 1950 con la creación del Test de Turing.

Según Arthur Samuel (1959), creador del primero algoritmo que jugaba damas, capaz de aprender de su propio juego, *“el machine learning es un campo de estudio que brinda a las computadoras la capacidad de aprender sin ser programadas explícitamente”*.

El machine learning es una rama dentro del campo de la Inteligencia Artificial que proporciona a los sistemas la capacidad de aprender y mejorar de manera automática, a partir de la experiencia. Estos sistemas transforman los datos en información, y con esta información pueden tomar decisiones. Es clave entonces, para que un modelo realice predicciones de manera robusta, necesita alimentarse de datos.

El Machine Learning tiene una amplia gama de aplicaciones, incluyendo motores de búsqueda, diagnósticos médicos, detección de fraude en el uso de tarjetas de crédito, análisis del mercado de valores, clasificación de secuencias de ADN, reconocimiento del habla y del lenguaje escrito, juegos y robótica. Ahora, cada uno de ellos representa un tipo diferente de problema a resolver por la máquina, de los cuales se categorizan como:

* Problemas de clasificación
* Problemas de clustering
* Problemas de regresión

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Ilustración 3. Tipos de problemas a resolver desde el Machine Learning

Fuente: página web oficial Scikit Learn [www.scikit-learn.org](http://www.scikit-learn.org)

Dependiente de la forma en la cual se busque resolver el problema es que se deberá escoger una estrategia de aprendizaje. Hoy día existe una gran cantidad de algoritmos prefijados como por ejemplo, los de Scikit-learn, que permiten resolver estos problemas, según un aprendizaje supervisado, no supervisado y de refuerzo.

Diagrama

Descripción generada automáticamente

Ilustración 4. Algoritmos de machine learning según tipo de problema

Fuente: página web oficial Scikit Learn [www.scikit-learn.org](http://www.scikit-learn.org)

### 4.1 Aprendizaje supervisado

En los problemas de aprendizaje supervisado se entrena al algoritmo a partir de datos que ya vienen etiquetados con la respuesta correcta.

Cuando el algoritmo aprende con el set de datos de entrenamiento, se prueba a través de otro set de prueba sin las etiquetas de las respuestas correctas, y el algoritmo, de acuerdo con su entrenamiento, busca predecir un resultado. Esto es similar al método de aprendizaje que se utiliza en las escuelas, donde se nos enseñan problemas y las formas de resolverlos, para que luego podamos aplicar los mismos métodos en situaciones similares.

Los algoritmos más comúnmente utilizados son :

* Regresión Lineal (Galton F.(1889) *“Natural Inheritance”*)
* Regresión Logística (Berkson (1944). “*Application of the logistic function to bioassay*”).
* Arboles de decisión (Quinlan (1986) *“Induction of Decision Trees”*)
* Random forest (Breiman (2001). “*Random Forests”*)
* K-nearest neighbor (Fix E, Hodges J (1951) *“Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties”.)*
* Redes neuronales (McCulloch W. Pitts W. (1943). *“A Logical calculus of the ideas immanent in nervous activity”*)
* Máquinas de soporte vectorial (Vapnik V., Chervonenkis A. (1964) *“The Method of Support Vectors”*)
* Clasificador Bayesiano (Duda R., Hart P. (1973). *"Pattern Classification and Scene Analysis"*)

Diagrama

Descripción generada automáticamente

Ilustración 5. Diagrama de aprendizaje supervisado

Fuente: Luna J. (2018)Tipos de Aprendizaje Automático. Portal Medium

### 4.2 Aprendizaje no supervisado

Los métodos no supervisados son algoritmos que basan su proceso de entrenamiento en set de datos sin etiquetas definidas. El aprendizaje no supervisado estudia los datos para identificar patrones, es el mismo algoritmo que determina las correlaciones y las relaciones mediante su interacción con los datos, finalmente describiendo su estructura.

Este tipo de aprendizaje está dedicado principalmente a las tareas de agrupamiento, conocidas como clustering o segmentación, donde su objetivo es encontrar grupos similares en el conjunto de datos.

Diagrama

Descripción generada automáticamente

Ilustración 6. Diagrama de aprendizaje no supervisado

Fuente: Luna J. (2018)Tipos de Aprendizaje Automático. Portal Medium

Entre los principales algoritmos de aprendizaje no supervisado destacan:

* K-means (Lloyd S. (1957). *“Least Squares Quantization in PCM”*)
* Mezclas Gaussianas (McLachlan G. Meng E (1987). *“Mixture Models: Inference and Applications to Clustering”*)
* Agrupamientos jerárquicos (MacQueen J. (1967) “*Some methods for classification and analysis of multivariate observations”*
* Mapas auto.organizados (Kohonen T (1982) *“Self-Organizing Maps“*

### 4.3 Aprendizaje por refuerzo

Los algoritmos de aprendizaje por refuerzo definen modelos y funciones enfocadas en maximizar una medida de “recompensas”, basados en “acciones” y al ambiente en el que el agente inteligente se desempeñará. Al definir estas reglas, el algoritmo intenta explorar diferentes opciones y posibilidades, evaluando cada resultado para determinar donde maximizará su recompensa.

El agente aprende a interactuar con su entorno, mediante el ensayo y error, aprendiendo de las opciones pasadas, recordando aquellas que maximizaron la recompensa. Dado que el agente interactúa con su entorno los datos se encuentran sin procesar, a diferencia de los otros tipos de aprendizaje.

Diagrama

Descripción generada automáticamente

Ilustración 7. Diagrama de aprendizaje por refuerzo

Fuente: Luna J. (2018)Tipos de Aprendizaje Automático. Portal Medium

Programación dinámica (Bellman R. [1950] *“Dynamic Programming”*) y Q-learning (Watkins C. [1989]. *“Learning from delayed rewards”*) son los algoritmos más utilizados en este campo, donde los robots son los principales agentes que se utilizan.

## Redes Neuronales

Una red neuronal es un modelo simplificado que copia el modo en que el cerebro humano procesa la información. El primer modelo de una red neuronal artificial surge en 1943, creado por Warren McCulloch y Walter Harry Pitts. Este modelo, pretende simular el funcionamiento de una neurona en el cerebro humano.

En 1949, el fisiólogo Donald Hebb presentó en su libro *The Organization of Behavior* la regla de aprendizaje. Su trabajo tenía que ver con las conexiones entre neuronas, lo cual es la base hasta el día de hoy de la forma en cómo se entiende que aprenden las neuronas. Este trabajo fue fundamental para que, en 1957, Frank Resenblatt, creara el Perceptrón, la primera red neuronal. Este modelo era capaz de generalizar, después de haber aprendido una serie de patrones podía reconocer otros similares, aunque no se le hubiesen presentado en el entrenamiento.

Una red neuronal artificial consiste en unidades de procesamiento interconectadas de manera densa, llamadas neuronas, por tener un comportamiento similar al de las neuronas biológicas. Las unidades de procesamiento reciben, procesan y transmiten señales, tal como las neuronas biológicas.

Según de Armas (2009), *“las redes neuronales artificiales están compuestas por muchos elementos sencillos denominados neuronas que operan en paralelo y son diseñadas para mostrar una función particular mediante el ajuste de los valores de peso de las conexiones, que actúan con terminada polarización , conocida como “bias”, sobre la función de activación , así se obtiene la salida adecuada en respuesta a la señal de entrada recibida”*.

Diagrama

Descripción generada automáticamente

Ilustración 8. Red neuronal

Fuente: Campbells E., Varela E. (2011) Revista Investigación y Desarrollo en TIC

La neurona puede tratar muchos valores de entrada como si fueran uno solo; esto recibe el nombre de entrada global. La función de entrada es la que permite que estas entradas se combinen con los pesos , dependiente del grado de influencia que tengan estos sobre los valores de entrada. Si bien normalmente la función se expresa con una sumatoria, podría tomar otra operatoria según sea el objetivo.

(1)

Una vez obtenido el valor de entrada con la regla de propagación de la función de entrada, se filtra a través de la función de activación, y es la que genera la salida de la neurona.

La función de activación calcula el estado de actividad de una neurona, transformando la entrada en un valor o estado de activación, que normalmente puede ir entre 0 y 1 o de -1 a 1.

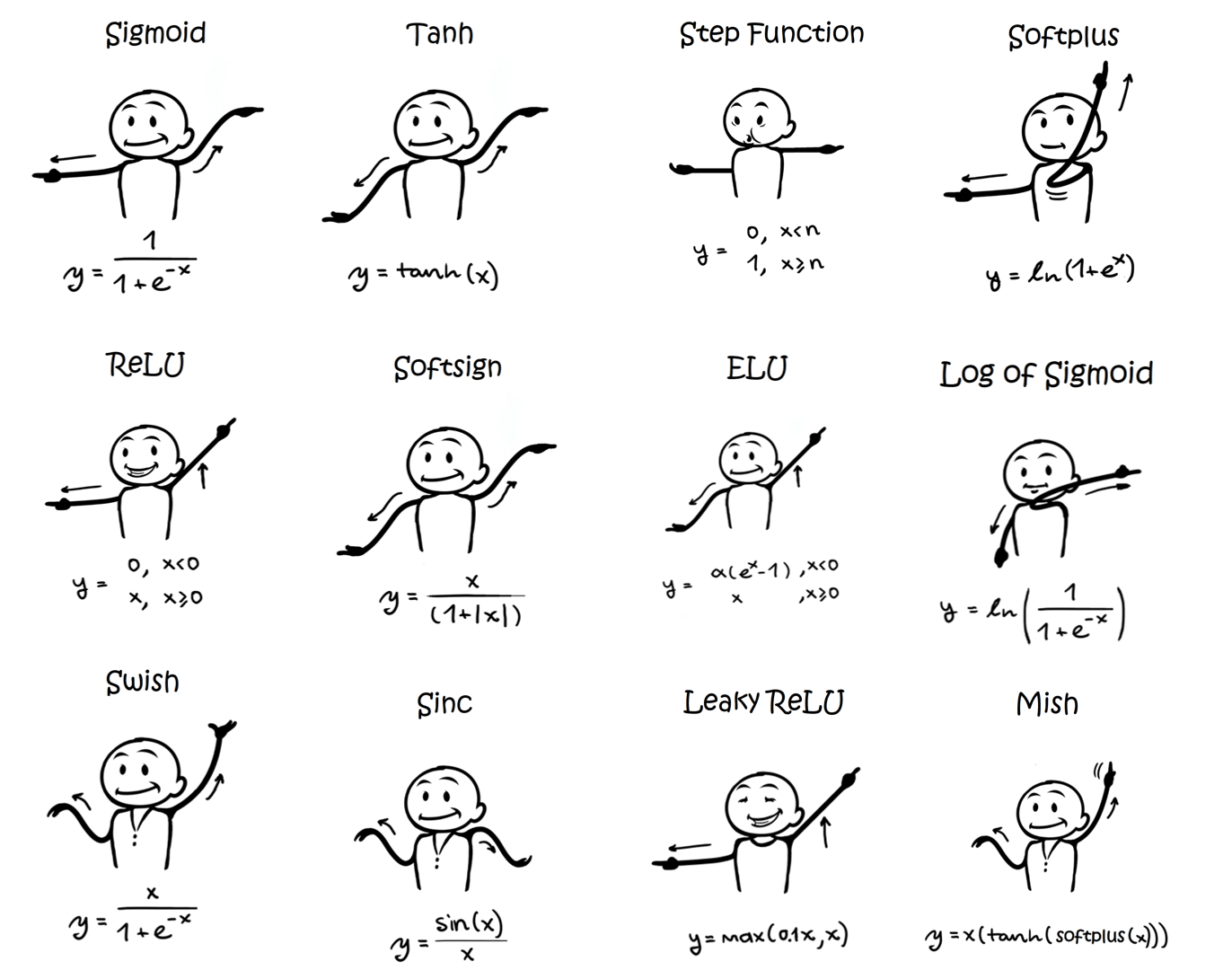


Ilustración 9. Ejemplos de funciones de activación

Fuente: Reddit (2021). Deep Learning Activations functions using dance moves.

Finalmente, el ultimo componente de la neurona es la función de salida. Esta determina que valor se transfiere al conjunto de redes neuronales vinculadas.

## Métricas de precisión

Los modelos de machine learning una vez implementados, deben pasar por una seria de pruebas que de acuerdo con que lo que están prediciendo o clasificando permita tener certezas, o al menos altas probabilidades de que sus resultados son correctos o al menos mejores que otro modelo, funcionando, así como una estrategia de discriminación de la eficiencia / efectividad entre distintos modelos.

Dentro de las métricas relevantes en los modelos de machine learning, podemos encontrar las siguientes:

* Matriz de Confusión. *(Pearson [1904])*
* Accuracy. *(Samuel [1950])*
* Recall. *(Raiffa [1961])*
* F1. *(Van Rijsbergen [1979])*
* Curva ROC *(Fawcett [1977])*
* MAE *(Pearson [1892])*
* MSE/RMSE *(Gauss [1821])*
* R2 *(Pearson [1892])*

Cada métrica presentada permite medir la eficiencia del modelo, en función de qué tipo de problema se busca resolver, sean estos de regresión o de clasificación.

Para efectos de este estudio, se considera un problema de regresión, por tanto, las métricas que se utilizarán para medir y finalmente discriminar por la eficiencia de los modelos se realizarán mediante las métricas MAE, MSE, RMSE, R2.

### 6.1 Error absoluto medio (MAE)

Representa el promedio de la diferencia absoluta entre los valores reales y estimados en el conjunto de datos. Mide el promedio de los residuos en el conjunto de datos, lo cual es la distancia que existe en los valores estimados por el modelo respecto a los valores originales.

Donde,

Como regla de decisión se considera que mientras menor sea su valor, mejores predicciones genera el modelo.

### 6.2 Error cuadrático medio (MSE)

El MSE es la media de las diferencias al cuadrado entre los valores estimados y los valores reales. Es una medida robusta de la precisión que no es sensible a los valores atípicos. Mide la varianza de los residuos del conjunto de datos.

Al elevar al cuadrado se evita la cancelación de términos negativos y da un mayor peso a los valores atípicos, lo que da como resultado un gradiente suave para errores pequeños.

### 6.3 Raíz del Error cuadrático medio (RMSE)

Es la raíz cuadrada de MSE. Es una medida del error medio entre los valores estimados y los valores reales. Es una medida de precisión más interpretable que el MSE.

Se usa ampliamente para evaluar el desempeño de modelos de regresión con otros modelos aleatorios, ya que tiene las mismas unidades que la variable dependiente.

### 6.4 Coeficiente de determinación (R2)

Representa la proporción de la varianza en la variable dependiente que se explica por el modelo de regresión lineal. Es una puntuación sin escala, es decir, independientemente de que los valores sean pequeños o grandes, el valor de R cuadrado será menor que uno.

Diagrama, Forma, Círculo

Descripción generada automáticamente

Ilustración 10. Bondad de ajuste en diagrama

Fuente: Gujarati, D., Porter, D. (2010). *Econometría*. (p, 74)

El coeficiente de determinación r2 , también conocido como bondad de ajuste, representa la fuerza de que tan bien se ajusta el modelo a las observaciones reales que tenemos, sin embargo, no dice nada sobre el modelo en sí, no indica si es bueno malo, si los datos están sesgados o si hemos elegido el método correcto.

Para obtener este coeficiente, se aplica la siguiente función a los datos:

Esta fórmula se define como el cociente de la diferencia entre los valores estimados y el promedio de los valores observados y la diferencia entre los valores observados respecto al promedio de los valores observados.

El coeficiente r2 tiene una regla de decisión asociada a los valores que puede tomar el coeficiente, que va entre 0 y 1.

|  |  |
| --- | --- |
| **Valores r2** | **Bondad de ajuste** |
| 0 – 0,5 | Ajuste débil |
| 0,51 – 0,8 | Ajuste moderado |
| 0,81 – 1 | Ajuste relevante |

Tabla 3. Reglas de decisión sobre coeficiente de determinación r2

Fuente: Elaboración Propia, en base a Gujarati (2010)

# CAPÍTULO III: METODOLOGÍA

La metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) fue creada en 1996 por ingenieros de la empresa SPSS. Estos, desde 1990 se encontraban trabajando con la empresa DaimlerChrysler (en ese entonces, Daimler-Benz) que ya tenía experiencia aplicando minería de datos en sus procesos comerciales y operacionales. Esta metodología pretende ser una guía de referencia para el desarrollo de proyectos en Minería de Datos.

A lo largo de la breve e intensa historia de la minería de datos se han propuestos diversas metodologías para enfrentar los proyectos, por ejemplo, la Knowledge Discovery in Databases (KDD), metodología SEMMA (Sample, Explore, Modify, Modelo, and Access) y la CRISP-DM, siendo esta última la más utilizada en proyectos.

La gran virtud de esta metodología, respecto a las otras y principal motivo por el cual se ha elegido para abordar el problema de este proyecto, es que se ha creado desde la experiencia, desde la aplicación real de diversas empresas, instituciones e investigaciones académicas aplicadas, recogiendo las buenas prácticas que durante años estos actores han adquirido por la experiencia real.

CRISP-DM está compuesta por seis procesos, interconectados, que de una manera jerárquica y cíclica permiten desarrollar de manera coherente un proyecto de minería de datos. Estos procesos se presentan en la *ilustración 9*.

Diagrama

Descripción generada automáticamente

Ilustración 11. Ciclo de la metodología CRISP-DM

Fuente: BSG Institute (2015). CRISPM-DM: Una metodología para proyectos de Data Mining

Se presenta a continuación una breve descripción de cada etapa de la metodología seleccionada.

## Comprensión del negocio

Implica definir los objetivos del negocio, el marco general de intervención, antecedentes generales y la situación actual. Esto permite comprender el contexto de negocio que es clave para determinar los objetivos y requerimientos de la minería de datos y los criterios para evaluar el éxito del proyecto, como lo son los riesgos y contingencias, requisitos, supuestos y restricciones y los recursos necesarios.

## Comprensión de los datos

En esta etapa se realizan principalmente cuatro acciones:

* Recolección de datos: generar el set de datos
* Descripción de los datos: Cantidad, tipo de datos, frecuencia
* Exploración de los datos: pruebas estadísticas, gráficas
* Verificación de calidad de los datos: datos perdidos, errores

## Preparación de los datos

El proceso de preparación se inicia con la *selección de los datos*, donde se puede excluir algunas filas y/o columnas que no sean relevantes para el análisis. De este paso se realiza la *limpieza de datos* relacionado al tratamiento de los datos perdidos, errores de datos.

Luego, en la *construcción de nuevos datos* se pueden crear nuevas columnas, a través de la aplicación de ingeniería de características, creando nuevos atributos que permitan complementar el análisis realizado.

En la siguiente etapa de *integración de datos se* involucra nuevas estructuras, creación de nuevos registros o fusión de tablas de datos. Finalmente, el este proceso se cierra con el *formato de datos,* donde se deben transformar los datos, si el análisis o el modelo a trabajar lo requiere.

## Modelamiento

Este proceso consta de cuatro acciones:

Selección de técnicas de modelaje: Consiste en seleccionar la técnica de minería de datos más adecuada, considerando el problema que se desea resolver.

Generación diseño de prueba: configurar parámetros, probar diferentes modelos, dividir los datos de entrenamiento de los datos de prueba.

Construir el modelo: De acuerdo con la técnica escogida y con los datos preparados se construyen los modelos, cada uno con su respectiva configuración de parámetros.

Evaluación del modelo: Una vez determinado el modelo se procede a evaluarlo y revisar los parámetros de configuración. Con esto se busca los más precisos y eficaces para las consideraciones finales.

## Evaluación

Las tres acciones de evaluación son:

Evaluar los resultados : se evalúan los resultados obtenidos del modelo y se comparan con los objetivos del negocio y las métricas planteados en la fase inicial.

Revisar el problema: es una revisión metodológica en busca de inconsistencia en el modelo o posibilidades de mejora.

Determinar pasos a seguir: se establece la manera de realizar el despliegue y que se puede hacer con el conocimiento generado.

## Despliegue

En esta última etapa se define el plan de despliegue, buscando aplicar el conocimiento generado en la resolución de los problemas identificados. También se define el plan de implementación y mantenimiento, documentando la mejor manera de llevar a cabo la propuesta y como asegurar su funcionamiento a través del tiempo.

Finalmente, se elabora el informe final y se generan los espacios para la revisión del proyecto. Aquí se buscar dejar el registro de la estrategia de trabajo que se realizó y los resultados asociados a esta metodología.

# CAPÍTULO IV: ANÁLISIS DE RESULTADOS

Para la presentación de los resultados obtenidos durante la ejecución del proceso de ciencia de datos realizado, se utilizará la metodología CRISPM-DM creada por el BSG Institute (2015) describiendo en cada etapa el trabajo realizado. La ejecución de este modelo se realizó mediante lenguaje Python.

## Comprensión del negocio

De acuerdo con la definición del problema del CAPITULO 1, la predicción de la mortalidad de peces en un centro de cultivos de salmónidos es una oportunidad importante para generar una herramienta de gestión que permita a la industria tomar decisiones oportunas en cuanto a uno de los grandes “dolores” que es la mortalidad.

Dentro de las enfermedades que causan la mortalidad de los salmones las más relevantes son de carácter infecciosas, y de acuerdo con la empresa, las más importantes a considerar son SRS, BKD, Tenacibaculosis, HSMI. Esto se valida a través del análisis realizado en el set de datos creado. En el *Gráfico 1.* Análisis de mortalidad por tipo de mortalidad

se aprecia, por ejemplo, un brote de enfermedad infecciosa en el centro Site 3Av, rondando los 200 días de permanencia en el agua.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico 1. Análisis de mortalidad por tipo de mortalidad

Fuente: Elaboración propia

## Comprendiendo los datos

### Recolección de datos

El set de datos ha sido obtenido a través de los servidores de la empresa y se han realizado consultas directamente para crear un set lo más limpio posible, generar una eficiencia en tiempo de preprocesamiento y para mantener la información lo más fidedigna posible.

El set de datos contiene 104.878 filas y 36 columnas.

### Descripción de los datos:

El set de datos presenta 36 columnas que representan los valores y atributos relevantes dentro del análisis de los centros de cultivos considerados en el estudio. Estos están distribuidos en 8 datos de tipo categórico y 28 datos de tipo numérico. En la *Tabla 4.* Se detallan las variables del modelo.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable del modelo** | **Tipo dato** | **Variable del modelo** | **Tipo dato** | **Variable del modelo** | **Tipo dato** |
| Barrio | numérica | MortalityPorc | numérica | Specie | categórica |
| Latitude | numérica | IsInfectiousTypeMortality | numérica | SiteName | categórica |
| Longditude | numérica | InfectiousMortalityCount | numérica | ModelName | categórica |
| Date | numérica | MortalityPorcInfectious | numérica | Unit | categórica |
| Year | numérica | MortalityByMechanicalDamage | numérica | ActiveSubstance | categórica |
| Month | numérica | MortalityByPredator | numérica | TreatmentMethod | categórica |
| Day | numérica | FeedDays | numérica | TreatmentProduct | categórica |
| Season | numérica | Density | numérica | TreatActiveSubstance Concentration | categórica |
| CloseCount | numérica | Temperature | numérica |  |  |
| CloseBiomass | numérica | Oxygen | numérica |  |  |
| CloseWeight | numérica | HasMovement | numérica |  |  |
| LiveDaysInProduct | numérica | LostFeed | numérica |  |  |
| MortalityCount | numérica | FeedAmount | numérica |  |  |
| MortalityBiomass | numérica | TreatAmount | numérica |  |  |

Tabla 4. Variables del modelo

Fuente: Elaboración Propia

Los datos de tipo numérico ( *Ver Gráfico 2.*) se describen de mejor manera en las siguientes gráficas de distribución. Para efectos de mejor comunicación visual se consideran solo aquellas que muestran relevancia para efectos del análisis posterior. Los días de vida, densidad, temperatura y nivel de oxígeno del agua se presentan con la mayor densidad en el set de datos.

Diagrama

Descripción generada automáticamente

Gráfico 2. Distribución de variables numéricas

Fuente: Elaboración propia

Así mismo, en el *Gráfico 3* se presentan 8 variables categóricas, siendo las especies de salmón (trucha, coho y salar), temporadas y nombre del centro las más relevantes.

Interfaz de usuario gráfica

Descripción generada automáticamente

Gráfico 3. Distribución de variables categóricas

Fuente: Elaboración propia

### Exploración de los datos

El primer análisis para realizar con los datos, desde la perspectiva del negocio, es conocer la cantidad de días que tienen los peces en el agua. Esto es relevante, dado que la mortalidad de los salmónidos tiene directa relación con la cantidad de tiempo que pasan en el agua, debido al ciclo natural de la especie.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Gráfico 4. Distribución de días de supervivencia promedio de los salmones por centro de cultivo

Fuente: Elaboración propia

En el *Gráfico 5* se aprecia justamente que la mayor cantidad de mortalidad se genera entre los 300 y 600 días en el agua, aproximadamente al año de vida. También se destaca los niveles de mortalidad temprana, entre 0 y 30 días, que dicen relación principalmente a pasar los salmones a los centros de agua mar.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Gráfico 5. Distribución de porcentajes de mortalidad por rangos de días

Fuente: Elaboración propia

Ahora bien, los peces ganan peso a través del tiempo, y esto queda establecido en el *Gráfico 6*. Es un dato más bien normal, dado que los centro de engorda justamente existen para que los salmones ganen el peso necesario para ser cosechados y transformados en productos de valor agregado.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Gráfico 6. Distribución de peso promedio por rango de días

Fuente: Elaboración propia

Se aprecia la dispersión respecto de los porcentajes de mortalidad en el *Gráfico 7* respecto de los pesos de los salmones. Se aprecia una dispersión en la mortalidad, focalizando la mayor mortalidad a los 6 kilos aproximadamente, que se corresponde con el tiempo de permanencia en el centro para obtener ese peso.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico 7. Dispersión entre porcentaje de mortalidad y el peso

Fuente: Elaboración propia

Dentro de las variables interesantes a reconocer son las enfermedades infecciosas. El *Gráfico 8* nos muestra un brote de enfermedad infecciosa generada entre los 180 y 240 días de supervivencia de los salmones.

Gráfico, Histograma

Descripción generada automáticamente

Gráfico 8. Mortalidad infecciosa y total

Fuente: Elaboración propia

Al realizar análisis de correlación de las variables se encuentran algunos datos altamente correlacionados los cuales generan ruido en los análisis posteriores, y en algunos casos, la no correlación es condición para que algunos modelos estadísticos, por ejemplo, la Regresión Lineal (1889).

Gráfico, Gráfico de barras

Descripción generada automáticamente

Gráfico 9. Escala de correlaciones de variables del modelo

Fuente: Elaboración propia

Estadísticamente se realiza el test de Shapiro-Wilk (1965) para determinar si existe normalidad en la distribución de los datos. Donde se debe probar que:

El *p-value* obtenido es menor que el nivel de significancia del 5% por tanto, no se rechaza la H0, por tanto, los datos tienen distribución normal.

### Verificación de calidad de los datos

Dado que en el proceso de extracción de datos se han realizado las consultas considerando la resolución de datos perdidos, duplicados y de errores, no se presentan éstos en el set de datos.

## Preparación de los datos

Dado la formulación del problema y la calidad de los datos que se poseen, se concluye que el enfoque debe ser para resolver un problema de aprendizaje supervisado, cuya variable objetivo es la columna ‘*MortalityPorc’.* En análisis previo, cuando se genera la base de datos se agrega la columna *‘LiveDaysInProduct*‘ para indicar los días que el grupo de peces está en un lugar determinado (centro) y poder hacer comparaciones temporales entre ellas. Así mismo, se agregó la variable *‘InfectiousTypeMortality’* para indicar la cantidad de la mortalidad que clasifica si la enfermedad que detecta el acuicultor corresponde a un tipo que sea considerada como un brote infeccioso.

Luego, se realiza la eliminación de variables altamente correlacionadas para mejorar el set de datos. Las variables *'Date', 'Unit', 'ModelName', 'CloseWeight', 'MortalityBiomass, 'IsInfectiousTypeMortality' y ‘CloseBiomass’.* Esto se expone de mejor manera en el *Gráfico 10.*

Gráfico, Gráfico de barras

Descripción generada automáticamente

Gráfico 10. Escala de correlaciones sin variables altamente correlacionadas

Fuente: Elaboración propia

Un aspecto importante para considerar son los outliers. Del análisis de distribución de las variables se han identificado variables con una cantidad importante de outliers, en algunas variables ambientales, por ejemplo. Para esto inicialmente, se quitan todos los registros de jaulas que tengan saldo negativo por ser considerados un error y los de porcentaje de mortalidad elevado, asumiendo una mala imputación de los datos. Esto fue corroborado por la empresa, dado que los datos de mortalidad se registran normalmente de manera manual.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico 11. Análisis días de vida vs mortalidad registrada

Fuente: Elaboración propia

En el *Gráfico 12.* se verifican algunos datos respecto a la temperatura y oxígeno del agua, dado que se estima puedan existir outliers.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Gráfico 12. Distribución de temperaturas y oxigeno por estación

Fuente: Elaboración propia

Efectivamente se aprecian valores anormales, que responden según los expertos de la empresa, a fallas en los sensores que capturan los datos.

Finalmente, para efectos de mejor la eficiencia de los modelos algoritmos a utilizar en el desarrollo del proyecto se desarrolla la conversión de variables categóricas a variables numéricas, mediante la técnica del Label Encoder (Scikit-Learn Library, 2007).

## Modelamiento

De acuerdo con el problema formulado, en cuanto a la necesidad de contar con un pronóstico respecto a la tasa de mortalidad de salmónidos en un centro de cultivos de agua mar orienta nuestro modelado hacia un problema de regresión. Para ello, se han seleccionado cuatro modelos. La regresión lineal, algoritmo XGBoost, Random Forest y una Red Neuronal Recurrente LSTM. Esta selección es arbitraria, respondiendo a una lógica de probar la efectividad de técnicas estadísticas, de machine learning y de redes neuronales para la resolución de este problema.

### Selección de técnicas de modelaje

### Regresión Lineal

La regresión lineal es una técnica estadística que se utiliza para modelar la relación entre una variable dependiente y una o más variables independientes (Gujarati, 2010). La variable dependiente es la variable que se está tratando de predecir, y las variables independientes son las variables que se utilizan para predecir la variable dependiente.

La regresión lineal es un modelo de regresión que utiliza una variable independiente para predecir una variable dependiente. Según Gujarati (2010), la ecuación de regresión lineal simple es:

Donde,

La pendiente de la línea de regresión representa la relación entre la variable independiente y la variable dependiente. Si la pendiente es positiva, entonces un aumento en la variable independiente está asociado con un aumento en la variable dependiente. Si la pendiente es negativa, entonces un aumento en la variable independiente está asociado con una disminución en la variable dependiente.

Cuando el modelo depende de más de una variable dependiente, se conoce como una regresión lineal múltiple.

### Random Forest

Random Forest o Bosques Aleatorios fue propuesto por Ho (1995) y consiste en crear muchos árboles de decisión para luego usarlos en la predicción de la variable de interés. En la *Ilustración 12* , se muestra una representación de la técnica.

Diagrama

Descripción generada automáticamente

Ilustración 12. Técnica del Random Forest

Fuente: Hernández F. (2023) Modelos Predictivos. Rescatado de <https://fhernanb.github.io/libro_mod_pred/index.html> el 18 de julio de 2023.

El algoritmo de Random Forest, desarrollado por Breiman (2001) es una técnica de aprendizaje supervisado que se aplica sobre un conjunto de datos de entrenamiento. Los resultados obtenidos se combinan a fin de obtener un modelo único más robusto en comparación con los resultados de cada árbol por separado.

Cada árbol generado por el algoritmo Random Forest contiene un grupo de observaciones aleatorias elegidas mediante *Bootstrap* (Efron, 1979), que es una técnica estadística para obtener muestras de una población donde una observación se puede considerar en más de una muestra. Las observaciones no estimadas en los árboles, conocidas como “out of the bag” se utilizan para validar el modelo. Las salidas de todos los árboles se combinan en una salida final denota por , llamado ensamblado, que se obtiene mediante alguna regla. En el caso de los problemas de regresión, se obtiene mediante la media aritmética de los resultados de cada ensamble.

La técnica que se destaca para el ensamble de los algoritmos en el Random Forest es el *Bagging (Boatstraping Aggregating).* Esta técnica creada por Breiman (1996) busca reducir la varianza dentro de un conjunto de datos a través de seleccionar una muestra aleatoria de datos en un conjunto de entrenamiento con reemplazo, lo que significa que los puntos de datos individuales se pueden elegir más de una vez.

Este algoritmo consta de tres pasos:

* *Bootstrapping:* Este método de muestreo genera diferentes subconjuntos del conjunto de datos de entrenamiento seleccionando puntos de datos al azar y con reemplazo.
* *Entrenamiento paralelo:* Las muestras de bootstrap luego se entrenan de forma independiente y en paralelo entre sí utilizando aprendizajes débiles o básicos.
* *Agregación:* Finalmente, se toma un promedio o la mayoría de las predicciones (según la tarea que se esté realizando) para calcular una estimación más precisa. Como fue comentado, para esta aplicación, se toma el promedio por ser un problema de regresión.

### XG Boost

Este algoritmo fue creado por Tianqi Chen (2014), quien en su publicación XGBoost: A Scalable Tree Boosting System describió que este algoritmo es una versión mejorada de Gradient Boosting Machine.

XGBoost funciona construyendo un conjunto de árboles de decisión. Cada árbol de decisión se entrena en un conjunto de datos diferente, que es una submuestra del conjunto de datos de entrenamiento. Los árboles de decisión se entrenan en secuencia, y cada árbol se entrena para corregir los errores cometidos por los árboles anteriores.

El proceso de entrenamiento de XGBoost se basa en el principio de descenso de gradiente. Comienza con un modelo inicial simple y luego se ajusta iterativamente para minimizar una función de pérdida. En cada iteración, se calcula el gradiente de la función de pérdida en relación con el modelo actual y se ajustan los parámetros del árbol de decisión para minimizar el gradiente. Esto se puede apreciar de gráficamente en la *Ilustración 13.*

Diagrama

Descripción generada automáticamente

Ilustración 13. Representación del algoritmo XGBoost

Fuente: Guo et al. (2020). Degradation State Recognition of Piston Pump bases on ICEEMDAN and XGBoost

La diferencia respecto al Random Forest, radica en la técnica que se utiliza se aprendizaje. Mientras que Random Forest utiliza la técnica del *Bagging*, donde el entrenamiento se genera de manera paralela de acuerdo con las submuestras obtenidas, el XGBoost utiliza la técnica del *Boosting.*

El *Boosting*, creado por Breiman (1996) es una técnica que permite reducir la varianza y el sesgo de un conjunto de datos A diferencia de otros enfoques de combinación de modelos, como el ensamblado simple, donde los modelos se entrenan de forma independiente y luego se promedia o vota sobre sus predicciones, como en el caso de los Random Forest. El *Boosting* se basa en la idea de construir modelos secuenciales, donde cada modelo se enfoca en corregir los errores cometidos por los modelos anteriores.

El proceso de *Boosting* comienza con un modelo inicial simple, que puede ser cualquier algoritmo de aprendizaje automático débil, como un árbol de decisión poco profundo o un clasificador débil. Luego, se ajusta iterativamente un nuevo modelo en función de los ejemplos de entrenamiento mal clasificados o los residuos del modelo anterior. En cada iteración, se asigna un mayor peso a los ejemplos mal clasificados, lo que permite que el nuevo modelo se centre más en corregir los errores cometidos anteriormente. Luego, se combina el modelo actualizado con los modelos anteriores, generalmente mediante la suma ponderada de las predicciones de cada modelo, donde los pesos se determinan en función de la precisión de cada modelo.

Este proceso de construcción secuencial de modelos se repite hasta que se alcanza un criterio de parada, como un número máximo de modelos o cuando se ha alcanzado un nivel deseado de precisión. El resultado final es un modelo robusto y preciso que combina las fortalezas de los modelos débiles individuales.

### Red neuronal recurrente LSTM

Las Redes neuronales recurrentes (RNN por sus siglas en inglés) son tipos de redes que permiten reconocer y predecir secuencias o datos a lo largo del tiempo. El text mining y el deep learning, son algunas disciplinas de aplicación de este tipo de red.

Según IBM (2023) *“Las Redes neuronales recurrentes LSTM se distinguen por su memoria, ya que obtienen información de entradas anteriores para influir en la entrada y salida actuales. Mientras que las redes neuronales profundas tradicionales asumen que las entradas y salida son independiente entre sí, la salida de las redes neuronales recurrentes depende de los elementos anteriores dentro de la secuencia”*.

Este tipo de red neuronal fue publicada en 1997 por Sepp Hochreiter y Jürgen Schmidhuber en su paper “long short-term memory”, donde presentan las celdas de memoria con una entrada y salida adicional, llamadas celdas de estado, respecto a las redes neuronales recurrentes básicas. Estas celdas están en las capas ocultas de la red neuronal que tienen tres compuertas: una compuerta de entrada, una compuerta de salida y una compuerta de olvido. Estas compuertas controlan el flujo de datos que se necesita para pronosticar la salida en la red.

Diagrama, Esquemático

Descripción generada automáticamente

Ilustración 14. Arquitectura de una celda de memoria

Fuente: Hochreiter S. Schmidhuber J. (1997) Long short-term memory

La celda de memoria se construye en torno a una unidad lineal central con conexiones a cada célula. La celda recibe entradas por una unidad multiplicativa de salida () y de otra unidad multiplicativa de entrada (). Finalmente, la activación de la entrada en el momento se denota por:

(1)

Donde,

, (2)

, (3)

Por tanto,

(4)

En las Redes neuronales básicas, la celda de memoria tendrá una estructura muy simple, como una sola capa de activación. Por su lado, las RNN LSTM también tienen esta estructura tipo cadena, pero la celda de memoria tiene una estructura diferente. En lugar de tener una sola capa de red neuronal, hay cuatro que interactúan de una manera muy especial.

El proceso de transferencia de datos es el mismo que el de las redes neuronales recurrentes básicas, sin embargo, la operación de propagación de la información es diferente. Cuando la información pasa a través de ella, la operación decide qué información procesar más y qué información debe dejar ir. La operación principal consiste en células y puertas. El estado celular funciona como una vía para transferir la información. Se puede considerar a las células como memoria.

Hay varias puertas en el proceso de LSTM. Cuando el estado celular lleva la información, estas puertas ayudan a que la nueva información fluya. Las puertas indicarán qué datos son útiles para guardar y qué datos no son útiles, por tanto, sólo los datos relevantes pasan a través de la cadena de secuencia para una fácil predicción.

Diagrama

Descripción generada automáticamente

Ilustración 15. Arquitectura de RNN Long short-term memory

Fuente: Ayala R. (2022). Series Temporales, Op3: LSTM Data Science Research Perú (<https://datasciencepe.substack.com/p/series-temporales-op-3-lstm>)

### Generación diseño de prueba

Una vez ya definidos los algoritmos a desarrollar, junto con la normalización de los datos que fue solicitada por la empresa para garantizar el resguardo de sus datos, el proceso de modelado y evaluación está listo para ser iniciado.

Mediante la instancia *Split,* se dividen los datos en set de entrenamiento y set de prueba. Se considera una partición 80/20, donde el 80% corresponde a datos de entrenamiento y el 20% a datos para test, quedado como se indica en la *Tabla 5*.

|  |  |
| --- | --- |
| Set | Tamaño del set |
| X\_train | 83.856 filas, 42 columnas |
| X\_test | 20.964 filas, 42 columnas |
| Y\_train | 83.856 filas |
| X\_test | 20.964 filas |

Tabla 5. Tamaño del set de entrenamiento y prueba

Fuente: Elaboración propia

Para efectos de la regresión lineal se trabaja directamente sobre el set de datos de entrenamiento, sin requerir mayor parametrización de los componentes del modelo.

### Construcción de los modelos

Se presentan en este apartado, los algoritmos utilizados para la resolución de los modelos presentados.

Para el caso del algoritmo Random Forest, como se aprecia en la *Ilustración 17.* se corrió el modelo con los siguientes parámetros:

Por su parte, el XGBoost,

Finalmente, la red neuronal recurrente LSTM

### Evaluación de los modelos

Una vez corridos los cuatro modelos se procede a evaluarlos de acuerdo con las métricas de precisión notadas en e*l* [*Capítulo II.6 Métricas de Precisión*](#_Métricas_de_precisión)*.*

***GRAFICA IGUAL A LA DE METRICAS CON PARAMETROS OPTIMIZADOS***

En esta primera instancia de evaluación, se concluye que el modelo Random Forest es el más robusto en cuanto a bondad de ajuste respecto a los datos de entrenamiento. Sin embargo, es necesario generar un estudio más, con la finalidad de asegurar el uso de los mejores hiperparametros de cada modelo.

Mediante la técnica *GridSearchCV* se busca la optimización de los hiperparámetros que permitan mayor efectividad en los modelos. Para el caso del modelo Random Forest, los hiperparámetros resultantes de la optimización, se presentan en la *Ilustración 21*.

El modelo XGBoost, utiliza 300 estimadores, un *max\_depth* de 3 y un *learning\_rate* 0,1.

Finalmente, el modelo de red neuronal recurrente se

## Evaluación

Finalmente, se obtienen las métricas de precisión de los modelos una vez que han sido optimizados los hiperparámetros. Como se aprecia en el *Gráfico 14.* El modelo desarrollado mediante el algoritmo Random Forest sigue siendo la mejor alternativa para realizar la predicción de mortalidad en los centros de cultivo de acuerdo con los datos utilizados.

El algoritmo Random Forest es más eficiente en todas sus métricas, obteniendo el menor valor en las métricas RSME, MSE y MAE, al mismo tiempo que obtiene un R2 superior a las otras técnicas probadas.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

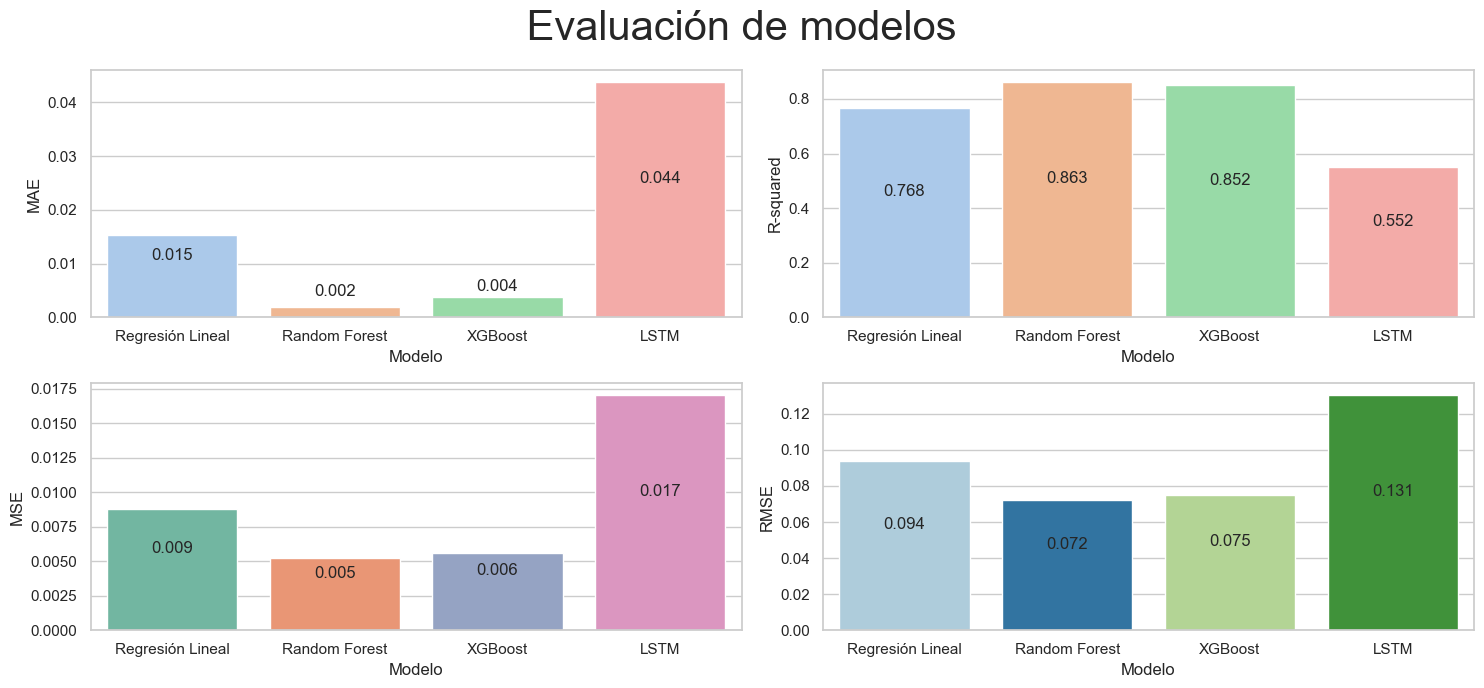


Gráfico 13. Comparativa de Métricas de precisión por modelo

Fuente: Elaboración propia

# CAPÍTULO V: CONCLUSIONES Y RECOMENDACIONES

Se recomienda para mejorar las metricas y publicar este mdelo en el ambito empresarial, contar con mas recurso de computo, Ya sea on premise o rentado en cloud.

Agregar mas caracteristicas, sobretodo ambientales desde iot y de componentes del alimentos entregado a los peces, que pudiesen aportar a predecir de mejor forma los porcentajes de mortalidad por dia de vida en la etapa de engorda de los salmones.

# BIBLIOGRAFÍA

* Bellman R. (1950). Dynamic Programming. Princeton University Press
* Berkson J. (1944). Application of the Logistic Function to Bio-Assay. Journal of the American Statistical Association: Vol. 39, No. 227, pp. 357-365.
* BSG Institute (2015). CRISP-DM: Una metodología para proyectos de data mining.
* Breiman, L. (1996). Bagging Predictors. Machine Learning 24, 123-140 Ed. Kluwer Academic Publishers, Boston
* Breiman, L. (2001). Random Forests. Machine Learning 45, 5–32 Ed. Kluwer Academic Publishers, Boston
* Camuñez J., Basulto J., Ortega F. (2009). La memoria de Daniel Bernoulli sobre la inoculación contra la viruela (1760) Un problema de decisión bajo incertidumbre. Historia de la probabilidad y la Estadística IV. Ed. Universidad de Huelva
* Chen T., Guestrin C. (2016). XGBoost: A Scalable tree boosting system. arXiv:1603.02754v3 [cs.LG] 10 Jun 2016
* Cox D. (1972). Analysis of Survival Data. Chapmann & Hall/Crc
* Cutler S., Ederer F., (1958). Maximum utilization of the life table method in analyzing survival, Journal of Chronic Diseases,Volume 8, Issue 6, Pages 699-712.
* Diamond S., Yuk M. (2014). Data Visualization for Dummies. Ed. John Wiley & Sons, Inc.
* Duda R., Hart P. (1973) Pattern Classification and Scene Analysis. Ed Wiley- Interscience.
* Fix, E. and Hodges, J.L. (1951) Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties. Technical Report 4, USAF School of Aviation Medicine
* Galton F. (1889) “Natural Inheritance”. Ed. Macmillan & Co
* García. J., Molina J., Berlanga A., (2018). Ciencia de datos. Técnicas analíticas y aprendizaje estadístico Ed. Alfaomega
* Gujarati, D., Porter, D., Monroy Alarcón, A. and Cortés Fregoso, J. (2010). Econometría. 5e. México: McGraw-Hill.
* Hochreiter S. Schmidhuber J. (1997). Long short-term memory. Neural Computation 9(8): 1735-1780, Technische Universität München
* Hernández F. (2023). Modelos Predictivos. <https://fhernanb.github.io/libro_mod_pred/index.html>
* IBM SPSS Modeler (2021). Guía de CRISP-DM de IBM SPSS Modeler
* Kaplan E., Meier P. (1958) Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association, Vol. 53, No. 282. Junio, pp. 457- 481
* Kohonen T (1990). The self-organizing map. Proceedings of the IEEE, vol. 78, N°9
* Lloyd S. (1957). Least Squares Quantization in PCM. IEEE transactions on information theory, vol IT-28, N°2
* MacQueen J. (1967) Some Methods for classification and analysis of multivariate observations. Berkeley Symp. on Math. Statist. and Prob., p. 281-297
* Malthus T. (1846). Ensayo sobre el principio de la población. Traducido por Eusebio María del Valle. Establecimiento Literario y Tipográfico de D. Lucas Gonzalez Matthews R. (2006). La introducción de los métodos estadísticos en la medicina de los siglos XIX y XX. Ars Medica. Revista de Humanidades. Fundación Pfizer
* McCulloch W. Pitts W. (1943). A Logical calculus of the ideas immanent in nervous activity. Bulletin of mathematical biophysics, Volume 5.
* McLachlan G. Meng E (1987). *“Mixture Models: Inference and Applications to Clustering”*
* Mina-Valdés A., (2009). Uso de las funciones de supervivencia en las ciencias sociales y en los estudios de población. Aplicación al caso de México. Papeles de Población, 15(61), 53-74
* Mitchell T. Carbonell J,Michalski R.(1986) Machine Learning: A Guide to Current Research, Ed. Kluwer Academic Publishers
* Quinlan J. (1986). Induction of Decision Trees. Kluwer Academic Publisher
* Salinas M. (2008). Modelos de Regresión VI. Análisis de Supervivencia. Ciencia y Trabajo. Abr-Jun; 75-78
* Soltveit, T. (30 de mayo de 2022). Salmonicultora se convierte en la segunda más grande del mundo tras fusión. *SalmonExpert*. Sección Empresas <https://www.salmonexpert.cl/article/salmonicultora-se-convierte-en-la-segunda-mas-grande-del-mundo-tras-fusion/>
* SPSS Company. (2000). CRISP-DM 1.0 Step by step data mining guide
* Watkins C. (1989). Learning from Delayed Rewards. King's College, Cambridge

# ANEXOS

**ANEXO 1. XXXXXX**

**ANEXO 2. XXXXXX**