NATIONAL UNIVERSITY OF ENGINEERING
COLLEGE ECONOMICS AND STATISTICAL ENGINEERING STATISTICAL ENGINERRING PROGRAM

INFORME DE CURSO CAPSTONE

TITLE

"DESIGN AND CONSTRUCTION OF PREDICTIVE MODELS THAT FIT THE GROWTH CURVE FOR WEIGHT OF MARINE SPECIES OF THE *VITAPRO COMPANY – ALICORP GROUP*"

TITULO

"CONSTRUCCIÓN DE MODELOS PREDICTIVOS QUE SE ADAPTAN A LA CURVA DE CRECIMIENTO DEL PESO DE LAS ESPECIES MARINAS DE LA EMPRESA VITAPRO - ALICORP"

By:

NESTOR JOEL SIMON LEYVA

GERALDINE GIANELA QUISPE TAPARA

SUPERVISOR:

AMÉLIDA PINEDO SANCHEZ

LIMA - PERÚ

2019

INDEX

INDICE

ABSTRACT

The objective of this research work is to find the prediction model that best fits the Gompertz growth curve for fish from crop centers of companies dedicated to marine species breeding, in addition to understanding and discovering factors that influence said growth. So in the research project it was considered as one of the main objectives in the first place, to build a model that makes predictions of the final average weights of marine species, for this the information provided by the company Alicorp was taken, which It contained information on productive and environmental variables of the cages, centers and clients in different periods about the fish harvests.

In the first instance, the data and the distribution of the numerical and categorical variables that will be used in the construction of the model were analyzed. Then the relationship between them and with the objective variable that would be the final average weight of the species was evaluated, consequently when reviewing the level of correlation two variables were created that affect the construction of the model, the first variable is the variation between the initial and final number of marine species by cages, centers and periods, the second variable is in relation to the depth and the amount of light hours they maintain, so a repetitive pattern was sought based on the depth and the average number of artificial light hours was calculated according to this pattern. Once the two new variables were created, the correlations between them and the objective variable were analyzed again.

In the second instance, the number of data lost per variable was analyzed, that is, missing data that comes in the database; Likewise, we reviewed whether there are atypical data, that is, information that is distantly related to the rest of the other values of the same variable, either because it has a value that is too high and too low. According to this, methods such as: imputation by means for the numerical variables and imputation by mode for the categorical variables were applied, this in order to eliminate the amount of lost data existing and avoid having variables that come without information, the Tukey's method of correcting atypical data problems, this method helped us replace very high outliers by the third quantile and very low outliers by the first quantile.

As a next step, we analyzed the objective variable (final average weight) and a standardization of the data was performed, the method of scaling the entire data was also carried out with the exception of the response variable, in order to have better results in terms of development of the model. The data was divided into two parts, the first part consists of 80% of information and was used to train the model, the second part consists of the remaining 20% and was used to test the effectiveness of the model. Models such as Linear Regression, Decision Tree, Random Forest and XGBoost were applied to determine the predicted weights. To know and find the best prediction model, three more relevant

indicators were calculated: mean square error (RMSE), mean absolute percentage error (MAPE), mean absolute error (MAE).

It should be noted that for the Linear Regression model, not all the information was used because records that had outliers or lost values had to be eliminated, since they affect the construction of the Linear Regression model considerably. In addition, compliance with the four assumptions of the Linear Regression model was verified: linearity, normality, autocorrelation and homocedasticity. Regarding the first assumption, it was verified that not all numerical variables have a linear relationship with the objective variable; however, the residues (difference between the real weight and predicted weight of the average weight of the fish) if they maintain a normal distribution, regarding the assumption of homocedasticity it was found that the variation of the residues is uniform throughout the range of the forecasts.

As mentioned in previous lines, after calculating the RMSE, MAPE and MAE indicators, a comparison was made for the Decision Tree, Random Forest and XGBoost models, finding that the best model that predicts the final average weight of the Fish is the Decision Tree model as it has lower values of RMSE, MAPE and MAE. Finally, taking into account these models, the predicted values were calculated and it was evaluated if they fit the growth curve according to the Gompertz model.

Comparison of indicators such as AIC, variability and R^2 was also performed. Although this last indicator was similar for all the models, the model was chosen based on the other indicators where it was obtained that the Random Forest model has lower values according to AIC and variability compared to the other models, that is, it is the one that best fits the Gompertz growth curve. It was also observed that the XGBoost model has lower values in its indicators compared to the Decision Tree model.

Finally, after obtaining different models as results for the adjustment of the final average weight values of fish and the one that best fits the Gompertz growth curve.

1. INTRODUCCIÓN

Alicorp es la compañía de consumo masivo más grande del Perú, y cuenta con operaciones en diversos países de Latinoamérica, dentro de las líneas de negocio de Alicorp, se tiene a la empresa Vitapro, empresa con más de 30 años de experiencia en el sector acuícola, dedicada a la producción y comercialización de alimento para especies marinas. La empresa Vitapro viene creciendo de forma sostenida teniendo un crecimiento en ventas de 15,4% anual en los últimos 4 años y proyecta crecer más de 12% en el 2019. En la revisión corporativa de estrategia de este año, se identificó una oportunidad única de que Vitapro lidere la transformación digital apoyada por la analítica avanzada, en la industria de acuicultura. En el trabajo de investigación presente se plantea construir modelos de predicción de la curva de crecimiento de peces para los centros de cultivos de empresas dedicadas a la crianza de especies marinas, además del entendimiento y descubrimiento de factores que influyen en dicho crecimiento.

Se conoce que el crecimiento de peces se desarrolla en promedio, en alrededor de 14 meses (ciclo productivo), y se encuentra condicionado a factores internos, como datos productivos del centro de crianza (p.ej. alimentación, genética, información foto-productiva, etc.) y externo-ambientales (p.ej. temperatura del agua, oxígeno), es por ello que durante la investigación se considera esencial realizar un análisis descriptivo y multivariados/experimentales que permita realizar una adecuada selección y estudio de variables, con la finalidad de describir de la mejor manera la curva de crecimiento.

Para este trabajo de investigación se tuvo como objetivo determinar si las variables de factores productivos y ambientales afectan al peso promedio final, como primer punto se realizaron análisis descriptivos de estas variables (categóricas y numéricas) y posteriormente se hizo el tratamiento a los datos faltantes con métodos de imputación adecuado, luego se procedió a crear variables, ya que se tenían altos niveles de correlación entre ellas. Hecho lo anterior, se emplearon distintos usos de software estadísticos como Python y Statistica, con el objetivo de construir el mejor modelo de predicción (p. ej. Regresión Lineal, Árbol de Decisión, Random Forest, etc.) de los pesos promedio final de los peces para así poder ajustar al modelo de crecimiento de Gompertz; estos modelos son elegidos mediante criterios de comparación de indicadores como el RMSE, MAPE y MAE para los modelos de predicción y para el caso del modelo de crecimiento se tienen indicadores como el AIC, R^2 y variabilidad.

A nivel de negocio, luego de encontrar el mejor modelo de predicción y el modelo que más se ajuste a la curva de crecimiento de Gompertz, en base a la importancia de las variables en los modelos, permitirá brindar aspectos de mejora a la empresa Álicorp, asimismo algunas estrategias para ampliar su mercado en la industria del cultivo acuícola.

1.1. PLANTEAMIENTO DEL PROBLEMA

La línea de negocio Vitapro perteneciente a la empresa ALICORP, en un contexto de mercado competitivo donde el sector acuícola en el Perú viene creciendo en los últimos años alrededor del 5% según el Ministerio de Producción (PRODUCE). Se plantea predecir la curva de crecimiento de peces para los centros de cultivos de empresas dedicadas a la crianza de especies marinas, además del entendimiento y descubrimiento de factores que influyen en dicho crecimiento. El crecimiento de peces se desarrolla en promedio, en alrededor de 14 meses (ciclo productivo), y se encuentra condicionado a factores internos, como datos productivos del centro de crianza y externo-ambientales.

Es así que, se busca construir el mejor modelo para la curva de crecimiento de los peces, dados ciertos factores. Esto permitirá posicionar a Vitapro como una de las pocas empresas en incorporar mejoras digitales para la mejora de la producción y la toma de decisiones corporativas.

FIGURA N° 1: Problema de la Investigación

Fuente: Elaboración Propia.

De esta premisa se desprenden las preguntas de la investigación:

1.1.1. Problema General

¿Qué modelo se ajusta a los pesos promedios finales de las especies marinas y cuál es la curva de crecimiento de los peces para los centros de cultivos dedicados a la crianza de especies marinas?

1.1.2. Problemas Específicos

- ¿Cómo influyen las variables de los factores ambientales en el crecimiento de peces?

- ¿Cómo influyen las variables de los factores de producción en el crecimiento de peces?
- ¿Cuáles son las variables más influyentes en relación al peso promedio final de las especies marinas?
- ¿Cuál es el mejor modelo que predice el peso promedio final de las especies marinas?
- ¿Qué modelo se ajusta mejor a la curva de crecimiento de Gompertz para los peces de los centros de cultivos dedicadas a la crianza de especies marinas?

1.2. JUSTIFICACIÓN

La investigación parte de la importancia de conocer los factores que influyen en el crecimiento de los peces en los criaderos de la empresa Vitapro, conociendo estas variables propondremos estrategias de negocio para que la empresa implemente y así mejorar su producción y comercialización de alimento para las especies marinas.

Debido a que la empresa Vitapro S.A. quiere aumentar y/o mantenerse en los pioneros de la producción y comercialización de alimentos para especies marinas, esto solo se puede medir viendo la producción de especies que obtiene por periodo. Dicho esto, la empresa Vitapro necesita tener conocimiento de aquellas variables que influyen en el crecimiento de peces para la mayor producción de su alimento, por ello en sus criaderos manejan variables relacionadas al ambiente (factor ambiental: oxígeno, cantidad de luz, etc.) y variables propias de producción (factor productivo: número de veces que se le da de comer a los peces, tamaños de jaulas, genética de los peces, etc.).

La finalidad de este trabajo de investigación es determinar si las variables con las que actualmente se tiene información para el peso de las especies marinas pueden mejorar la producción del alimento mediante estrategias de negocio o hay algún otro tipo de variables de otro factor desconocido que la empresa no está tomando en cuenta para la producción, si el caso es lo segundo se debería utilizar algún otro tipo de base de datos adicional como son variables del factor económico como por ejemplo el índice de inflación del país, aunque, se debería ver en primer lugar si este tipo de variables del factor económico son muy determinantes para explicar la producción de especies marinas de la empresa Vitapro.

1.3. OBJETIVO

Guardando relación con lo señalado en el problema central y los problemas específicos del problema de la investigación, se establecen los siguientes objetivos:

1.3.1. Objetivo General

Construir el modelo de predicción y el que mejor se ajuste a la curva de crecimiento de Gompertz para los peces de los centros de cultivos de empresas dedicadas a la crianza de especies marinas.

1.3.2. Objetivos Específicos

- Determinar cómo influyen las variables de los factores ambientales en el crecimiento de peces, se realizará análisis exploratorio de los datos.
- Determinar cómo influyen las variables de los factores de producción en el crecimiento de peces.
- Determinar las variables más influyentes con respecto al peso promedio final de los peces.
- Encontrar el mejor modelo de predicción que mejor se ajuste al peso promedio final de las especies marinas.
- Encontrar cuál de los modelos de predicción empleados se ajusta a la curva de crecimiento de Gompertz para las especies marinas.

1.4. PLANTEAMIENTO DE LA HIPÓTESIS

En este trabajo de investigación, nos enfocamos en construir y encontrar una curva de predicción de crecimiento de peces para la empresa Vitapro, haciendo el uso de comparación de indicadores de modelos y análisis de variables del factor ambiental y el factor de producción. En consecuencia, estos factores son los que pueden influir en crecimiento de peces.

1.4.1. Hipótesis General

El mejor modelo de predicción que se ajuste a la curva de crecimiento de Gompertz para las especies marinas, así como la importancia de las variables de los factores productivos y ambientales influyentes en el crecimiento de especies marinas en la empresa Vitapro.

1.4.2. Hipótesis Específicas

- Los variables de los factores productivos influyen en el crecimiento de peces en la empresa Vitapro.
- Los variables de los factores ambientales influyen en el crecimiento de peces en la empresa Vitapro.
- El peso inicial y la alimentación que posee cada unidad de análisis (jaula por periodo) está relacionada potencialmente con el peso final de cada unidad de análisis.
- Los modelos de Regresión Lineal, Árbol de Decisión, Random Forest y XGboost son adecuados para predecir los pesos promedios de las especies marinas.
- Los modelos de Regresión Lineal, Árbol de Decisión, Random Forest y XGboost se ajustan a la curva de crecimiento de Gompertz para las especies marinas.

2. FUNDMENTOS TEÓRICOS

2.1.ANTECEDENTES DE LA INVESTIGACIÓN

Se presentan los fundamentos teóricos que sustentan los métodos y análisis realizados durante todo el trabajo de investigación, con este fin se obtuvo información de los siguientes antecedentes (3 internacionales y 3 nacionales):

AGUDELO D.A. (2007). "Modelación de funciones de crecimiento aplicadas a la producción animal", Universidad de Antioquia, Colombia.

Este artículo tiene como objetivo indicarle al lector la aplicación de los modelos no lineales y no lineales mixtos en el análisis del crecimiento animal, ya que este puede ser descrito por medio de funciones matemáticas que predicen el desempeño de la evolución del peso vivo, dichas funciones permiten realizar evaluaciones sobre el nivel de producción en las empresas ganaderas, pudiendo clasificar de forma sencilla la productividad de una raza específica para una zona determinada. También permiten calcular los valores máximos de los crecimientos medio y corriente, pudiendo determinar las edades de sacrificio que permitan obtener el máximo beneficio económico. Además, proveen información que permite realizar programaciones de alimentación, de capacidad de carga y medir cambios genéticos de una generación a otra que estén relacionados con el nivel de producción.

El desconocimiento de las curvas de crecimiento y de parámetros productivos de interés económico, ha limitado la implementación de programas de mejoramiento zootécnico que permitan aumentar la productividad, como lo son la velocidad de crecimiento, la tasa de madurez a diferentes edades y la edad al sacrificio. Estos factores se pueden analizar con base en la información zootécnica de los animales siendo indispensable para ello contar con registros de producción.

Conclusión: En conclusión, el crecimiento animal es uno de los aspectos más importantes al momento de evaluar la productividad en las explotaciones dedicadas a la producción y en algunos casos es usado como criterio de selección. Para medir el crecimiento animal se han usado diferentes modelos matemáticos lineales, no lineales y logarítmicos, entre otros, eligiéndolos por su bondad de ajuste. Los criterios más usados para elegir la curva que mejor ajusta a los datos son: el coeficiente de determinación, el porcentaje de curvas significativas y atípicas encontradas para cada función; además se pueden aplicar criterios como el criterio de información Akaike y el criterio de información Bayesiano.

CASTILLO T. (2008). "Modelos matemáticos en la evaluación del crecimiento de vaquillas cruzadas en clima cálido húmedo y su caracterización productiva a primera gestación", Universidad Veracruzana, México.

En esta investigación presentan el crecimiento de los animales domésticos utilizando distintas funciones matemáticas que relacionan el cambio del peso en función de la edad de los animales. Con el objetivo de determinar la curva de crecimiento de los diferentes genotipos de vaquillas en un sistema de doble propósito y, ubicar en ella la edad y peso a la que se realiza la primera gestación, utilizan los modelos Logístico y Gompertz con el fin de obtener modelos de predicción para el comportamiento reproductivo de futuras becerras validándolos con los datos obtenidos en campo.

Conclusión: En conclusión la cuantificación del crecimiento en peso hasta una edad determinada bajo condiciones óptimas de manejo, alimentación, control sanitario y condiciones climáticas en las cuales se ha desarrollado una raza bovina, constituye el patrón de crecimiento normal de la raza o grupo genético; estos patrones de crecimiento pueden ser utilizados como referencia para determinar la eficiencia productiva y reproductiva de un grupo o grupos específicos de animales con respecto al patrón de la raza; también pueden utilizarse para determinar o predecir si un animal es excepcional o si es necesario adoptar medidas correctivas en la explotación. Los resultados mostraron que los ajustes de los modelos Logistic y Gompertz representan de manera eficiente el comportamiento de los pesos en las diferentes edades de los grupos de vaquillas.

LUQUIN M. (2016). "Función de verosimilitud conjunta basada en distribuciones de probabilidad normal y multinomial para analizar la variabilidad fenotípica en el crecimiento de almeja de sifón panopea globosa", Centro de Investigaciones Biológicas del Noroeste, México.

En este artículo de investigación se quiere analizar la variabilidad fenotípica individual de la Panopea globosa, para ello se analizan la longitud de la concha y la edad de la Panopea. Para ello se emplearon estudio de datos de frecuencias de la longitud mediante la función negativa de verosimilitud conjunta, luego estos fueron ajustados mediante modelos de crecimiento como el de Gompertz, Von Bertalanffy, Jhonson, Logístico y Richards; estos modelos incluyeron una estimación de la varianza para cada edad observada. Para seleccionar el mejor modelo se basaron en el criterio de información de Akaike (AIC), el modelo que crecimiento más adecuado fue el de Jhonson.

Conclusión: En conclusión, el uso de la función negativa de verosimilitud que incluye la variabilidad individual de la longitud de concha a la edad, mostró ser adecuado para analizar el crecimiento de la P. globosa, el uso de la función de verosimilitud conjunta proporcionó mayor información sobre el parámetro t_0 que fue usado en los modelos de crecimiento. El modelo Von Bertalanffy fue el que mostró mayor variabilidad y no fue el más adecuado; incluir este análisis de variabilidad permitió demostrar que la población de la P. globosa presenta un crecimiento despensatorio, lo que indica que la variabilidad se denota mayor en los individuos más viejos.

ALDAVA J. (2009). "Evaluación de la densidad de cultivo del hibrido (piaractus brachypomus ♀ x colossoma macropomum ♂) "pacotana" en sistema semi-intensivo en selva alta", Universidad Nacional Agraria de la Selva, Perú.

En este trabajo de investigación estudian el crecimiento del pez híbrido "Pacotana" en la Selva del Perú (Tingo María) mediante el análisis de la densidad (peces/m2) de espejo de agua por tres tratamientos, cada uno de estos tenía variaciones en sus variables de condición ambiental y propias de la especie. Para ello hacen uso principalmente de las variables peso y longitud del pez en su etapa de alevino, asimismo estudia otras variables

ambientales del agua como la salinidad, oxigenación, pH, etc. Finalmente hace contrastes con otros estudios que plantea en su investigación como por ejemplo el tiempo de vida de los alevines, el número de veces que se alimentan al pez, etc.

Conclusión: Finalmente, el autor elige que la óptima densidad en sistema de cultivo semi-intensivo en selva alta sembrados en fase de alevino en base a los méritos productivos y económicos es de 2 peces/m2 (tratamiento II) de espejo de agua. Aunque en los otros tratamientos también se observaron otros indicadores; por ejemplo: la ganancia de peso, ganancia de longitud y la velocidad de crecimiento en peso fue superior para el tratamiento I (1 pez/m2 de espejo de agua); la tasa de crecimiento específico en peso, tasa de crecimiento específico en longitud, factor de conversión alimenticia, eficiencia alimentaria son superiores en el tratamiento I (1 pez/m2 de espejo de agua), sin embargo, el factor de condición mostro superioridad en el tratamiento II (2 peces/m2 de espejo de agua), y el rendimiento del cultivo mostro mejor resultado en el tratamiento III (3 peces/m2 de espejo de agua).

MONTENEGRO K. (2016). "Edad y crecimiento de ethmidium maculatum (machete) desembarcado en las caletas de la región La Libertad durante el año 2012 - 2013", Universidad Nacional de Trujillo, Perú.

En este documento como primer punto se determinó las edades de la especie Ethmidium maculatum que se capturaron en la Región de la Libertad durante los años 2012 y 2013. Esta edad se determinó por los anillos de crecimiento de los otolitos sagita; asimismo, se separaron muestras por género. Se estimaron parámetros mediante el modelo de crecimiento de Von Bertalanffy, con las variables talla-edad lo que permitió agrupas a las especies en cinco grupos teniendo como mayor representatividad los individuos que tenían entre 2 y 3 años.

Conclusión: Finalmente, fue muy adecuado tomar en cuenta la estimación de los otolitos sagita para saber la edad de la especie Ethmidium maculatum, asimismo se determinó que para hembras y machos no presentó diferencias estadísticas en la edad por lo que se pudo describir el crecimiento considerando ambos géneros, esto es al usar el modelo Von Bertalanffy. Luego de analizar el crecimiento de los otolitos y asignación de edades se tuvo como resultado cinco grupos de edad, donde se obtuvo que los que tienen entre 2 y 3 años tienen mayor representatividad.

PARDO M. (2015). "Crecimiento de mugil cephalus (lisa) en agua dulce en el valle del medio Piura en los años 2013 y 2015", Universidad Nacional de Piura, Perú.

En esta investigación se tuvo como objetivo saber la tasa de crecimiento de la lisa en agua dulce en el Valle del Medio Piura, para ello se tuvo un estanque donde se dividieron en dos zonas con diferentes condiciones de clima, alimentación y densidad de agua. Adicionalmente luego de la división del estanque se evaluaron indicadores con respecto al agua (temperatura, concentración de oxígeno y pH).

Conclusión: Finalmente, el autor halló que la tasa de crecimiento en peso promedio es del 2.83% diario y la talla promedio es del 0.46%, esto a condiciones de pH entre 8 y 10.5. Asimismo, se aclimató al Mugil cephalus a condiciones de salinidad de 305 a 0% obteniendo 100% de sobrevivencia.

2.2. MÉTODOS ESTADÍSTICOS

2.2.1. Regresión Lineal Múltiple

Este modelo es una extensión del modelo de regresión lineal simple, en donde se tenía que una variable dependiente influye solo una variable independiente. Por lo general este caso en la práctica son escasos, por lo que se generaliza en donde hay más de una variable independiente que influyen sobre una dependiente, a lo cual se denomina regresión lineal Múltiple. En general se puede relacionar la variable respuesta y con k variables independientes $x_1, x_2, x_3, \dots, x_k$, en ese caso el modelo está dado por:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \qquad \text{…….(1)}$$

Donde los coeficientes β_j, $j = 0, 1, 2, \dots, k$ son constantes desconocidos y son los parámetros del modelo. Cada β_j , representa el cambio esperado en la respuesta y por el cambio unitario en x_j cuando todas las demás variables independientes $x_i (i \neq j)$ se mantienen constantes. ϵ es un componente de error aleatorio.

En el caso de los modelos de regresión múltiple es preferible usar la notación matricial, pues dicha forma permite expresar el modelo en una forma más compacta y que con un poco de conocimiento del álgebra matricial los resultados se simplifican considerablemente.

Forma Matricial: El modelo de Regresión Lineal Múltiple en su forma matricial es la siguiente:

$$y = X\beta + \epsilon, \qquad \text{…….(2)}$$

$Donde$:

- y: es un vector n × 1 observable.
- X: es una matriz n × p que contiene los valores de las variables independientes.
- β: es un vector p × 1 de parámetros no observables.
- ϵ: es un vector n×1 de variables aleatorias no observables conocido como el vector de errores aleatorios.

Si se reescriben los vectores y las matrices en la ecuación anterior, se obtiene:

$$y = \begin{bmatrix} y_1 1 x_{12} \cdots x_{1k} \\ y_2 1 x_{22} \cdots x_{2k} \\ \vdots \ \vdots \ \vdots \ \ddots \ \vdots \\ y_n 1 x_{n2} \cdots x_{nk} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_k \end{bmatrix} \qquad \text{…….(3)}$$

Estimación de parámetros del modelo:

Estimación del β:

El estimador de mínimos cuadrados de β, denotado por $\hat{\beta}$, es el valor de β que minimiza:

$$S(\beta) = \sum_{I=1}^{n} \epsilon_i{}^2 = \epsilon\epsilon' = (y - X\beta)'(y - X\beta) \qquad \text{…….(4)}$$

Por lo tanto, lo que se debe hacer es derivar la expresión anterior y buscar el valor de β que la hace igual a cero. Antes de derivar note que la expresión anterior se puede escribir como:

$$S(\beta) = yy' - \beta'X'y - y'X\beta + X'\beta'\beta X$$
$$= yy' - 2\beta'X'y + X'\beta'\beta X \qquad \text{…….(5)}$$

Ahora si derivando e igualando a cero se obtiene:

$$\frac{\delta S}{\delta \beta}|_{\hat{\beta}} = -2X'y + X'X\hat{\beta} = 0 \qquad \text{…….(6)}$$

que se simplifica a:

$$X'X\hat{\beta} = X'y \qquad \text{…….(7)}$$

las cuales se conocen como las ecuaciones normales de mínimos cuadrados. Para hallar la expresión de β se premultiplica la ecuación anterior por la inversa de $X'X$ (que en este caso se asume que existe). Por lo tanto, el estimador de β por mínimos cuadrados es:

$$\hat{\beta} = (X'X)^{-1}X'y \qquad \text{…….(8)}$$

Estimación del σ^2:

Al igual que en el caso de la regresión lineal simple, el estimador de σ^2 se puede obtener a partir de la suma de cuadrados de los residuales:

$$SC_{Res} = \sum_{I=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{I=1}^{n} r_i{}^2 = r'r \qquad \text{…….(9)}$$

Sustituyendo $r = y - X\hat{\beta}$ se obtiene:

$$SC_{Res} = (y - X\hat{\beta})'(y - X\hat{\beta})$$
$$= y'y - \hat{\beta}'X'y - y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta}$$
$$= y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta} \qquad \text{…….(10)}$$

Como $X'X\hat{\beta} = X'y$, la última ecuación se transforma en:

$$SC_{Res} = y'y - \hat{\beta}'X'y \qquad \text{…….(11)}$$

la cual tiene $n - p$ grados de libertad (pues hay que p parámetros en el modelo de regresión múltiple). Por lo tanto, el cuadrado medio del residual es:

$$CM_{Res} = \frac{SC_{Res}}{n-p} \quad \text{…….(12)}$$

cuyo valor esperado es σ^2. Entonces, un estimador insesgados de σ^2, denotado por $\hat{\sigma}^2$ es:

$$\hat{\sigma}^2 = CM_{Res} \quad \text{…….(13)}$$

Propiedades de los estimadores

1. Son estimadores insesgados. En la demostración anterior se probó que $\hat{\sigma}^2$ es un estimador insesgados de σ^2. Por lo tanto, solo falta probar con $\hat{\beta}$.

   $$E(\hat{\beta}) = E((X'X)^{-1}X'y) = (X'X)^{-1}X'E(y)$$
   $$= (X'X)^{-1}E(X\beta + \epsilon)$$
   $$= (X'X)^{-1}X'X\beta = \beta \quad \text{…….(14)}$$

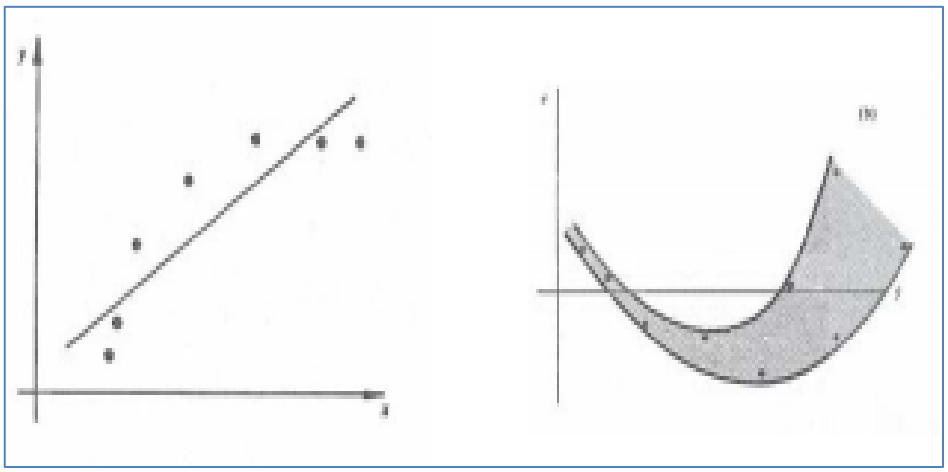
2. $Cov(\hat{\beta}) = \sigma^2(X'X)^{-1}$.
3. $\hat{\beta}$ y $\hat{\sigma}^2$ son independientes.
4. Si se supone que los errores son normales se tiene que $\hat{\beta}$ también se distribuye normal y que una función de $\hat{\sigma}^2$ se distribuye chi cuadrado. Además $\hat{\beta}$ y $\hat{\sigma}^2$ son los estimadores de máxima verosimilitud.

Supuestos del modelo de regresión lineal

Para realizar un modelo de regresión lineal, es necesario que se cumplan cinco supuestos muy importantes, si alguno de ellos no cumple se aplican algunos métodos para que se pueda utilizar este modelo, sobre todo en el supuesto de Normalidad, Homocedasticidad e Independencia.

- **Linealidad**, si no se tiene linealidad se dice que tenemos un error de especificación. En el caso de que sean varias variables independientes, todos los gráficos parciales nos da los diagramas de dispersión parcial para cada variable independiente. En ellos se ha eliminado el efecto proveniente de las otras variables y así la relación que muestran es la relación neta entre las variables representadas. A continuación, se muestran gráficas que no cumplen este supuesto:

FIGURA N° 2: Regresiones que no cumplen el supuesto de Linealidad



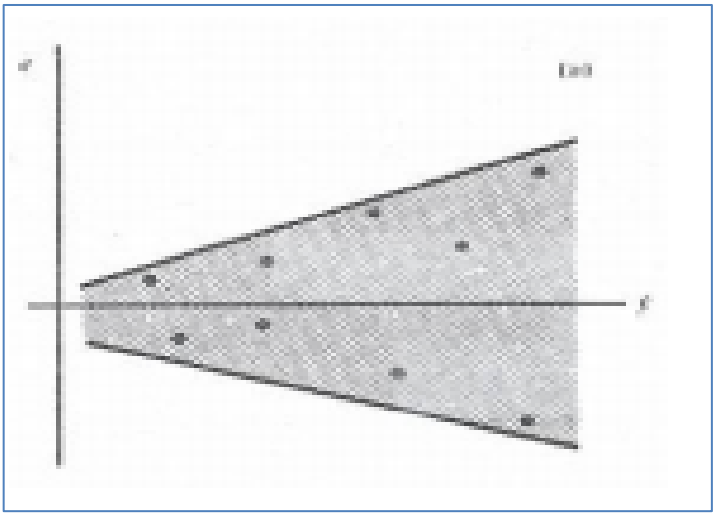
Fuente: Materiales del Departamento de Estadística e Investigación Operativa – USC.

- **Independencia**, de la variable aleatoria "residuos" (especialmente importante si los datos se han obtenidos siguiendo una secuencia temporal). Independencia entre los residuos mediante el estadístico de Durbin-Watson que toma valor 2 cuando los residuos son completamente independientes (entre 1.5 y 2.5 se considera que existe independencia), $DW < 2$ autocorrelación positiva y $DW > 2$ autocorrelación negativa.

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} {e_i}^2}, 0 \leq DW \leq 4 \qquad \text{…….(15)}$$

- **Homocedasticidad**, o igualdad de varianzas de los residuos y los pronósticos. Esta condición se estudia utilizando las variables: pronósticos tipificados y residuos tipificados mediante: el estadístico de Levene o un gráfico de dispersión. Este supuesto de homocedasticidad implica que la variación de los residuos sea uniforme en todo el rango de valores de los pronósticos. A continuación, se muestra una gráfica donde no presenta homocedasticidad.

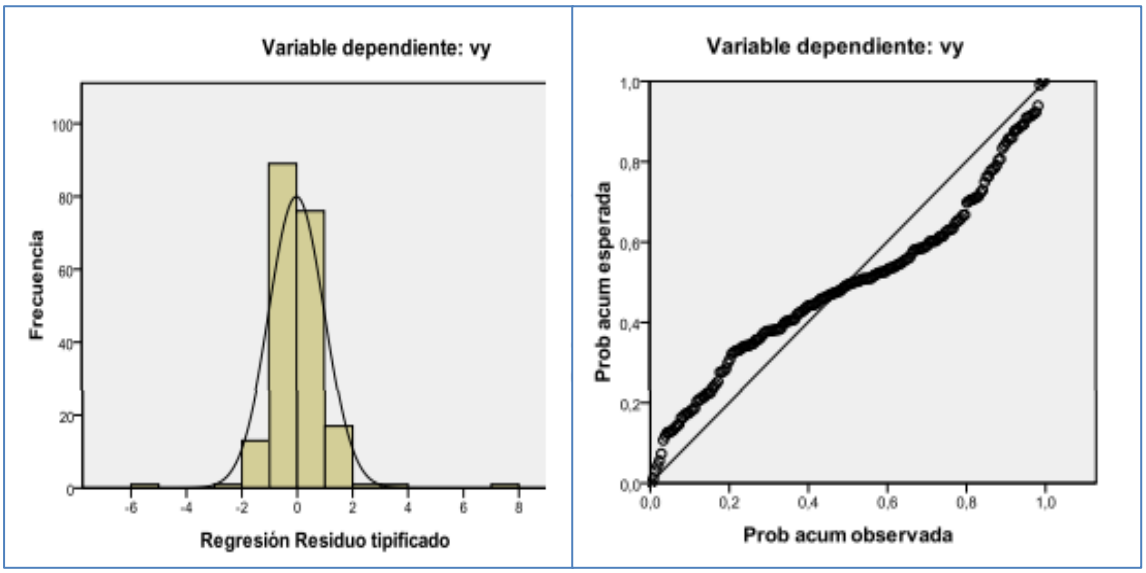
FIGURA N° 3: Regresión que no cumple supuesto de Homocedasticidad



Fuente: Materiales del Departamento de Estadística e Investigación Operativa – USC.

- **Normalidad**, de los residuos tipificados. Podemos contrastarla mediante: la prueba de Kolmogorov -Smirnov, con gráficos de normalidad de tipo Q-Q (cuantiles) o P-P(proporciones). Gráficamente en un histograma se añade una curva $N \sim (0,1)$ o un gráfico de Probabilidad Normal de tipo P-P, donde se representa las proporciones acumuladas de la variable respecto a las proporciones acumuladas de la variable observada. A continuación, mostramos la Figura N° 4 como ejemplo el histograma con la curva Normal y la gráfica P-P (proporciones) normal de regresión en donde se puede observar que los residuos tipificados no se ajustan a la distribución normal, lo cual puede ser resuelta utilizando una Transformación Box-Cox.

FIGURA N° 4: Histograma y Gráfica (P-P) de regresión que no cumple el supuesto de Normalidad



Fuente: Materiales del Departamento de Estadística e Investigación Operativa – USC.

- **No-colinealidad**, es decir la inexistencia de colinealidad, esta puede ser: colinealidad perfecta si una de las variables independientes tiene una relación lineal con otra/as independientes, colinealidad parcial si entre las variables independientes existen altas correlaciones. Algún método para detectar la multicolinealidad es el Factor de inflación de la varianza.

$$FIV(\beta_i) = \frac{k-2}{1-{R_{x_i}}^2} \leftrightarrow T(\beta_i) = 1 - {R_{x_i}}^2 \qquad \text{…….(16)}$$

De manera que podemos contrastar con la Hipótesis si existe o no colinealidad para la i-ésima variable, habrá colinealidad cuando:

$$FIV(\beta_i) = \frac{k-2}{n-k+1} F_0 + 1 \qquad \text{.....(17)}$$

Siendo F_0 el cuantil de distribución F-Snedecor con $k-2$ y $n-k+1$ grados de libertad.

2.2.2. Árbol de decisión y clasificación

El algoritmo CART es el acrónimo de Classification And Regression Trees (Árboles de Clasificación y de Regresión) fue diseñado por Breiman et al. (1984). Este modelo admite variables de entrada y de salida nominales, ordinales y continuas, por lo que se pueden resolver tanto problemas de clasificación como de regresión. Los árboles de decisión o clasificación no son considerados como modelos estadísticos basados en la estimación de los parámetros de la ecuación propuesta, por tanto, no tenemos que estimar un modelo estadístico formal.

En los árboles de decisión se usa la segmentación jerárquica, esta técnica explicativa y descomposicional utiliza un proceso de división secuencial, iterativo y descendente que, partiendo de una variable dependiente, forma grupos homogéneos definidos específicamente mediante combinaciones de variables independientes en las que se incluyen la totalidad de los casos recogidos en la muestra.

Suponemos que se dispone de una muestra de entrenamiento que incluye la información del grupo al que pertenece cada caso y que sirve para construir el criterio de clasificación. Se comienza con un nodo inicial, dividiendo la variable dependiente a partir de una partición de una variable independiente que se escoge de modo tal que dé lugar a dos conjuntos homogéneos de datos (que maximizan la reducción en la impureza). Se elige, por ejemplo, la variable x_1 y se determina un punto de corte, por ejemplo c, de modo que se puedan separar los datos en dos conjuntos: aquellos con $x_1 \leq c$ y los que tienen $x_1 > c$. De este nodo inicial saldrán ahora dos: uno al que llegan las observaciones con $x_1 \leq c$ y otro al que llegan las observaciones con $x_1 > c$. En cada uno de estos nodos se vuelve a repetir el proceso de seleccionar una variable y un punto de corte para dividir la muestra. El proceso termina cuando se hayan clasificado todas las observaciones correctamente en su grupo.

En los árboles de decisión se encuentran los siguientes componentes: nodos, ramas y hojas. Los nodos son las variables de entrada, las ramas representan los posibles valores de las variables de entrada y las hojas son los posibles valores de la variable de salida. Como primer elemento de un árbol de decisión tenemos el nodo raíz que va a representar la variable de mayor relevancia en el proceso de clasificación. Todos los algoritmos de aprendizaje de los árboles de decisión obtienen modelos más o menos complejos y consistentes respecto a la evidencia, pero si los datos contienen incoherencias, el modelo se

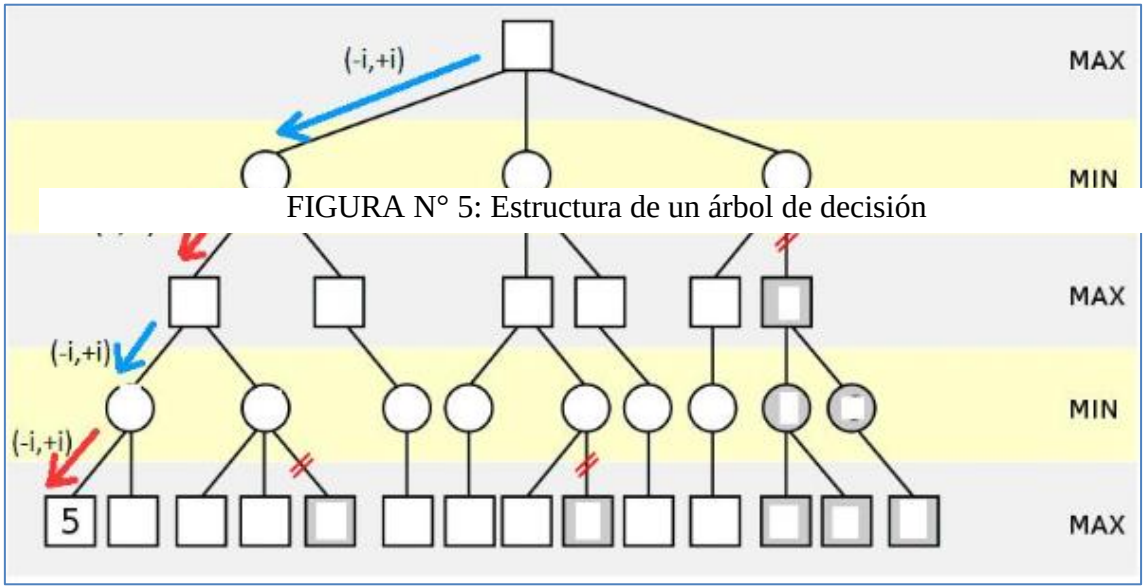


FIGURA N° 5: Estructura de un árbol de decisión

ajustará a estas incoherencias y perjudicará su comportamiento global en la predicción, es lo que se conoce como sobreajuste. Para solucionar este problema hay que limitar el crecimiento del árbol modificando los algoritmos de aprendizaje para conseguir modelos más generales. Es lo que se conoce como poda en los árboles de decisión.

Fuente: Materiales de Trabajo Arboles de decisión y Random Forest – UCUENCA.

Las reglas de parada tratan de preguntar si merece la pena seguir o detener el proceso de crecimiento del árbol por la rama actual, se denominan reglas de prepoda ya que reducen el crecimiento y complejidad del árbol mientras se está construyendo:

- **Pureza de nodo.** Si el nodo solo contiene ejemplos o registros de una única clase se decide que la construcción del árbol ya ha finalizado.
- **Cota de profundidad.** Previamente a la construcción se fija una cota que nos marque la profundidad del árbol, cuando se alcanza se detiene el proceso.
- **Umbral de soporte**. Se especifica un número de ejemplos mínimo para los nodos, y cuando se encuentre un nodo con ejemplos por debajo del mínimo se para el proceso, ya que no consideramos fiable una clasificación abalada con menos de ese número mínimo de ejemplos.

El algoritmo utiliza el índice de Gini para calcular la medida de impureza:

$$G(A_i) = \sum_{j=i}^{M_i} p(A_{ij}) * G(C/A_{ij}) \quad \text{…….(1)}$$

Siendo, $G(C/A_{ij})$ igual a:

$$G(C/A_{ij}) = -\sum_{j=i}^{M_i} p(C_k/A_{ij}) * (1 - p(C_k/A_{ij})) \quad \text{…….(2)}$$

- A_{ij}: es el atributo empleado para ramificar el árbol.
- j: es el número de clases.
- M_i: es el de valores distintos que tiene el atributo A_i.
- $p(A_{ij})$: constituye la probabilidad de que A_i tome su $j - ésimo$ valor.
- $p(C_k/A_{ij})$: representa la probabilidad de que un ejemplo sea de la clase $C_k$cuando su atributo A_i toma su $j - ésimo$ valor.

El índice de diversidad de Gini toma el valor cero cuando un grupo es completamente homogéneo y el mayor valor lo alcanza cuando todas las $p(A_{ij})$ son contantes, entonces el valor del índice es $\frac{(J-1)}{J}$.

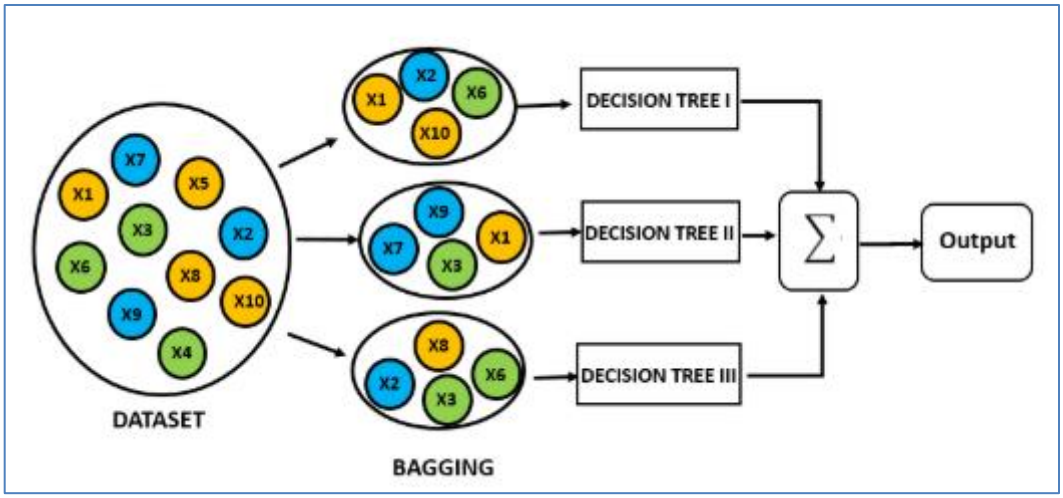
2.2.3. Random Forest

Random Forest es útil para regresión y clasificación, es la combinación de muchos árboles de decisión y una técnica para reducir las dimensiones, en cada árbol da una clasificación (vota por una clase) y el resultado es la clase con mayor número de votos en todo el bosque (forest); para el caso de una regresión, se toma el promedio de las salidas (predicciones) de

todos los árboles. La construcción de árboles en el Random Forest se realiza de la siguiente manera:

- Dado que el número de casos en el conjunto de entrenamiento es N. Una muestra de esos N casos se toma aleatoriamente, pero **con reemplazo**, esta muestra será el conjunto de entrenamiento para construir el árbol i.
- Si existen M variables de entrada, un número $m < M$ se especifica tal que, para cada nodo, m variables se seleccionan aleatoriamente de M. La mejor división de estos m atributos es usado para ramificar el árbol. El valor m se mantiene constante durante la generación de todo el bosque, de tal manera que cada árbol crece hasta su máxima extensión posible y **no hay proceso de poda**.
- Finalmente, nuevas instancias se predicen a partir de la agregación de las predicciones de los x árboles (i.e., mayoría de votos para clasificación, promedio para regresión).

FIGURA N° 6: Proceso de Creación de un Random Forest



Fuente: Materiales de Trabajo Arboles de decisión y Random Forest – UCUENCA.

En Random Forest es importante revisar el proceso de muestreo de los datos con reemplazo, el cual se denomina **bootstrap**. Por ejemplo, un tercio de los datos no se usan para el entrenamiento y pueden ser usados para test, este conjunto se denomina muestras out of bag (OOB).

FIGURA N° 7: Proceso Bootstrap en Random Forest

Fuente: Materiales de Trabajo Arboles de decisión y Random Forest – UCUENCA.

El error estimado en estas muestras out of bag se conoce como out of bag error (OOB error), usar este conjunto de test (OOB) es tan preciso como si se usara un conjunto de test del mismo tamaño que el de entrenamiento. Asimismo, sería posible no usar un conjunto de test adicional.

2.2.4. XGBoost

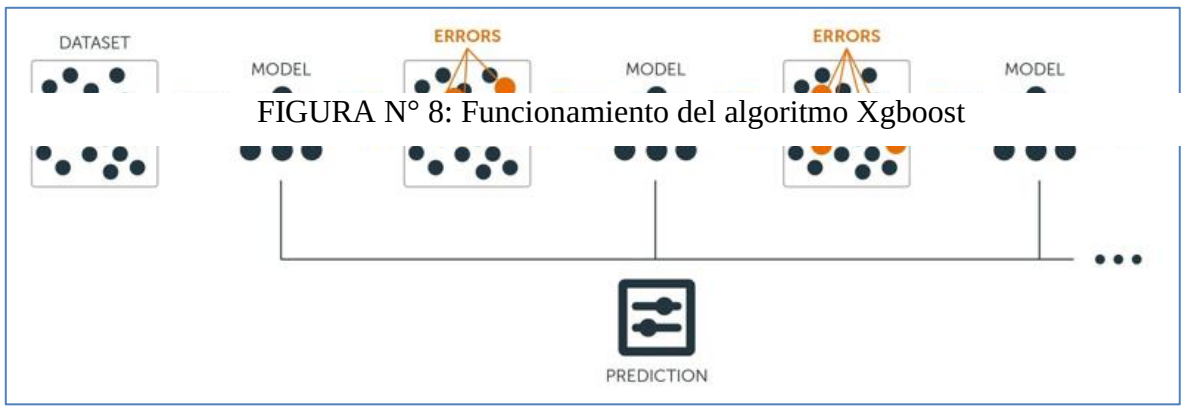
Boosting es una familia de técnicas de ensamble empleada habitualmente, en Boosting se entrenan modelos utilizando todos los datos de entrenamiento y cada modelo se entrena teniendo en cuenta la salida de los modelos anteriores. Intuitivamente, podemos entender boosting como un método que va aprendiendo de los datos "lentamente". Se ha demostrado que los modelos estadísticos de este tipo, que aprenden "progresiva y lentamente" tienen muy buen performance. Además, la predicción del Boosting es la suma ponderada de las predicciones de los modelos.

Los algoritmos de tipo Boosting tienen los siguientes hiperparámetros principalmente:

- El número de modelos es crítico, ya que, si es muy grande, el modelo puede llegar a hacer overfitting. Una forma de escoger el número óptimo es con cross-validation.

- Un parámetro $\boldsymbol{\lambda}$ que determina la "velocidad" a la que aprende el modelo. Típicamente se usan valores entre 0.01 y 0.001. Valores menores de $\boldsymbol{\lambda}$ requieren de

FIGURA N° 8: Funcionamiento del algoritmo Xgboost



valores altos del número de modelos para llegar a un buen performance. (este parámetro solo existe en algunas implementaciones).

Fuente: Materiales de la Facultad de Ciencias Económicas y Empresariales – UV.

El XGboost (Extra Gradient boosting) es uno de los algoritmos más usados, el cual está referido a la potenciación de gradientes, este es un algoritmo de aprendizaje supervisado que intenta predecir de forma apropiada una variable objetivo mediante la combinación de estimaciones de un conjunto de modelos más simples y débiles. Cabe resaltar que el refuerzo de gradiente es una técnica de aprendizaje automático para problemas de regresión y clasificación, que produce un modelo de predicción en forma de un conjunto de modelos de predicción débiles, típicamente árboles de decisión. Construye el modelo de manera escalonada como lo hacen otros métodos de refuerzo, y los generaliza permitiendo la optimización de una función arbitraria de pérdida diferenciable.

Cuando se utiliza la potenciación del gradiente para la regresión, los aprendices débiles son árboles de regresión y cada árbol de regresión mapea un punto de datos de entrada en una de sus hojas que contiene una puntuación continua. XGBoost minimiza una función de objetivo regularizada (L1 y L2) que combina una función de pérdida convexa (según la diferencia entre las salidas de destino y previstas) y un plazo de penalización para la complejidad de modelos (es decir, las funciones de árboles de regresión). La capacitación avanza de forma iterativa, agregando nuevos árboles que predicen los residuos de errores de los árboles anteriores que se combinan después con los árboles anteriores para realizar la predicción final. Se denomina potenciación del gradiente porque utiliza un algoritmo de gradiente descendente para minimizar la pérdida cuando se agregan nuevos modelos. En conclusión, el modelo o algoritmo XGBoost reduce la velocidad de ejecución y maximiza el rendimiento.

2.2.5. Modelo de Crecimiento de Gompertz

El modelo de Gompertz es uno de los modelos sigmoideos utilizados con mayor frecuencia adaptados a los datos de crecimiento y otros datos, tal vez solo por detrás del modelo logístico (también llamado modelo Verhulst), muchos investigadores han adaptado el modelo de Gompertz a todo, desde el crecimiento de las plantas, el crecimiento de las aves, el crecimiento de los peces y el crecimiento de otros animales, hasta el crecimiento de tumores y el crecimiento bacteriano, etc. El modelo de Gompertz es un caso especial del modelo Richards de cuatro parámetros y, por lo tanto, pertenece a la familia Richards de modelos de crecimiento sigmoidal de tres parámetros, junto con modelos familiares como el exponencial negativo (incluido el Brody), el logístico y el Von Bertalanffy (o solo Bertalanffy.

El modelo Gompertz ha estado en uso como modelo de crecimiento incluso más tiempo que el modelo logístico. El modelo, referido en ese momento como la ley teórica de

mortalidad de Gompertz, fue sugerido por primera vez y aplicado por primera vez por el Sr. Benjamin Gompertz en 1825. Lo ajustó a la relación entre el aumento de la tasa de mortalidad y la edad, a lo que se refirió como "los agotamientos promedio del poder de un hombre para evitar la muerte", o la "porción de su poder restante para oponerse a la destrucción". La industria de seguros rápidamente comenzó a usar su método de proyectar el riesgo de muerte. Sin embargo, Gompertz solo presentó la función de densidad de probabilidad.

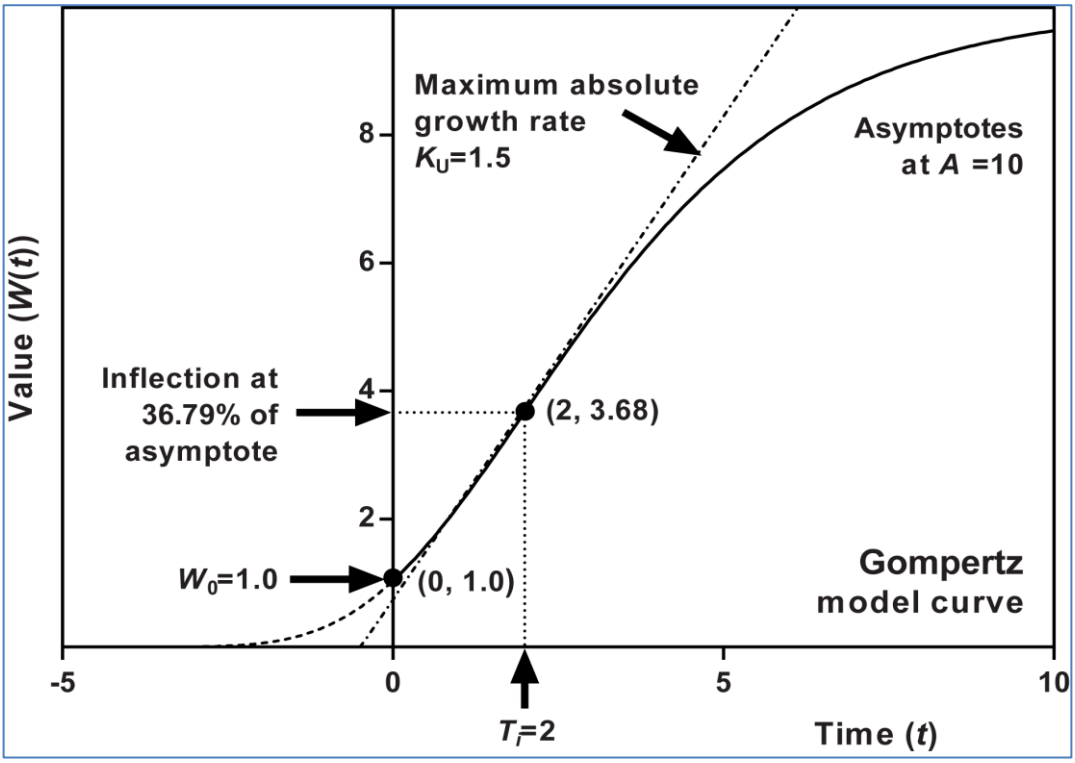
Fue Makeham quien declaró por primera vez este modelo en su conocida forma acumulativa, y por lo tanto se hizo conocido como el modelo Gompertz-Makeham. Se intentó utilizar un método de mínimos cuadrados para el modelo de Gompertz y así encontrar la mejor curva. Sin embargo, no linealizaron el modelo, como se hace más adelante, sino que solo transformaron los valores (variable dependiente) para facilitar la determinación de la suma de los cuadrados. Este método parece haberse utilizado hasta la década de 1940, cuando Hartley propuso y primero explicó cómo linealizar el modelo de Gompertz.

A partir de la década de 1920, el modelo acumulativo de Gompertz-Makeham también se convirtió rápidamente en un favorito en otros campos además de la mortalidad humana, por ejemplo, al pronosticar el aumento de la demanda de bienes y servicios, las ventas de tabaco, el crecimiento del tráfico ferroviario y la demanda de automóviles. Wright fue el primero en proponer el modelo de Gompertz para el crecimiento biológico, y el primero en aplicarlo a los datos biológicos fue probablemente Davidson en su estudio del crecimiento de la masa corporal en el ganado. En 1931, Weymoth, McMillin y Rich informaron sobre el modelo de Gompertz para describir con éxito el crecimiento del tamaño de una concha en navajas, Siliqua patula y Weymouth y Thompson informó lo mismo para el berberecho del Pacífico, Cardium corbis. Pronto, los investigadores comenzaron a ajustar el modelo a sus datos por regresión, y con el paso de los años, el modelo Gompertz común se convirtió en un modelo de regresión favorito para muchos tipos de crecimiento de organismos, como los dinosaurios aves y mamíferos, incluidos los de marsupiales. El modelo de Gompertz también se aplica con frecuencia al crecimiento del modelo en número o densidad de microbios, el crecimiento de tumores y la supervivencia de pacientes con cáncer.

Se utilizan varias re-parametrizaciones diferentes del modelo tradicional acumulativo de Gompertz. Uno de los más importantes fue sugerido por Zwietering y colegas para modelar el crecimiento en el número de bacterias, y actualmente es uno de los modelos más comunes en el crecimiento microbiano. Otra re-parametrización prominente del modelo de Gompertz es el modelo de Gompertz-Laird, propuesto por Laird y ajustado a los datos de crecimiento tumoral. Sin embargo, los parámetros del modelo no son fácilmente interpretables sin convertirse en mediciones más útiles.

A continuación, revisamos los modelos de Gompertz, centrándonos en cómo sus parámetros afectan las características de la curva (Ilustración 4.). Se presentan los modelos utilizando una notación típica para estudios de crecimiento de organismos, describiendo mediciones biométricas como funciones del tiempo; $W(t)$. Varios campos usan diferentes notaciones, para el valor medido, por ejemplo, supervivencia: $S(t)$, número de células / bacterias o tamaño de la población: $N(t)$, densidad de células o microorganismos; $D(t)$, concentración de organismos $C(t)$, volumen $V(t)$, masa corporal: $M(t)$ y longitud: $L(t)$. La variable dependiente (lado izquierdo de la ecuación) también puede expresarse como valores relativos, por ejemplo, como $W(t)/A$, donde A es la asíntota superior, o $W(t)/W_0$, donde W_0 es el valor inicial (o punto de partida en el eje x). Este último representa el valor relativo al valor inicial (descrito como una medición "adimensional"). A veces, la variable dependiente se transforma logarítmicamente, en particular cuando se modela el

FIGURA N° 9: Características de la forma del modelo de Gompertz (línea continua)



crecimiento microbiano.

Fuente: Tjørve KMC, Tjørve E (2017) The use of Gompertz models in growth analyses, and new Gompertz-model approach: An addition to the Unified-Richards family.

En la Figura N° 9 también se puede observar el valor de inflexión fijo en 36.79% de la asíntota superior. Aquí la asíntota superior (A) se establece en 10, la tasa de crecimiento absoluta máxima K_U a 1.5, el tiempo en la inflexión T_i a 2 y el punto de inicio W_0 a 1.0.

Con una asíntota y una tasa de crecimiento establecidas, el tiempo de inflexión se sigue desde un punto de partida dado o viceversa. La tasa de crecimiento máxima está representada por la tangente en la inflexión (línea discontinua).

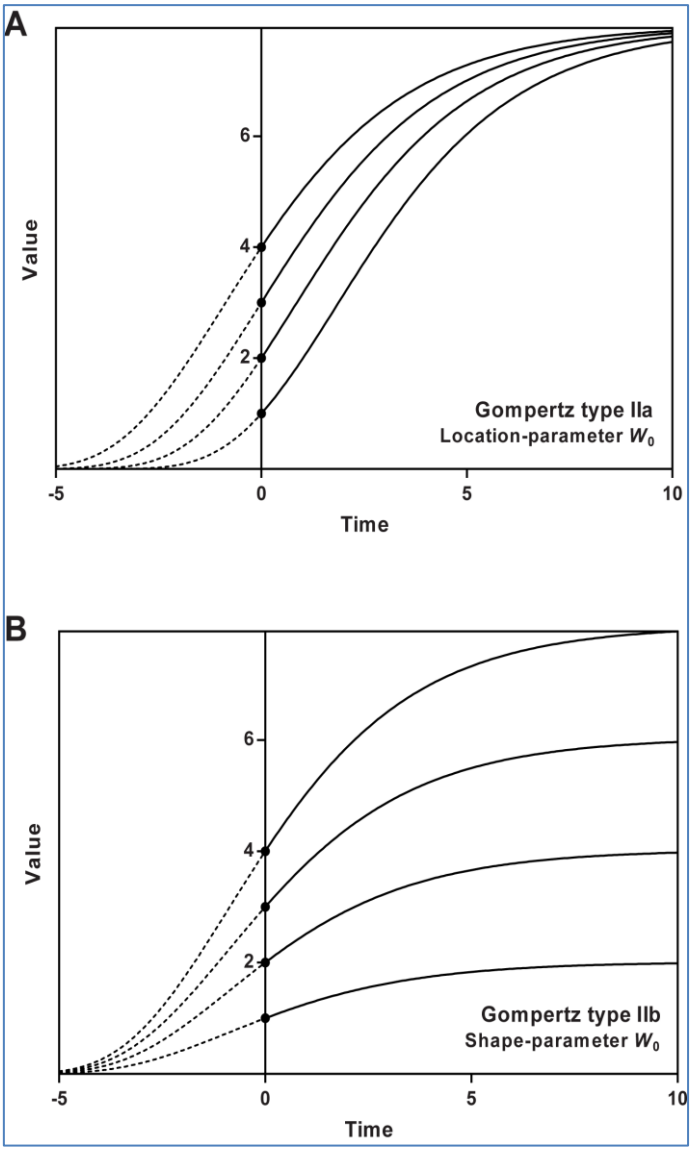
Dos tipos principales de modelos Gompertz

La mayoría de los modelos de Gompertz de tres parámetros tienen dos parámetros de "forma" que afectan la forma de la curva y un parámetro de "ubicación" que desplaza la curva horizontalmente sin cambiar su forma. Los parámetros de forma cambian la forma de la curva, pero dejan el valor del parámetro de ubicación sin alterar. El valor del parámetro se mantiene constante en relación con el eje x o con respecto al eje y, caracterizando el tipo I y el tipo II de los modelos Gompertz, respectivamente.

En los modelos de tipo I, un único parámetro controla el tiempo (es decir, el valor x) en el que se produce un punto específico en la curva. El punto representa una proporción fija (o porcentaje) de la asíntota superior, y el tiempo en que ocurre este punto no se ve afectado por los otros parámetros (aunque todos los demás puntos a lo largo de la curva sí lo son). En algunos modelos, este punto cae en la inflexión, que en el modelo de Gompertz ocurre en el 36.8% de la asíntota superior (Figura N° 9). En otros modelos, cae en algún otro porcentaje fijo de la asíntota.

En los modelos tipo II, un solo parámetro controla el valor inicial de la curva (es decir, la intersección con el eje y). En estas reconfiguraciones, los otros parámetros no afectan el punto de partida. Se mostrará cómo el parámetro de forma cambia la curva en un modelo de tipo I (Figura N° 10- A) y en un modelo de tipo II (Figura N° 10 - B).

FIGURA N° 10: Dos tipos de modelo tipo I & II



Fuente: Tjørve KMC, Tjørve E (2017) The use of Gompertz models in growth analyses, and new Gompertz-model approach: An addition to the Unified-Richards family.

En la Figura N° 10, ambos paneles muestran curvas de Gompertz con cuatro valores de punto de partida diferentes W_0. El panel "A" ilustra cómo el parámetro W_0 afecta la curva en los modelos de tipo II A (donde W_0 actúa como un parámetro de ubicación, manteniendo constante la asíntota superior), y el panel "B" ilustra cómo el parámetro W_0 afecta la curva en el tipo- Modelos II B (donde W_0 actúa como parámetro de forma, cambiando la asíntota superior).

Revisión de modelo

Algunas de las re-parametrizaciones del modelo de Gompertz comúnmente encontrada es:

$$W(t) = A * exp\ exp\left(-\ exp\ exp\left(-k_G(t - T_i)\right)\right) \qquad \text{.....(1)}$$

Donde $W(t)$ es el valor esperado (masa o longitud) en función del tiempo (por ejemplo, días desde el nacimiento o la eclosión) y t es el tiempo, A representa la asíntota superior

(valor adulto), k_G es un coeficiente de tasa de crecimiento (que afecta la pendiente), y T_i representa el tiempo en la inflexión. El parámetro T_i desplaza la curva de crecimiento horizontalmente sin cambiar su forma y, por lo tanto, es lo que a menudo se denomina parámetro de ubicación (mientras que A y k_G son parámetros de forma), lo que significa que este es un modelo de tipo I. Sin embargo, más específicamente nos referiremos al modelo (i) como una forma T_i , porque T_i es uno de los parámetros, en oposición a la forma W_0 (que no incluye T_i). Tenemos una forma W_0 de un modelo en el caso de que W_0 sea el valor (punto de inicio / intersección) en el eje y (intersección). Todos los modelos W_0 son de tipo II. En un trabajo anterior sistematizamos una serie de T_i y W_0 formas para otros modelos de crecimiento en la familia Richards: el exponencial negativo, el logístico y el Von Bertalanffy.

La mayoría de las otras parametrizaciones del modelo de Gompertz encontradas en la literatura son menos útiles, ya que sus parámetros son más difíciles de interpretar, por ejemplo:

$$W(t) = A * exp\ exp\ (-\ exp\ exp\ (-k_G t + b)\ ) \qquad \text{…….(2)}$$

y

$$W(t) = A * exp\ exp\ (-\ exp\ exp\ (-k_G t)\ ) \qquad \text{…….(3)}$$

que son ambos modelos de tipo II, pero donde el parámetro b y el parámetro c hacen que el punto de inicio se comporte como un valor relativo (un porcentaje de la asíntota superior), y ninguno de los dos representa el valor relativo para El punto de partida (que por lo tanto se ha derivado de alguna ecuación). Por lo tanto, no es correcto como, por ejemplo, Kurnianto y colegas afirman que el parámetro c (en el modelo (iii)) "no tiene una importancia biológica específica". Vemos que uno puede convertir los valores del parámetro de ubicación entre los modelos (i), (ii) y (iii) a partir de las siguientes ecuaciones: $b = ln(c)$ para que $c = exp(b)$, $b = k_G * T_i$ para que $T_i = b/k_G$ y $c = exp(k_G * T_i)$ para que $T_i = ln\ ln\ (c)\ /k_G$. Sin embargo, tenemos que concluir que el modelo (i) es más útil que los otros dos, a medida que la T_i Parámetro directamente, en lugar de tener que calcularlo.

2.2.6. Método Quasi Newton

Los métodos Quasi-Newton consisten en aproximar la matriz Hessiana de cada iteración mediante fórmulas de recurrencia que la relacionen con el valor que toma en iteraciones precedentes, ver Bonnans et al. (2002). La dirección de búsqueda en el método de Newton requiere del cálculo de la matriz Hessiana y que esta sea invertible, cuestión que no se puede garantizar en el curso de las iteraciones. Esto conlleva de un gran esfuerzo del punto de vista computacional en el cálculo de esta matriz. Con el fin de soslayar estas dificultades los métodos Cuasi-Newton aproximan la matriz $\nabla^2 f(\theta_k)$ por una matriz definida positiva B, que se modifica en cada iteración y que converge a la verdadera matriz Hessiana; ver Coleman (1984), Frandsen et al. (2004), Lange (2004). Los métodos Cuasi-Newton han demostrado ser bastante eficientes en optimización no lineal y juegan un papel importante en muchas implementaciones. Además, este tipo de métodos, a diferencia de los de Newton, tienen una tasa de convergencia super lineal, lo que frecuentemente desde el punto de vista computacional resulta ser más eficiente que el método analítico de Newton; ver De la Fuente O'Conor (1995), Luenberger (1984). En estos métodos las iteraciones pueden ser más costosas computacionalmente; sin embargo, la información almacenada en la aproximación del Hessiano podría reducir el número total de iteraciones

comparado con otros métodos tradicionales (Nocedal y Wright, 1999). Consideremos la solución del sistema

$$B_k d_k = -\nabla f(\theta_k) \quad \text{…….(1)}$$

Donde B_k es una matriz cuadrada definida positiva.

Otra forma de presentar los métodos Cuasi-Newton es a través de la aproximación de la inversa del Hessiano, es decir, $B = H^{-1}$. Como todo método iterativo, y según lo mencionado anteriormente, este necesita de una aproximación inicial. En este caso, además, se necesita una aproximación para el Hessiano, es decir B0 inicial, la que frecuentemente se puede tomar como la matriz identidad $B_0 = I$ si no existe más información. También B_0 se puede considerar como un múltiplo de la matriz identidad, es decir, $B_0 = \eta I$, para un $\eta > 0$. La matriz Hessiana se actualiza de acuerdo a la siguiente estructura (Nocedal y Wright, 1999):

$$B_{k+1} = B_k + U_k \quad k = 0,1,2, \ldots \quad \text{…….(2)}$$

Donde U_k es la expresión que aproxima a la verdadera matriz Hessiana. Veamos dos estrategias posibles. Una condición para definir Bk es (ver Frandsen et al., 2004; Luenberger, 1984; Nocedal y Wright, 1999):

$$B_{k+1}(\theta_{k+1} - \theta_k) = \nabla f(\theta_{k+1} - \nabla f(\theta_k)) \quad \text{…….(3)}$$

Esta condición se conoce como la condición secante que se basa en una generalización del método de la secante unidimensional, donde la matriz Hessiana $\nabla^2 f(\theta_k)$ se reemplaza por una aproximación B_k. Definiendo $S_k = \theta_{k+1} - \theta_k = \alpha_k d_k$ y $v_k = \nabla f(\theta_{k+1}) - \nabla f(\theta_k))$ se obtiene $v_k = B_k + S_k$. La condición secante (Kelley, 1995) se satisface si:

$$S_k{}^T y_{k>0} \quad \text{…….(4)}$$

que se conoce como Condición de Curvatura (Frandsen et al., 2004). La matriz de actualización Bk se puede calcular mediante diferentes métodos. A continuación, se presentan dos métodos para la actualización de dicha expresión. Este método fue desarrollado por Broyden, Fletcher, Goldfard y Shanno, conocido como BFGS, y toma la siguiente forma. Ver Fletcher (1980), Frandsen et al. (2004), Luenberger (1984):

$$U_k = -\frac{(B_k S_k)(B_k S_k)^T}{S_k{}^T B_k S_k} + \frac{y_k y_k{}^T}{y_k{}^T S_k} \quad k = 0,1,2, \ldots. \quad \text{…….(5)}$$

Uno de los esquemas más inteligentes para la construcción de la inversa del Hessiano, fue propuesto originalmente por Davidon y más tarde desarrollado por Fletcher y Powell, conocido actualmente como DFP. La actualización está dada por: ver Fletcher (1980), Frandsen et al. (2004), Luenberger (1984);

$$U_k = H_k + \frac{S_k S_k{}^T}{S_k{}^T y_k} + \frac{H_k y_k y_k{}^T H_k}{y_k{}^T H_k S_k} \quad k = 0,1,2, \ldots. \quad \text{…….(6)}$$

2.3. INDICADORES DE SELECCIÓN DE MODELOS DE PREDICCIÓN

2.3.1. La Raíz del Error Cuadrático Medio o RMSE

La raíz del error cuadrático medio o RMSE por sus siglas en inglés, mide el valor medio cuadrático del error. Es presentado por (Hyndma & Koehler, 2006; Bergmeir & Benítez,

2012) y usado en (Liu et al., 2010; Karamirad et al., 2013; Gaiser et al., 2010; Wu et al., 2008). Éste índice es el más usado para la validación de modelos de sistemas físicos en la literatura revisada. Su resultado tiene las unidades de la variable medida y pondera los pronósticos que están más alejados del valor medido.

$$RMSE = \sqrt{\frac{1}{n} * \sum_{i=!}^{n} (y_i - \hat{y}_i)^2} \qquad \text{…….(1)}$$

Donde y_i es el valor medido, $\hat{y}_i$ el valor estimado y n el número de muestras.

2.3.2. El Error Medio Absoluto o MAE

El error medio absoluto o MAE por sus siglas en inglés, mide la magnitud promedio del error entre los datos medidos y los datos estimados por el modelo. El MAE es presentado por Hyndman y Koehler (2006) y es usado por Karamirad et al. (2013) y Gaiser et al. (2010). Su valor mínimo es cero y ocurre cuando los datos medidos y las estimaciones son iguales en todo el rango de mues-tras. Conserva las unidades de los datos medidos.

$$MAE = \frac{1}{n} * \sum_{i=!}^{n} |y_i - \hat{y}_i| \qquad \text{…….(1)}$$

Donde y_i es el valor medido, $\hat{y}_i$ el valor estimado y n el número de muestras.

2.3.3. El Error Medio Absoluto Porcentual o MAPE

El error medio absoluto porcentual o MAPE por sus siglas en inglés, mide el porcentaje de error promedio de las estimaciones. Es presentado por Bergmeir y Benítez (2012) y usado por Liu et al. (2010). Su valor mínimo es cero, y significa que las estimaciones y los datos medidos son iguales en todo el conjunto de datos.

$$MAPE(\%) = \frac{100}{n} * \sum_{i=!}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \qquad \text{…….(1)}$$

Donde y_i es el valor medido, $\hat{y}_i$ el valor estimado y n el número de muestras.

2.3.4. Máxima verosimilitud

Este método de ajuste selecciona los calores de los parámetros que maximizan la probabilidad de que las actuales observaciones hubieran ocurrido si los parámetros fueran verdaderos (Hilborn y Walters, 1992). Para la estimación de los parámetros se maximizó la función del logaritmo negativo de la verosimilitud a través del algoritmo de Newton (Hilborn y Mangel, 1997) asumiendo un error multiplicativo de los residuales (Neter et al., 1996).

$$-lnL(datos) = \sum_{i=1}^{n} \left(-\frac{1}{2} * Ln(2\pi) \right) - \left(\frac{1}{2} * Ln(\sigma^2) - \frac{(lny_i - ln\hat{y}_i)^2}{2\sigma^2} \right) \qquad \text{…….(1)}$$

Donde θ hace referencia a los parámetros del modelo, y_i son los pesos observados y $\hat{y}_i$ son los valores calculados mediante el modelo. La desviación estándar de los errores (σ) se determinó a través de la siguiente ecuación:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (lny_i - \hat{y}_i)^2}{n}} \qquad \text{…….(2)}$$

2.4. CRITERIOS DE SELECCIÓN DEL MODELO DE CRECIMIENTO

2.4.1. Suma de cuadrados residuales

SCR es un método basado en la búsqueda de la combinación de parámetros que permitan una menor distancia cuadrática de los residuales, los cuales resultan de la diferencia entre los valores observados y_i y los calculados ($\hat{y}_i$). Lo anterior se resume en la siguiente ecuación (Hilborn y Mangel, 1997).

$$SCR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad \text{…….(1)}$$

La búsqueda de esta combinación de parámetros se realizó mediante el algoritmo de Newton asumiendo un error multiplicativo de los residuales (Neter et al., 1996; Rodríguez – Domínguez et al., 2012).

2.4.2. Criterio de Akaike

La selección del modelo que mejor describe el crecimiento de pesos de los peces, se realizó con base en el criterio de Akaike (AIC: Akaike Information Criterion). El AIC se basa en las distancias relativas esperadas entre el modelo ajustado y los datos observados; por lo tanto, el mejor modelo candidato será aquel que posea el menor valor AIC ($AIC_{mín}$ Burnham y Anderson, 2002). Dependiendo de la forma de abordar el ajuste de los modelos a los datos varía la ecuación del AIC. Para SCR se utilizó la siguiente ecuación (Johnson y Omland, 2004; Katsanevakis, 2006; Mercier et al., 2011).

$$AIC = n * ln\ ln\left(\hat{\sigma}^2\right) + 2k \qquad \text{…….(1)}$$

Mientras que, para la máxima verosimilitud, se tiene lo siguiente:

$$AIC = (2 * L) + 2\theta_i \qquad \text{…….(2)}$$

Donde $\hat{\sigma}^2$ es la razón de la suma de cuadrados residuales y el número de datos ($\hat{\sigma}^2 = \mathrm{SCR/n}$). Los argumentos $2\mathrm{k}$ y $2\theta_\mathrm{i}$ son la penalización que se otorga al modelo por cada parámetro utilizado, donde k es el número de parámetros que son empleados en cada modelo (θ_i) incluyendo $\hat{\sigma}^2(\mathrm{k} = \theta_\mathrm{i} + 1)$. Finalmente $-\mathrm{lnL}$ es el logaritmo negativo de la verosimilitud. De esta manera, el modelo que resulta seleccionado es el que obtiene menor AIC entre todo el conjunto de candidatos, resultando favorecido sobre las bases de ajuste y parsimonia (Link y Barker, 2006).

2.4.3. Coeficiente de determinación ($\mathbf{R^2}$)

El coeficiente de determinación, se define como la proporción de la varianza total de la variable explicada por la regresión, este refleja la bondad del ajuste de un modelo a la variable que pretender explicar. Es importante saber que el resultado del coeficiente de determinación oscila entre 0 y 1. Cuanto más cerca de 1 se sitúe su valor, mayor será el ajuste del modelo a la variable que estamos intentando explicar. De forma inversa, cuanto más cerca de cero, menos ajustado estará el modelo y, por tanto, no es factible usar el modelo con un coeficiente de determinación muy pequeño.

$$\mathrm{R}^2 = \frac{\sum_{t=1}^{T}(\widehat{Y_t} - \overline{Y})^2}{\sum_{t=1}^{T}(Y_t - \overline{Y})^2} \qquad \text{…….(1)}$$

3. MATERIALES, MÉTODOS Y PROCEDIMIENTOS

3.1. TIPO DE INVESTIGACIÓN

La modalidad de la investigación es exploratoria, se construyen modelos para predecir la curva de crecimiento de los peces, y se observa tanto las variables categóricas y numéricas con el fin de encontrar el mejor modelo y determinar que variables son las que influyen más en esta curva de crecimiento.

3.2. PROCEDENCIA DE LA INFORMACIÓN

La información que se requerirá para el desarrollo de la investigación fue proporcionada por la compañía Alicorp, teniendo así datos de enfoque cuantitativo y cualitativo.

3.3. VARIABLES DE ANÁLISIS

A continuación, se describirá las variables a utilizar en el trabajo de investigación:

3.3.1. Variable dependiente: Peso Promedio Final

La piscicultura es una actividad pecuaria que ha ido aumentando en todos los países latinoamericanos (nos enfocaremos en la especie llamada Tilapia), uno de los principales indicadores que mide este aumento de esta actividad es mediante la medición de los pesos de los peces cultivados; es decir, se realiza un estudio desde el nacimiento del pez hasta su cultivo como pez adulto.

En este estudio el peso promedio final se medirá en "kg", resaltando que es un peso promedio ya que el valor medido es por el tamaño de la jaula y el tamaño de los peces no es el mismo para todos.

3.3.2. Variables independientes: Factores Ambientales

En sus hábitats originarios, los peces están sometidos a cambios estacionales, que influyen en la duración del día, en las temperaturas máximas y mínimas del medio, en la variación de pH y dureza del agua, etc. Algunos peces incluso se ven impelidos a reproducirse al notar un descenso del nivel del agua, acompañado por un aumento de temperatura, y sus huevos requieren reposo en un sustrato húmedo, fuera del agua y a oscuras.

Por estas causas, cuando deseemos criar con éxito alguna especie, la primera medida ha de consistir en adquirir toda la información sobre estos peces. Principalmente tener información de los lugares de origen y su climatología nos ayudará a comprender los mecanismos de integración biología-medio ambiente, válidos para la especie. Asimismo, existen algunos estudios sobre algunos factores ambientales que influyen en el crecimiento de estas especies; tenemos, por ejemplo:

- **Temperatura:**

Este parámetro tiene un rol importante en el funcionamiento de los ecosistemas acuáticos. Es un ente regulador de factores abióticos como: pH, densidad, viscosidad, solubilidad de nutrientes y gases, entre otros. Los peces por su parte no tienen capacidad propia de regulación de su temperatura corporal y ésta depende del medio acuático en que viven. Mayores temperaturas menores son las cantidades de oxígeno disuelto en el agua. La interacción de algunos de los factores físico-químicos dependientes de la temperatura

puede adicionalmente afectar la nutrición, el crecimiento, y el metabolismo en general de especies acuáticas (Becker, C. D., & Genoway, R. G. 1979).

- **Salinidad**

En aguas continentales la salinidad corresponde a la concentración de todos los iones disueltos en el agua, especialmente el contenido de cloruros. Los arrastres de la erosión de suelos, la mineralización de rocas son fuentes que pueden incrementar la presencia de iones en cuerpos acuáticos. Para obtener mejores índices de sobrevivencia y crecimiento, un rango óptimo de salinidad debe mantenerse en el agua del estanque. Si la salinidad es muy alta, los peces pueden alterar sus procesos de crecimiento y reproducción (Matthews, W. J. 2012).

- **Oxígeno**

El oxígeno disuelto corresponde al parámetro más importante en la calidad del agua e imprescindible para el desarrollo de la acuicultura. Influye en la generación de energía, y la movilización del carbono en la célula mediante la respiración aeróbica de los peces y organismos presentes en el agua. Por lo que si hay déficit se afecta el crecimiento y la conversión alimenticia de los peces y demás organismos acuáticos.

Por lo que a medida que aumenta la profundidad, disminuye encontrándose en mayores concentraciones en la zona superficial. La fotosíntesis a su vez depende de la presencia de luz. Aguas turbias, o con alta presencia de sólidos o color afectan considerablemente el proceso. La presencia de grandes cantidades de materia orgánica como los restos de alimentos y excretas de la cría de peces consumen elevadas cantidades en los procesos oxidativos. Existen otros factores como la temperatura que también inciden negativamente, pues a mayor temperatura mayor consumo de oxígeno.

- **Turbidez**

Está dada por el material en suspensión en el agua, bien sea mineral u orgánico, sedimentos procedentes de la erosión, etc. En cuanto al grado de turbidez varía de acuerdo a la naturaleza, tamaño y cantidad de partículas suspendidas. La turbidez originada por el plancton es una condición necesaria en acuicultura, limita la habilidad de los peces para capturar el alimento y por consiguiente éste irá al fondo del estanque incrementando la cantidad de materia orgánica en descomposición lo que va en detrimento del oxígeno disuelto (Matthews, W. J. 2012).

- **Potencial de Hidrógeno (pH)**

La estabilidad del pH viene dada por la llamada reserva alcalina o sistema de equilibrio que corresponde a la concentración de carbonato o bicarbonato. Los extremos letales de pH

para la población de peces en condiciones de cultivo, están por debajo de 4 y por encima de 11. Además, cambios bruscos de pH pueden causar la muerte. Las aguas ácidas irritan las branquias de los peces, las cuales tienden a cubrirse de moco llegando en algunos casos a la destrucción histológica del epitelio. La sobresaturación de CO2 acidifica aún más el agua causando alteraciones de la osmorregulación y acidificando la sangre (Matthews, W. J. 2012).

- **Otras sustancias. Nitritos ($NO2^{-}$)**

Los aumentos de las concentraciones de $NO2^{-}$ en el agua inducen la acumulación de $NO2^{-}$ en la sangre y en los tejidos y, a través de reacciones complejas, producen derivados tóxicos que afectan los procesos fisiológicos hematológicos (Jensen, 1995). Esto puede comprometer la supervivencia y el crecimiento de peces en ambientes caracterizados por altas temperaturas y bajo contenido de oxígeno disuelto, como en las regiones tropicales donde la temperatura del agua permanece entre 24 y 28 °C (Williams et al., 1997). Si bien es cierto se mencionan distintos factores ambientales que pueden ser muy influyentes en el crecimiento de los peces, para este trabajo de investigación se cuenta con la información diaria del oxígeno y la temperatura del ambiente donde se está realizando el cultivo de los peces.

3.3.3. Variables independientes: Factores de Producción

Existen muchos factores de producción, estos a diferencia de los factores ambientales pueden ser manipulados por el ser humano. Dentro de los principales factores que se cuenta con información, tenemos a:

- **Alimento para peces**

Esto mide qué cantidad de alimento en “kg” necesita el pez según su crecimiento, ya que entendemos por lógica del ser vivo que mientras más crezca más alimento necesitará para sobrevivir.

- **Número de días que se alimentan a los peces**

Esto indica el número de días por el cual los peces son alimentados, ya que al inicio de su crianza se alimentan pocas veces, es decir mientras pase el tiempo para el pez mayor será las veces que se alimente.

- **Especie**

Tenemos 3 tipos de especies, esto quiere decir que son conjunto de individuos que tienen características comunes, esto es importante ya que antes de empezar a criar a un pez debemos saber su procedencia, su hábitat original para tener un desarrollo óptimo de la especie.

- **Profundidad**

Como los peces son criados en jaulas, estas son sumergidas en aguas y este factor de la profundidad medida en metros sobre el nivel del mar (m.s.n.m.) es importante medir para el crecimiento y cultivo de peces.

- **Genética**

La genética estudia los caracteres hereditarios de los peces, se puede tener una especie de pez con diferentes genéticas, por ello es importante tener en cuenta esta variable.

- **Sistema de oxígeno**

Esta variable es importante medir ya que según el hábitat de los peces se requiere un nivel de oxígeno, por ellos en las granjas a pesar de tener indicadores de oxígeno ambientales, suelen emplear métodos de oxigenación artificial.

- **Tipo de alimentación**

Esta variable contribuye a la dieta de los peces, en este caso se diferencian dos tipos de alimentación, mencionados como “mono” y “micro”.

- **Estación**

Esta variable se puede decir que es ambiental, pero las temporadas como el invierno, primavera, otoño y verano pueden ser elegidas por el ser humano para iniciar la crianza, por lo que pasaría verse como un factor de producción que depende del ser humano que evaluará la crianza.

- **Tipo de jaula**

Esta variable nos brinda información de las dimensiones de las jaulas donde se efectúa la crianza de los peces, en este estudio se ven las dimensiones de “30cm x 30cm” y “40cm x 40cm”

- **Luz Artificial**

Esta variable nos indica sobre si en la granja se está efectuando el uso de energía, ya que a veces mantienen a los peces despiertos durante más tiempo y así tener un tipo diferente de crianza.

- **Horas Luz**

Esta variable nos indica el tiempo en horas de luz artificial que se le brinda a las especies en estudio.

3.4. OPERACIONALIZACIÓN DE LAS VARIABLES

TABLA N° 1: Matriz de operacionalización de variables

CLASIFICACIÓN DE LA VARIABLE	VARIABLES	DIMENSIÓN	DEFINICIÓN CONCEPTUAL	INDICADORES	ESCALA DE DIMENSIÓN
Variable Dependiente (Cuantitativa)	Factor de producción	Peso de las especies	Peso promedio de las especies al final del mes de la fecha correspondiente según jaula después de los 14 meses de su ciclo productivo	PESO_PROMEDIO_FINAL ¿Cuál es el peso promedio de las especies?	Numérica
Variables Independientes	Factor de producción	Información descriptiva	Nombre de la empresa de crianza	NOM_EMPRESA ¿Cuál es el nombre de la empresa de crianza?	Categórica
			Nombre del centro de crianza	NOM_CENTRO ¿Cuál es el nombre del centro de crianza?	Categórica
			Nombre de la jaula de crianza	NOM_JAULA ¿Cuál es el nombre de la jaula de crianza?	Categórica
			Año correspondiente al inicio de cosecha	AÑO_INICIO_CICLO ¿En qué año inicio la cosecha?	Numérica
			Fecha correspondiente al mes / año en el que se encuentra la cosecha	MES_INFORMACION ¿Cuál es el mes en el que se encuentra la cosecha?	Categórica
		Cantidad y peso de las especies	Número de especies que murieron en la fecha correspondiente	N_MORTALIDAD ¿Cuántas especies murieron hasta la fecha?	Numérica
			Número de especies que se encontraron al inicio del mes de la fecha correspondiente según jaula	N_INICIAL ¿Cuántas especies había al inicio de la cosecha?	Numérica
			Número de especies que se encontraron al final del mes de la fecha correspondiente según jaula	N_FINAL ¿Cuántas especies hay al final del mes?	Numérica
			Peso promedio de las especies al inicio del mes de la fecha correspondiente según jaula	PESO_PROMEDIO_INICIAL ¿Cuál es el peso promedio de las especies al inicio del mes?	Numérica
		Alimentación y características	Cantidad de alimento utilizado en Kg.	ALIMENTO_USADO_KG ¿Cuál es la cantidad de alimento usado?	Numérica
			Número de días en el que se alimentó	DIAS_ALIMENTADOS ¿Cuántos días se alimentó a las especies?	Numérica
			Tipo de especie del pez	ESPECIE ¿Cuál es el tipo de especie del pez?	Categórica
			Tipo de alimentación del pez	TIPO_ALIMENTACION ¿Qué tipo de alimentación tiene?	Categórica
			Tipo de genética del pez	GENETICA ¿Cuál es el tipo de genética del pez?	Categórica

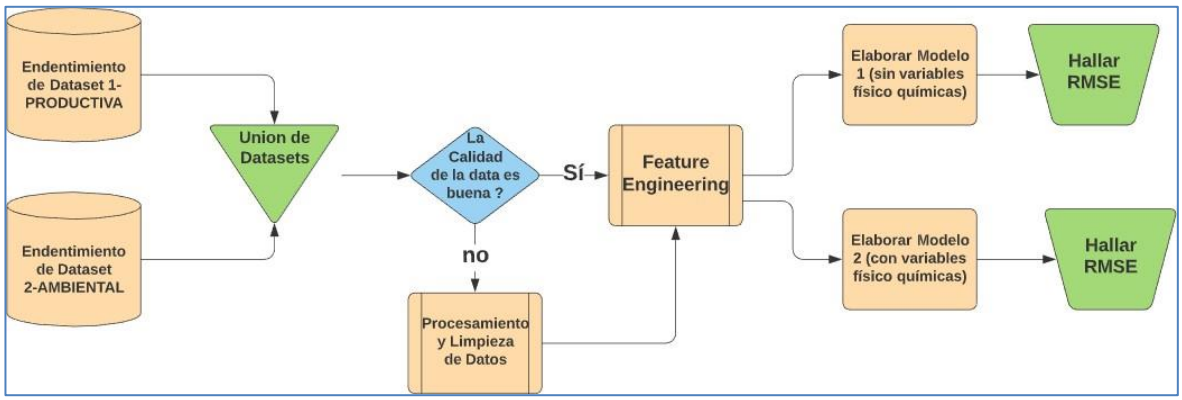
		Características de las condiciones de la jaula	Profundidad de la jaula en el mar en mts.	PROFUNDIDAD ¿Cuál es la profundidad de la jaula en la que se encuentran?	Numérica
			Estación del año	ESTACION ¿En qué estación del año se encuentra?	Categórica
			Tipo de jaula	TIPO_JAULA ¿Cuál es el tipo de jaula?	Categórica
			Uso de sistema de oxígeno	SISTEMA_OXIGENO ¿Usa algún sistema de oxígeno?	Categórica
			Uso de fotoperiodo	FOTOPERIODO ¿Usa fotoperiodo?	Categórica
			Cantidad de luz artificial en horas	HORAS_LUZ ¿Cuál es la cantidad de luz artificial?	Numérica
	Factor ambiental	Información descriptiva	Fecha correspondiente al día / mes / año en el que se toma información de las condiciones atmosféricas del ambiente	FECHA ¿Cuál es la fecha en la que se tomó información acerca de las condiciones atmosféricas?	Categórica
			Nombre de la empresa de crianza	CENTRO ¿Cuál es el nombre del centro de crianza?	Categórica
		Condiciones atmosféricas	Temperatura en grados centígrados a nivel de centro de las empresas de crianza de peces	TEMPERATURA ¿Cuál es la temperatura?	Numérica
			Oxígeno en milígramos / litros a nivel de centro de las empresas de crianza de peces	OXIGENO ¿Cuál es el oxígeno?	Numérica

Fuente: Elaboración propia.

3.5. DISEÑO DE LA INVESTIGACIÓN

Para el desarrollo del presente trabajo de investigación se utilizó la metodología internacional para el desarrollo de proyectos de minería de datos CRISP-DM descrita en el apéndice de la presente investigación. Los procesos y tareas desarrolladas en el trabajo se organizan de la siguiente manera:

FIGURA N° 11: Procedimiento usado para el desarrollo del estudio



Fuente: Elaboración Propia.

- **Comprensión del Negocio**

El objetivo de negocio de ALICORP S.A es poder brindarle a la empresa Vitapro una herramienta para que estos puedan contar con información oportunidad que les permita mejorar su proceso productivo y así incrementar sus ventas. Específicamente en este trabajo se busca obtener información relevante acerca de la etapa de engorde de los peces haciendo uso de variables poco comunes en los modelos clásicos como las variables físico químicas de la calidad del agua y otras propias del proceso productivo.

La búsqueda de solución a los problemas planteados se realizará a través de la construcción de modelos señalados anteriormente, los cuales, permitirán no solo predecir solo el peso promedio final de los peces sino también permite encontrar cuales son las variables más relevantes que ayudan a predecir mejor el peso promedio final de los peces.

- **Comprensión de los Datos**

Como se mencionó anteriormente el conjunto de datos de las variables productivas y las variables ambientales fueron brindados por la propia empresa Alicorp, dicha empresa recolectó la información en los propios centros de crianza en intervalos de tiempo mensual.

En esta etapa se hizo un análisis de las dimensiones de cada conjunto de datos, se estudiaron los conceptos de cada variable, se hizo un análisis de datos perdidos para posteriormente darles un tratamiento, y por último se hizo un análisis grafico univariado para entender el comportamiento de cada variable.

Con los datos proporcionados por la empresa, se puede observar que estos cumplen con los criterios mencionados por el estándar ISO-19115, que básicamente busca verificar la información geográfica de los datos proporcionados.

- **Preparación de los Datos**

En primer lugar, una vez que se estudió el comportamiento univariado de los datos en la etapa anterior se procedió a realizar el análisis bivariado muy importante para encontrar el comportamiento de las variables versus la variable objetivo del caso de estudio (peso promedio final de los peces) encontrando relaciones que nos dan un indicio de las variables con mayor relevancia para el estudio, este análisis se hizo tanto para las variables categóricas como a las continuas.

En segundo lugar, se hizo la unión de los datasets disponibles creando identificadores 'LLAVE JOIN' en cada conjunto de dato. Una vez que se tuvo todas las variables en un solo dataset se procedió a darle tratamiento a los formatos de las variables cambiándoles los tipos de datos para que vayan de acurdo a las necesidades de modelamiento.

En tercer lugar, se analizó las variables con valores faltantes decidiéndose eliminar todos los registros con valores perdidos para no crear relaciones entre variables al imputar los datos. Se realizó un análisis de valores atípicos donde se decidió eliminar los registros con valores atípicos para las variables más notorias como el peso promedio final, número de días en que se alimentó, etc. Después del tratamiento se decidió transformar la variable

objetivo peso promedio final de los peces para darle un comportamiento normal, para esto se aplicó la transformación logarítmica.

Posteriormente, se hizo el estudio de correlaciones con la variable objetivo y multicolinealidad entre variables, este estudio fue muy importante para seleccionar las variables que iban entrar al modelo.

Por último, se hizo un escalamiento de variables aplicando el método de Standard Scaler donde cada variable se resta por su media y se divide entre su desviación estándar, llevando así todas las variables a una solo escala para que el modelamiento pueda ser correcto.

- **Modelamiento**

Después de realizar todo el procesamiento correspondiente a los datos se decidió trabajar con los modelos: Regresión Lineal Múltiple, Árbol de decisión, Random Forest y Xgboost, ya que lo que se busca es predecir el peso promedio final de los peces, es decir se trata de un problema de regresión. Los modelos Random Forest y Xgboost hoy en día son bastante utilizados en el campo de la analítica y va en sincronía con la idea de Alicorp de buscar desarrollar el caso planteado con nuevas y mejores técnicas a los modelos clásicos.

- **Evaluación**

Una vez obtenidas las métricas para cada modelo, se procedió a realizar un cuadro comparativo donde se pudo observar que al trabajar con las variables físico-químicas (temperatura y oxigeno) se reduce el RMSE ayudando de esta forma a mejorar la predicción del peso promedio final de los peces. Además, también se graficaron la importancia de variables según la ganancia de información para cada modelo, donde se pudo observar que es recomendable trabajar con las variables físico químicas de la calidad del agua ya que estas toman buena relevancia solo por debajo del peso promedio inicial, los días y la cantidad de alimentación y la cantidad inicial y final de peces en la jaula.

4. PRESENTACION Y ANALISIS DE RESULTADOS

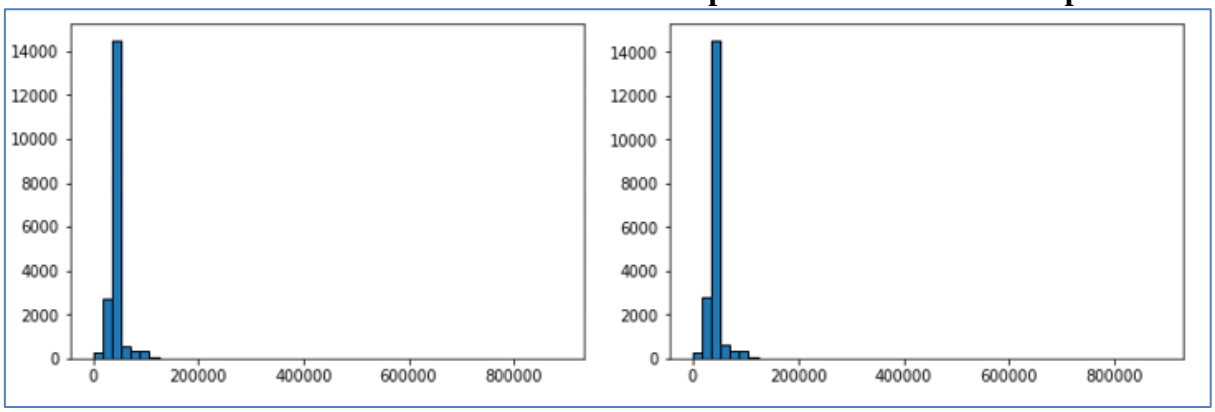
- **Análisis de los datos**

Inicialmente se contó con 2 bases de datos, la primera cuanta con información mensual de producción (ciclos productivos) y logística de los peces. Cuenta con un total de 19931 registros y 21 variables, que abarca información de un total de 68 jaulas, correspondiente a 8 empresas de crianza de peces desde setiembre 2016 a setiembre 2019. La segunda cuenta con información diaria de condiciones atmosféricas como la temperatura y oxígeno a nivel de centro de las empresas de crianza de peces, teniendo un total de 103740 registros con información atmosférica desde 2013 al 2018. La descripción de cada variable se encuentra dentro de la matriz de operacionalización.

La variable objetivo se encuentra dentro de la base de producción, y para obtener mejores predicciones en cuanto al peso promedio final de los peces, se optó por unir los factores

ambientales a la base de producción ya que las condiciones atmosféricas también afectan a la variable objetivo. En consecuencia, dentro de la base ambiental se construyó una nueva variable en función del centro y de la fecha diaria, la cual sirvió como código clave para la unión de ambas bases (Producción y Ambiental). De acuerdo a esto se analizó los datos que tienen mayor relevancia en base al negocio que se quiere enfocar. A continuación, se muestra las distribuciones de las variables numéricas:

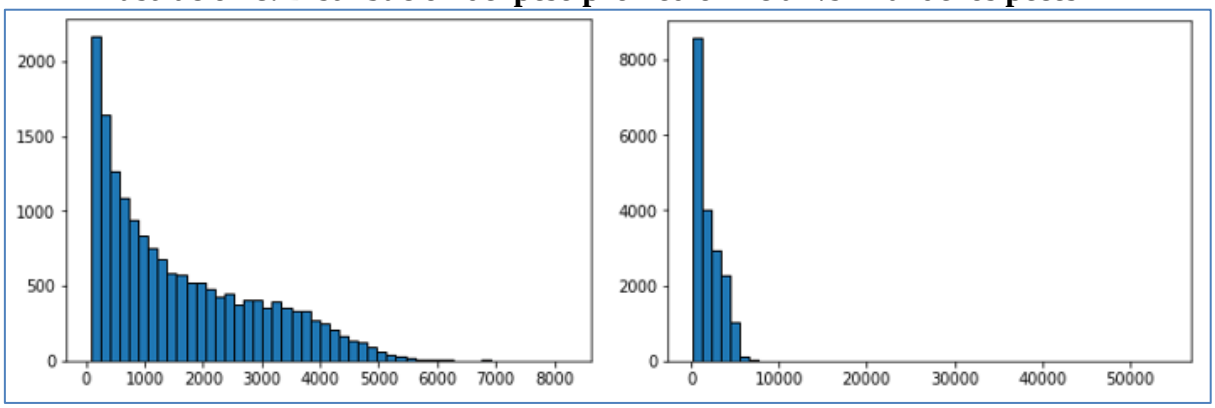
Ilustración 7. Distribución de Número Inicial de peces vs Número Final de peces



Fuente: Elaboración propia.

Se observa que no existe cambios significativos entre el número inicial y número final de peces.

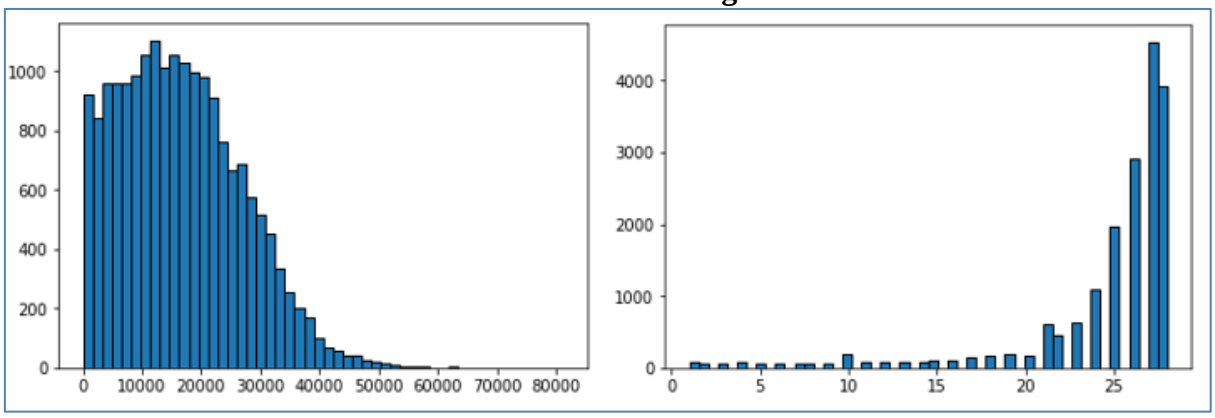
Ilustración 8. Distribución del peso promedio inicial vs final de los peces



Fuente: Elaboración propia.

Se observa el cambio en la distribución de los pesos promedios debido a los factores ambientales.

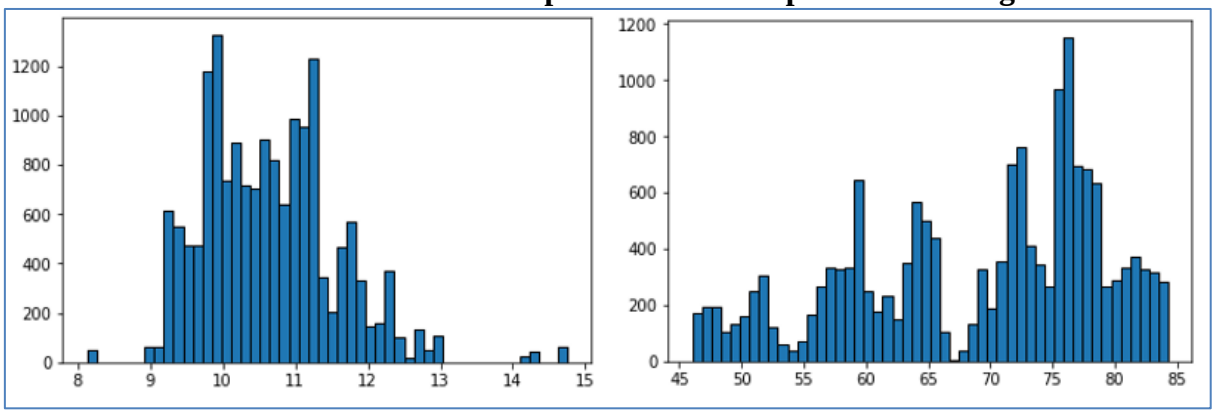
Ilustración 9. Distribución de Alimento kg & Días alimentados



Fuente: Elaboración propia.

En la primera gráfica, se observa la distribución de alimentos en kg vs días alimentados; en la segunda gráfica se puede observar la distribución de días alimentados vs la cantidad de alimentos en kg.

Ilustración 10. Distribución del promedio de Temperatura & Oxígeno

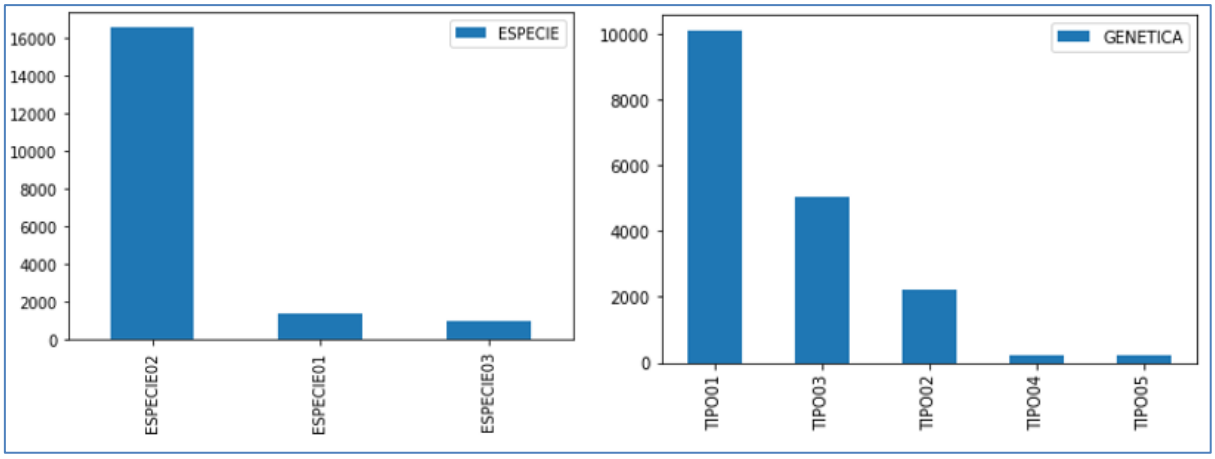


Fuente: Elaboración propia.

En la primera gráfica, se observa la distribución del promedio de Temperatura vs la cantidad de peces; en la segunda gráfica se puede observar la distribución de oxígeno vs la cantidad de peces.

A continuación, se muestras las gráficas de distribución de las variables categóricas:

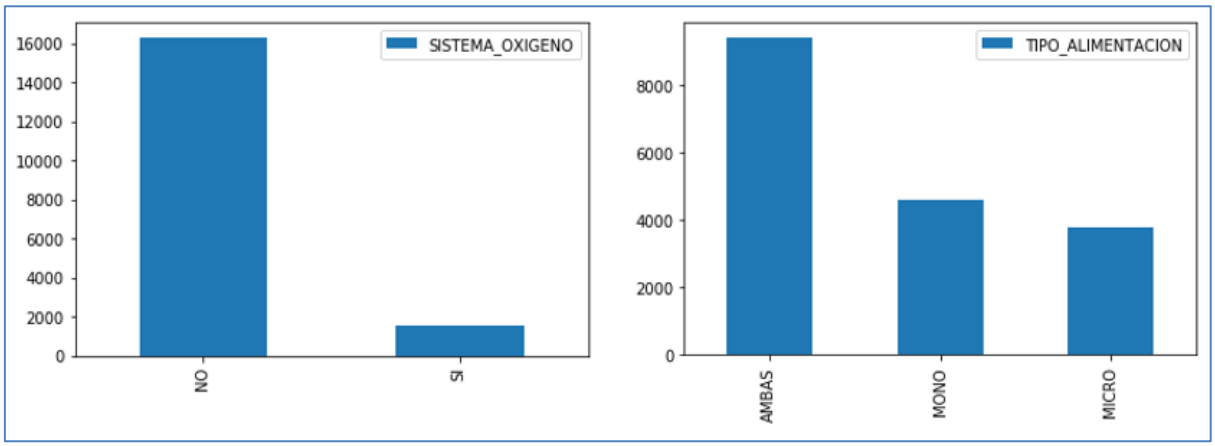
Ilustración 11. Distribución de Especie & Genética de los peces.



Fuente: Elaboración propia.

En el primer gráfico se observa que la mayor cantidad de peces son de la especie 02; y en la segunda gráfica se observa que la mayoría de los peces son de tipo 01.

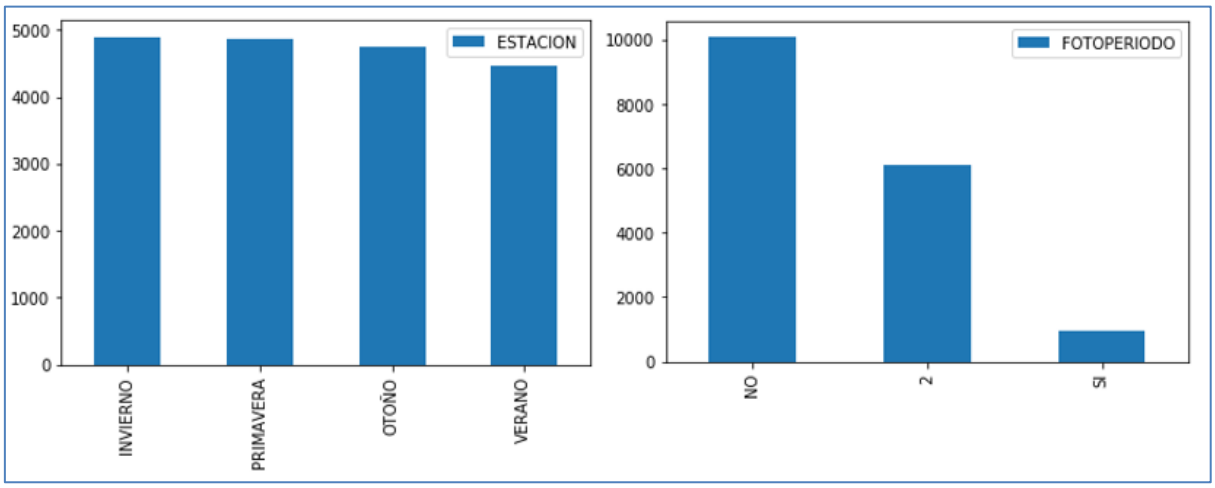
Ilustración 12. Distribución del Sistema de Oxígeno & Tipo de Alimentación



Fuente: Elaboración propia.

En la primera gráfica, se observa la mayoría de peces no recibe el Sistema de Oxígeno; y en la segunda gráfica se observa que la mayoría de peces recibió de ambas comidas.

Ilustración 13. Distribución de la Estación y el Fotoperiodo.



Fuente: Elaboración propia.

En la primera gráfica, se observa que no se tiene cambios significativos según la estación del año; en la segunda gráfica se puede observar que en la mayoría de peces prefieren no estar expuestos a la luz.

De acuerdo a los análisis realizados a las distribuciones de las variables numéricas y categóricas, podemos concluir que la variable objetivo (peso promedio final) presenta una distribución exponencial por lo que se le aplicará transformación logarítmica con el fin modelar de forma adecuada. Asimismo, existe una especie más predominante en relación a las demás al igual que su genética, también la mayoría de las especies marinas no usan sistema de oxígeno y el tipo de alimentación que mantienen puede ser de dos formas por separado (Mono y Micro) y en conjunto (ambas). Para obtener un análisis más completo en cuanto a las variables numéricas y la relación que tienen entre ellas, se procedió a realizar una matriz de correlaciones.

Con el fin de obtener información relevante y con datos que guarden relación se procede a determinar las variables que poseen datos perdidos y atípicos en relación al resto de datos, para poder tomar acciones que mejoren los datos. Por ejemplo, en caso de tener datos perdidos, es decir que vengan vacíos y sin información alguna se tomara la media y moda por variable numérica y categórica respectivamente como reemplazo de estas variables. En caso de tener datos atípicos se procederá a reemplazarlo por el cuantil mínimo y máximo.

Tabla 2. Porcentaje de datos perdidos por variables

VARIABLE	DATOS PERDIDOS	
N_MORTALIDAD	143	0.75%
N_INICIAL	184	0.97%
N_FINAL	4	0.02%
PESO_PROMEDIO_INICIAL	1698	8.93%
ALIMENTO_USADO_KG	181	0.95%
DIAS_ALIMENTADOS	816	4.29%
PROFUNDIDAD	124	0.65%
PROM_TEMPERATURA	2501	13.16%
PROM_OXIGENO	2501	13.16%

Fuente: Elaboración propia mediante el uso del Software Python.

Tabla 3. Porcentaje de datos atípicos por variables

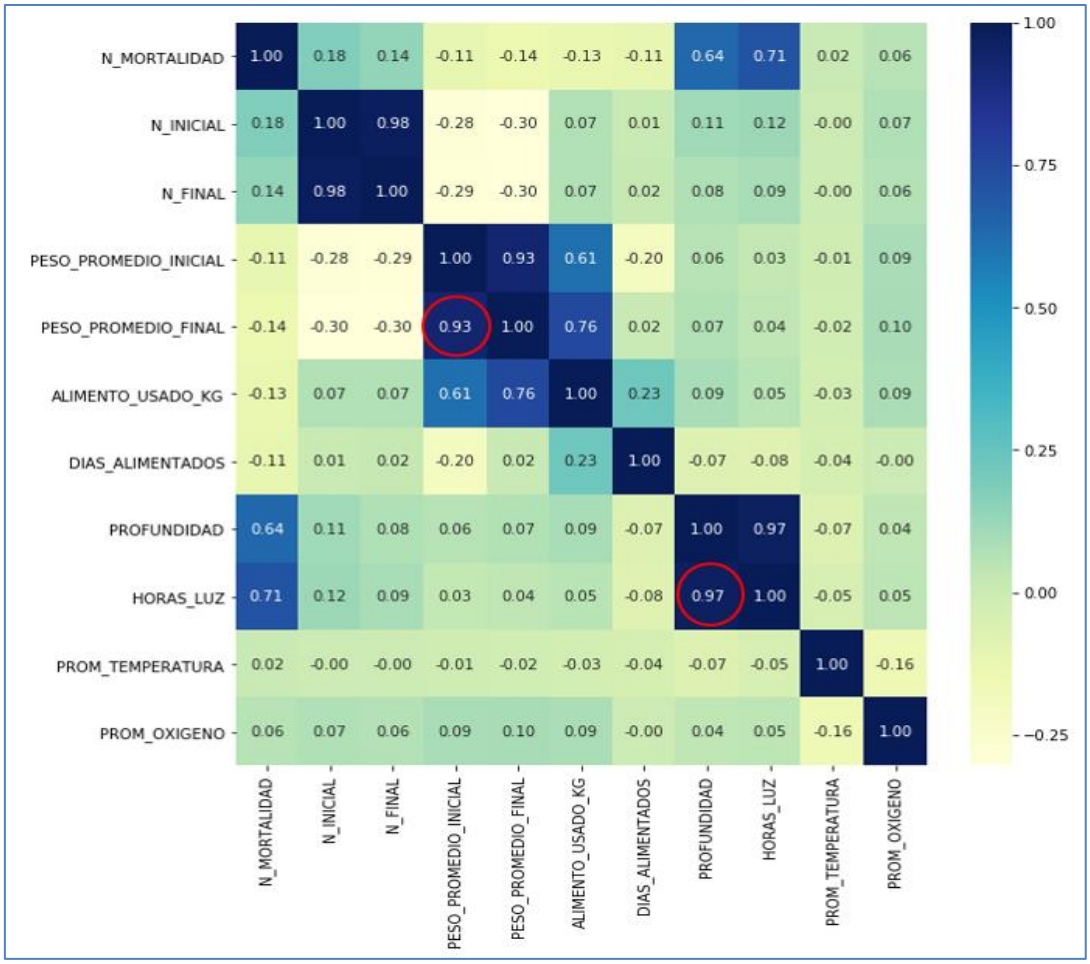
VARIABLE	DATOS ATIPICOS	
N_MORTALIDAD	1922	10.11%
HORAS_LUZ	1910	10.05%
DIAS_ALIMENTADOS	1866	9.82%
PROFUNDIDAD	1856	9.77%
N_INICIAL	1713	9.01%
N_FINAL	1697	8.93%
PROM_TEMPERATURA	335	1.76%
ALIMENTO_USADO_KG	131	0.69%
PESO_PROMEDIO_INICIAL	129	0.68%
PESO_PROMEDIO_FINAL	40	0.21%

Fuente: Elaboración propia mediante el uso del Software Python.

Continuando con el análisis de los datos, se determinó que existen altos niveles de correlación entre las variables referidas a la cantidad de luz artificial en horas y la profundidad en metros a la que la jaula está expuesta, realizando un análisis más exhaustivo se vió que ambas variables presentan el mismo valor en un 70% de la data, y además existe un patrón repetido por profundidad y periodo de cosecha, por lo que próximamente se procederá a crear una variable que contemple el promedio de horas luz por nivel de profundidad en metros para cada cliente, centro, jaula y periodo de cosecha.

Adicionalmente, existen dos variables que presenta fuerte correlación, esto es evidente ya que las variables referidas al número inicial y final de peces guardan relación aparente, la nueva variable vendría a ser la variación absoluta entre la cantidad inicial y final de peces por mes. Próximamente, también se procederá a crear una variable en relación a estas dos mencionadas.

Ilustración 14. Matriz de correlaciones - Variables cuantitativas



Fuente: Elaboración propia.

Se puede observar que existe una alta correlación entre los pesos iniciales y finales, esto se debe a que son los mismos individuos en estudio, la información relevante se podría observar la correlación de las horas luz con respecto a la profundidad.

Luego de la creación de nuevas variables, se tendrá una nueva matriz de correlaciones, en donde se nota la fuerte correlación del número inicial de especies con la variable objetivo, esto será favorable para la predicción ya que será la variable que más aporta:

Ilustración 15. Matriz de correlaciones - Nuevas variables

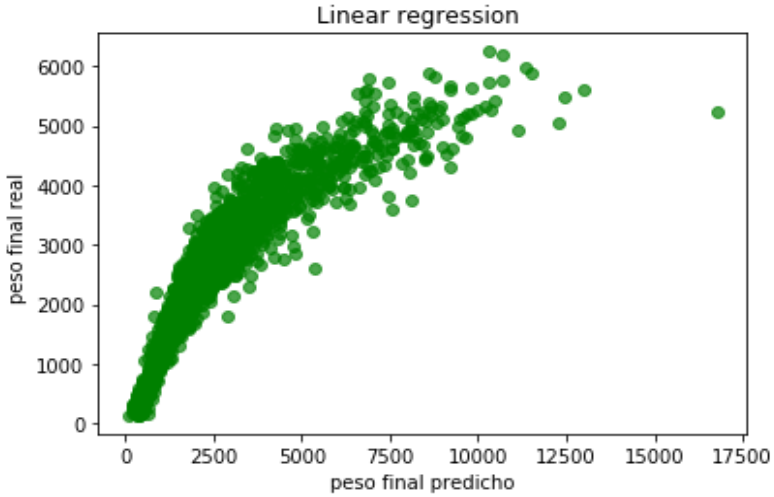
Fuente: Elaboración propia.

Se puede observar que las variables han sido reagrupado en base a su alta correlación.

A. MODELO DE REGRESIÓN LINEAL:

Se usó solo las variables numéricas, por lo que se cuenta con 9 variables explicativas (N_Mortalidad, N_Inicial, N_Final, Peso_Promedio_Inicial, Alimento_Usado_Kg, Dias_alimentados, Profundidad, Horas_luz, Prom_Temperatura, Prom_Oxigeno).

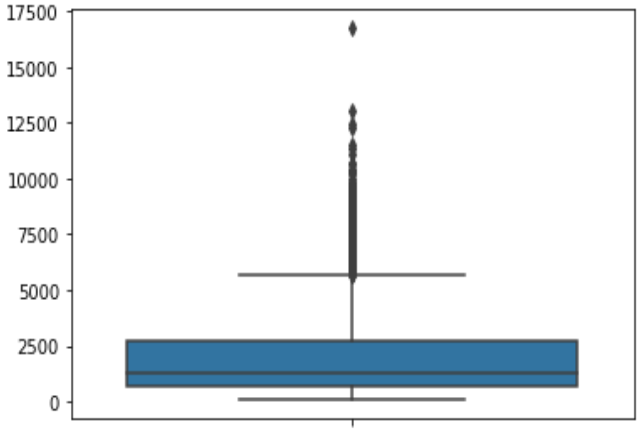
Ilustración 16. Gráfico Peso promedio final Real vs Predecido - Regresión Lineal



Fuente: Elaboración propia.

Los valores predichos por el modelo de regresión lineal tienen una media de 1400 gramos de peso promedio de las especies al final de las especies, además los valores se encuentran entre 900 y 2800 gramos.

Ilustración 16. Diagrama de caja - Peso promedio final predicho

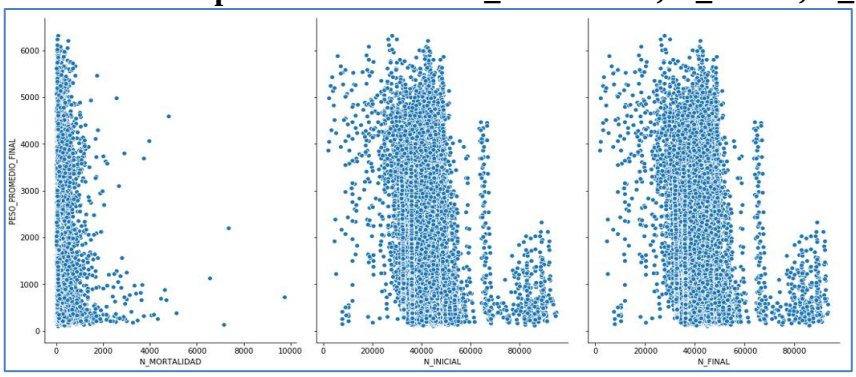


Fuente: Elaboración propia.

Adicionalmente se calculó los supuestos que debería cumplir para aplicar el modelo de manera correcta:

A.1. Linealidad:

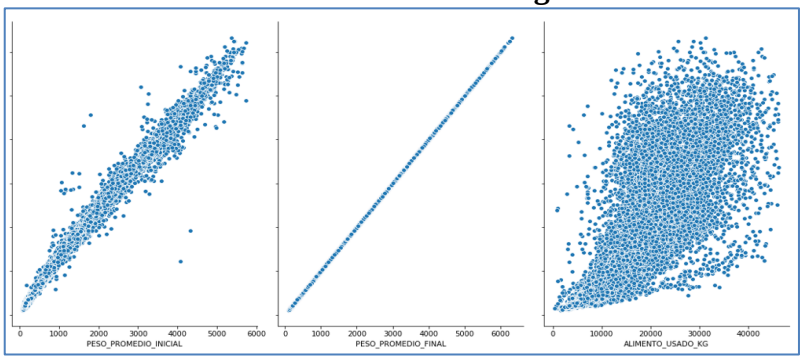
Ilustración 17. Peso promedio final vs N_Mortalidad, N_Inicial, N_final



Fuente: Elaboración propia.

En la primera gráfica, se puede observar que no existe linealidad entre las 4 variables en estudio.

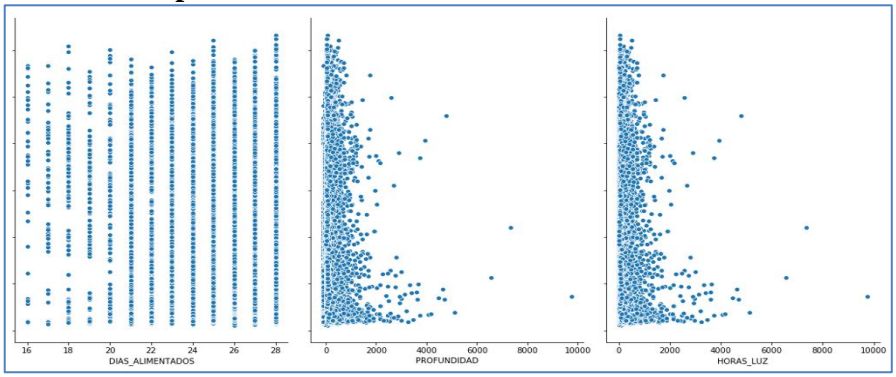
Ilustración 18. Peso promedio final vs Peso promedio inicial, Peso promedio final, Alimento usado en kg



Fuente: Elaboración propia.

Para este gráfico, los pesos promedio se distribuyen de manera lineal; sin embargo, el alimento usado no muestra linealidad en su distribución.

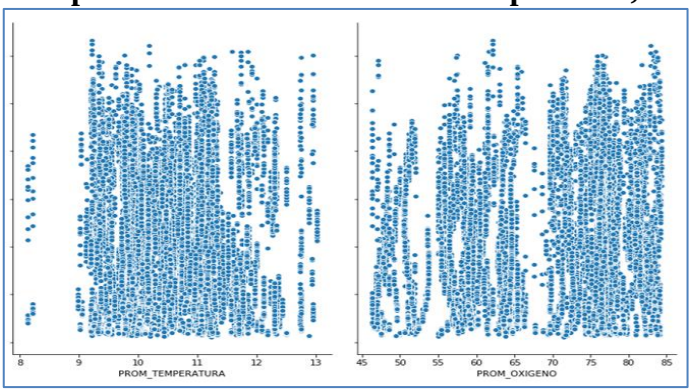
Ilustración 19. Peso promedio final vs Dias alimentados, Profundidad, Horas luz



Fuente: Elaboración propia.

En esta gráfica, tampoco se observa linealidad en las variables.

Ilustración 20. Peso promedio final vs Promedio temperatura, Promedio oxigeno

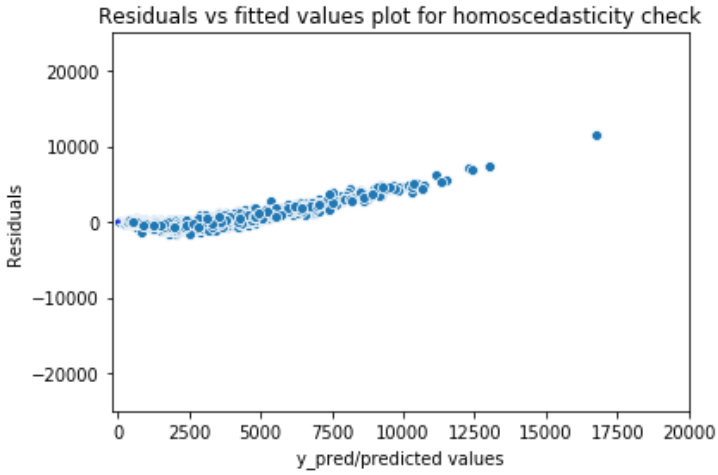


Fuente: Elaboración propia.

El peso final, promedio temperatura y promedio, tampoco se distribuyen de manera lineal.

A.2. Homocedasticidad:

Ilustración 21. Residuales vs Valores predecidos

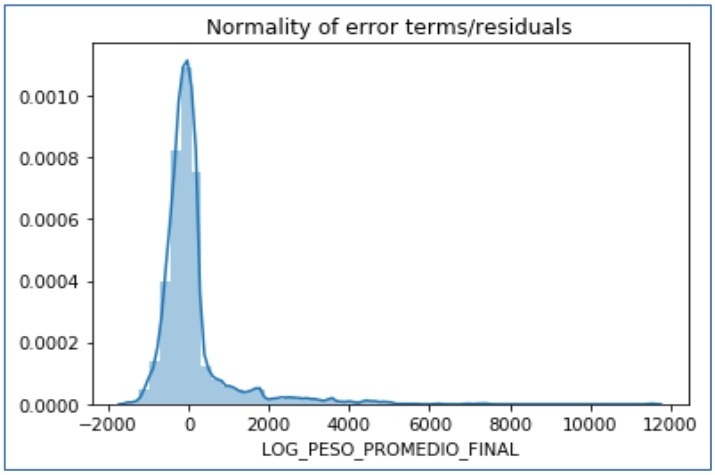


Fuente: Elaboración propia.

Se puede observar, que los residuales tienen varianza constante, respecto a los valores predecidos.

A.3. Normalidad:

Ilustración 21. Gráfica de la normalidad de los residuales

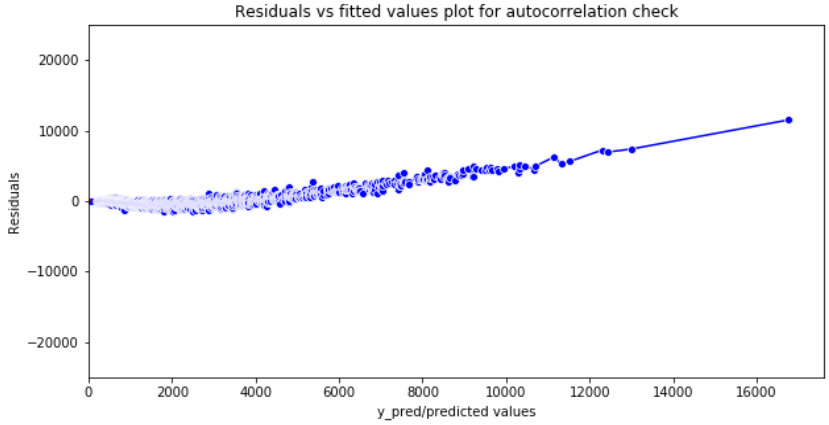


Fuente: Elaboración propia.

Se puede visualizar, que los residuales se distribuyen de manera normal.

A.4. Autocorrelación:

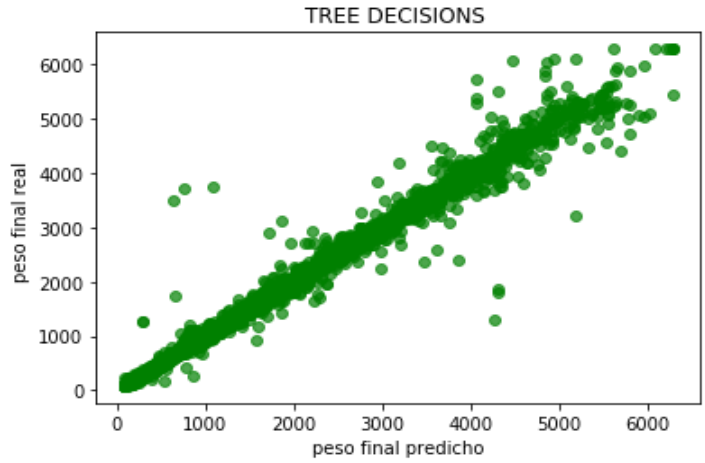
Ilustración 22. Residuales vs valores ajustados



Fuente: Elaboración propia.

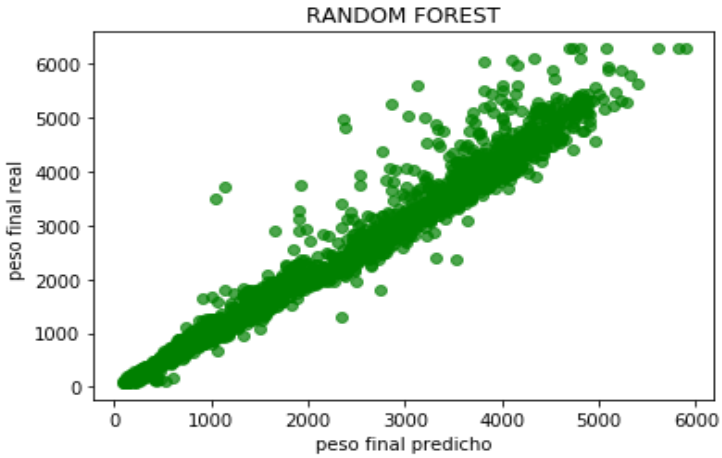
Relación de residuales versus los valores ajustados.

B. MODELO ARBOL DE DECISION:



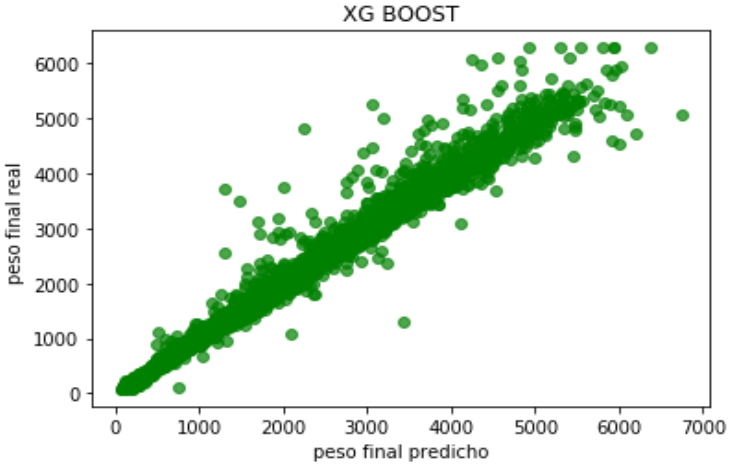
Fuente: Elaboración propia.

C. MODELO RANDOM FORST:



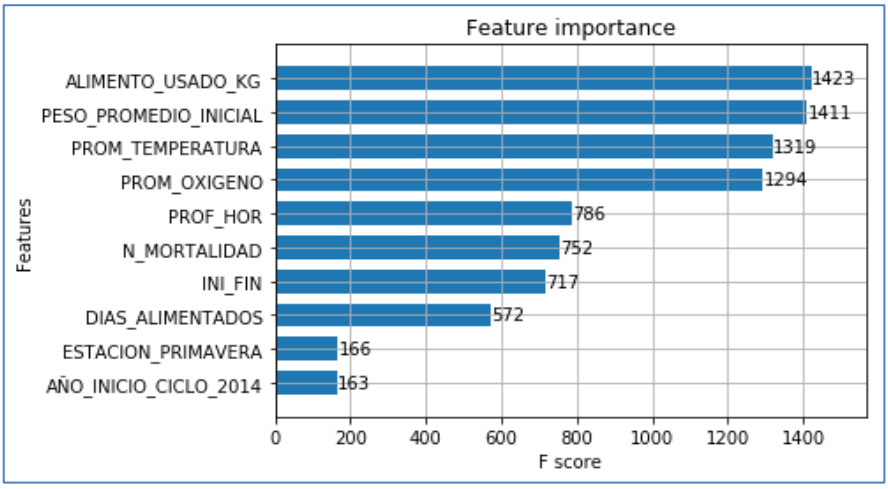
Fuente: Elaboración propia.

D. MODELO XGBOOST:



Fuente: Elaboración propia.

E. IMPORTANCIA DE VARIABLES:



Fuente: Elaboración propia.

Se observa una ponderación de los pesos de cada variable dentro del modelo

F. COMPARACIÓN DE INDICADORES:

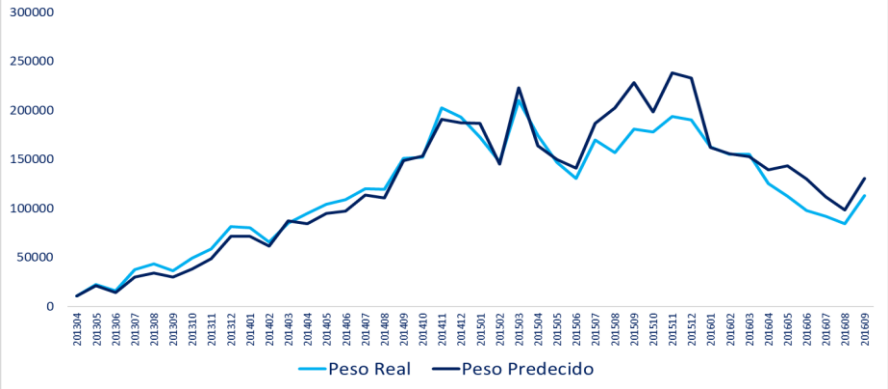
INDICADOR	REGRESIÓN LINEAL	ARBOL DE DECISION	RANDOM FOREST	XGBOOST
RMSE	921.47	188.07	262.23	210.56
MAPE	25.43	5.2	8.66	7.05
MAE	482.63	75.39	127.7	101.33

Fuente: Elaboración propia.

En conclusión, el mejor modelo que predice el peso promedio final en gramos es el árbol de decisión de acuerdo a los tres indicadores calculados.

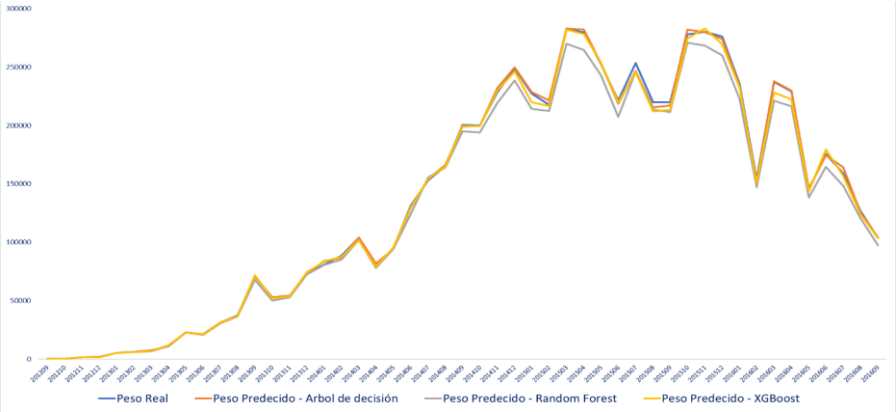
G. COMPARACIÓN TEMPORAL DE MODELOS:

G.1. Modelo Regresión Lineal:



Fuente: Elaboración propia.

G.2. Modelos (Arbol de decisión, Random Forest, XGBoost):

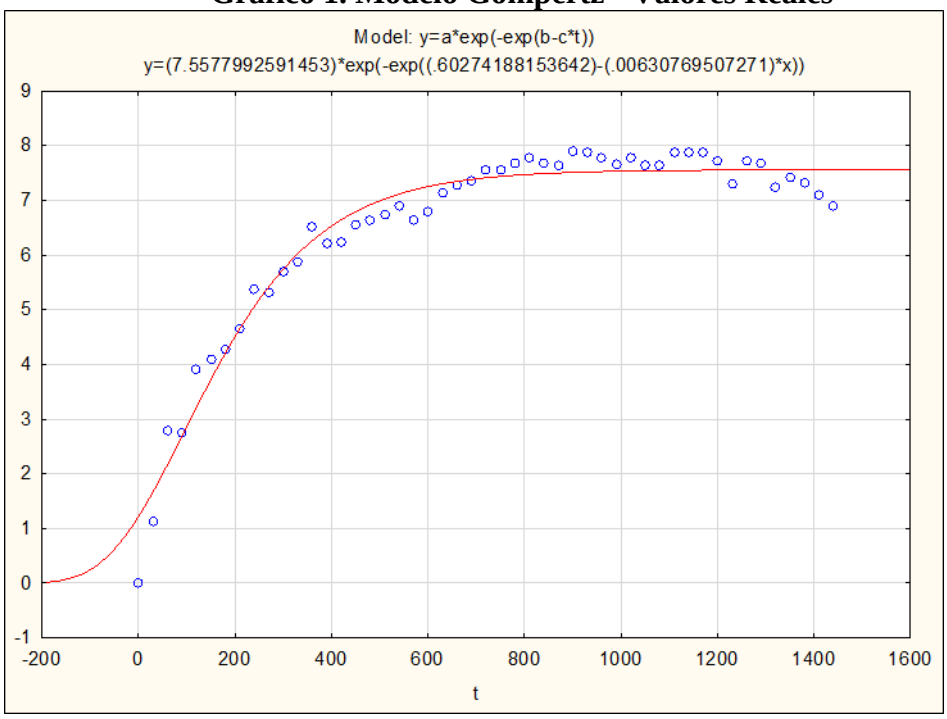


Fuente: Elaboración propia.

H. MODELO DE CRECIMIENTO:

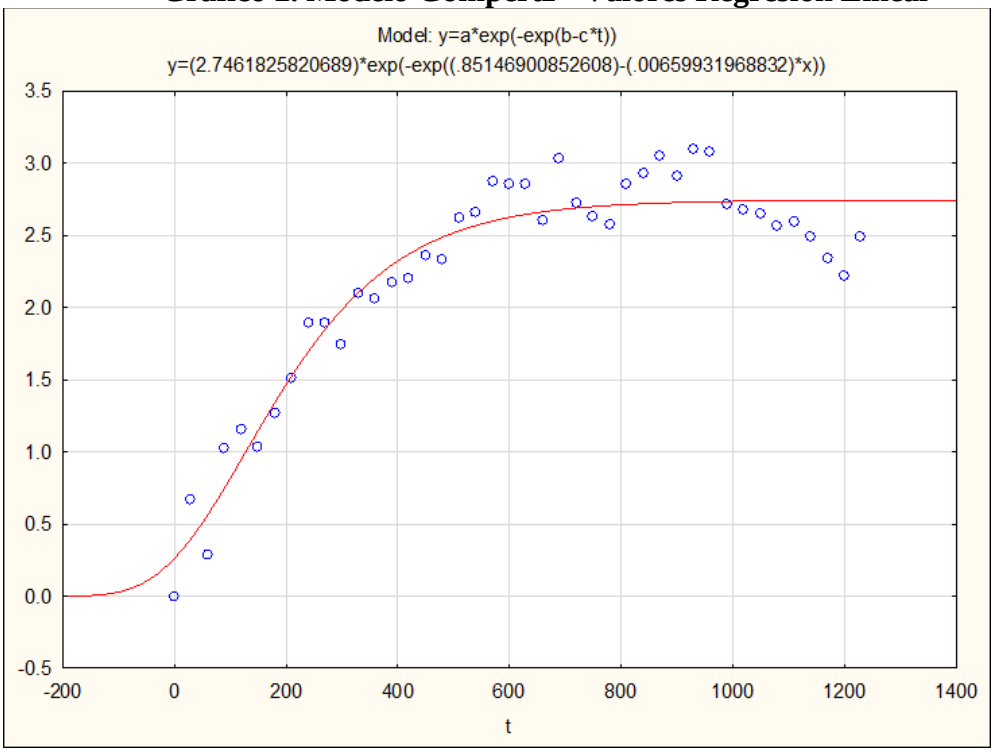
A continuación, se mostrarán las gráficas de crecimiento utilizando el modelo Gompertz para los valores reales y los que se obtuvieron usando los modelos de predicción de Regresión Lineal, Árbol de decisión, Random Forest y XGBoost.

Gráfico 1. Modelo Gompertz - Valores Reales



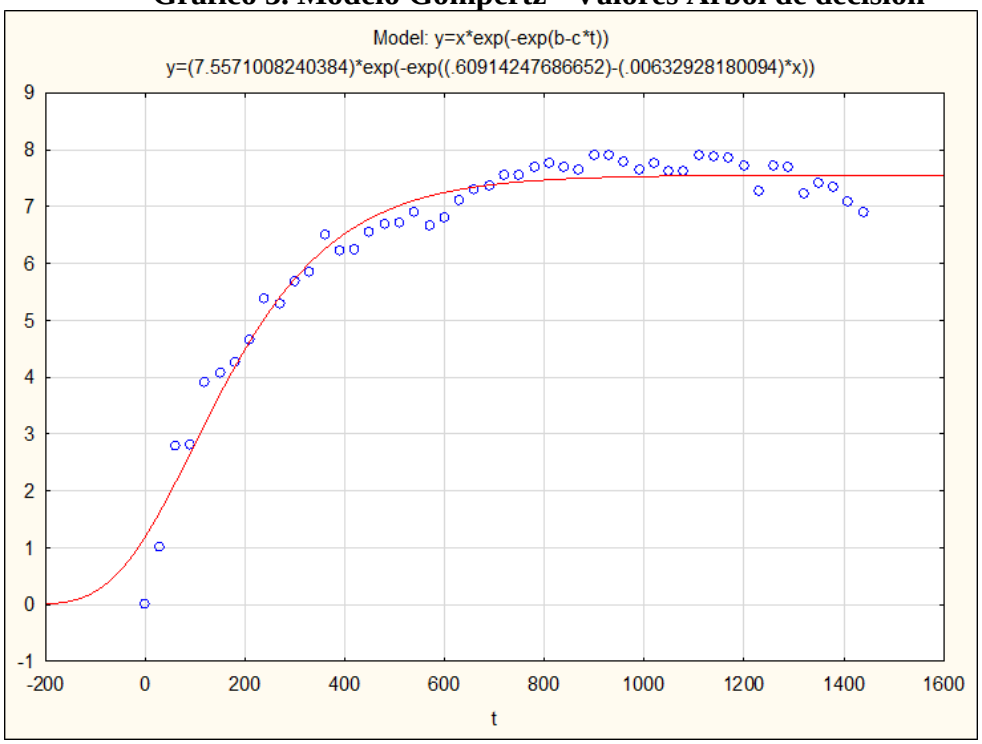
Fuente: Elaboración propia – uso del software Statisca 12

Gráfico 2. Modelo Gompertz - Valores Regresión Lineal



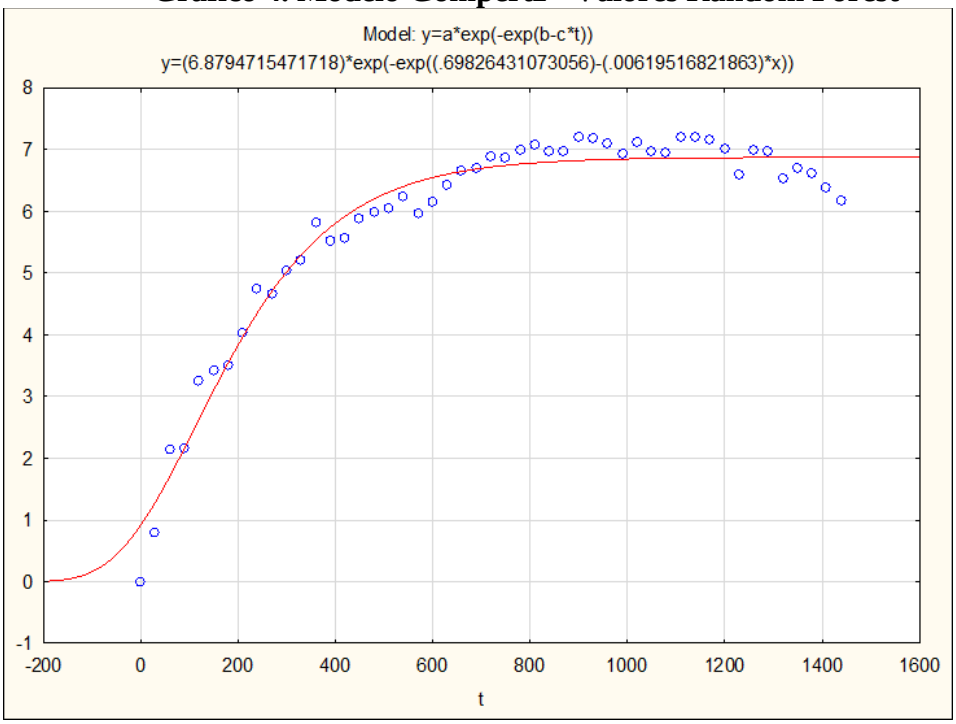
Fuente: Elaboración propia – uso del software Statisca 12

Gráfico 3. Modelo Gompertz - Valores Árbol de decisión



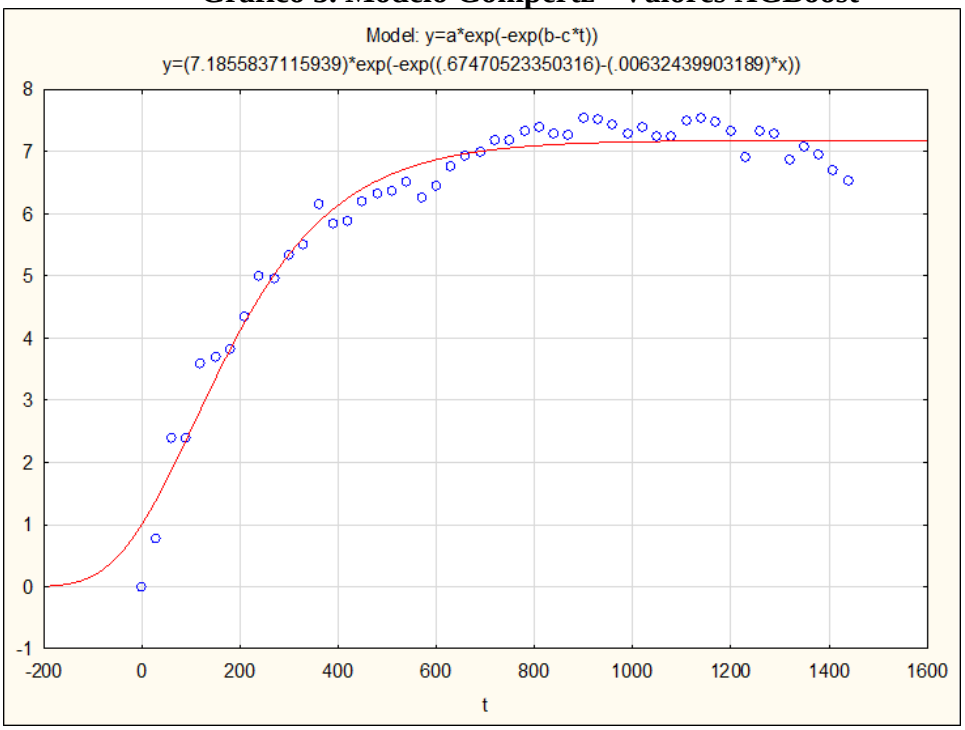
Fuente: Elaboración propia – uso del software Statisca 12

Gráfico 4. Modelo Gompertz - Valores Random Forest



Fuente: Elaboración propia – uso del software Statisca 12

Gráfico 5. Modelo Gompertz - Valores XGBoost



Fuente: Elaboración propia – uso del software Statisca 12

Se puede observar que las gráficas se comportan de manera similar, lo único que cambia de forma mínima son sus parámetros que intervienen en el modelo de crecimiento de Gompertz, estos parámetros son obtenidos en el software "Statistica 12", donde utilizan

la estimación No Lineal de Quasi Newton. Ahora, se mostrarán los valores de estos parámetros.

En primer lugar, se tiene el modelo de crecimiento de Gompertz:

$$y = a * \exp(-\exp(b - c * t))$$

Tabla 2. Parámetros según modelos empleados

		Parámetros		
Modelo de Crecimiento	Modelo de Predicción	a	b	c
Gompertz	Pesos Reales	7.55780	0.60274	0.00631
	Regresión Lineal	2.74618	0.85147	0.00660
	Árbol de Decisión	7.55710	0.60914	0.00633
	Random Forest	6.87947	0.69826	0.00620
	XGBoost	7.18558	0.67471	0.00632

Fuente: Elaboración propia – uso del software Statisca 12

Ahora observaremos los indicadores R^2, AIC y SCR para el modelo real y los cuatro modelos de predicción que se emplearon para contrastar, como bien se puede observar estos valores son muy cercanos. A pesar que, los valores del modelo de Regresión Lineal tienen mejores indicadores, cabe resaltar que no entra en nuestro contraste debido al tamaño de muestra, ya que se tiene menor número de data de acuerdo a la base que nos fue proporcionada por la empresa Alicorp. Finalmente, entre los otros tres modelos restantes elegiremos al modelo Random Forest ya que presenta indicadores AIC y SCR un poco menores que los otros modelos, asimismo con el R^2 no se puede concluir nada ya que todos los modelos presentan valores adecuados.

Tabla 3. Parámetros según modelos empleados

Modelo de Crecimiento	Modelo de Predicción	Tamaño de muestra (n)	R^2	SCR	AIC
Gompertz	Pesos Reales	49	96.19%	6.02	-96.71
	Regresión Lineal	42	91.95%	1.98	-122.32
	Árbol de Decisión	49	96.16%	6.11	-96.00
	Random Forest	**49**	**96.65%**	**4.85**	**-107.35**
	XGBoost	49	96.43%	5.47	-101.45

Fuente: Elaboración propia – uso del software Statisca 12

5. CONCLUSIONES

Las variables número inicial y final de peces, peso promedio inicial y final de las especies marinas, el alimento en kg, días alimentados tienen una distribución exponencial, además las variables en relación al promedio de temperatura y oxigeno presentan distribución normal.

En relación a las variables categóricas, se notó una mayor proporción de peces pertenecientes a una especie en comparación a las otras dos especies. Se tuvieron dos casos en cuanto a los datos perdidos y atípicos, en primera instancia se eliminó a todos aquellos que presentan estas casuísticas, esto se decidió para obtener mejores resultados al evaluar el modelo de Regresión Lineal. En segunda instancia, se reemplazó a todos los datos perdidos con la media por variable, y a los datos atípicos se remplazó con el mínimo y máximo cuartil, esto se hizo para evaluar los modelos Árbol de decision, Random Forest y XGBoost.

Las variables promedio de temperatura y oxigeno tuvieron mayor cantidad de datos perdidos con 13% aproximadamente, la otra variable con mayores datos perdidos es el peso promedio inicial. Al igual que en los datos perdidos, también se tiene un alto porcentaje de datos atípicos en las variables relacionadas al número de especies que murieron, cantidad de horas luz artificial, días alimentados, profundidad en metros, cantidad inicial y final de peces.

Al analizar la matriz de correlación se tuvo como estrategia crear dos nuevas variables, la primera en relación al número inicial y final de peces, la segunda en función a la profundidad y horas luz. En conclusión, las variables que se encuentran más correlacionadas con la variable objetivo (peso promedio final de peces) son peso promedio inicial y alimento usado en kg.

Desarrollando el modelo de regresión lineal se obtuvo un error cuadrático medio de 921, un error porcentual absoluto medio de 25.43, y un error absoluto medio de 482. También se analizó el cumplimiento de los supuestos de linealidad, homocedasticidad, normalidad y autocorrelación. El supuesto de linealidad no se cumple en su totalidad ya que no todas las variables que entran al modelo tienen una relación lineal con el peso promedio final de los peces. En cuanto a la homocedasticidad, la variación de todos los residuos no es uniforme en todo el rango de valores de los pronósticos, sin embargo, en la mayor parte del rango si cumple el supuesto. En cuanto a la normalidad, de acuerdo al grafico calculado se muestra el cumplimiento de este supuesto.

Desarrollando los modelos de Arbol de decision, Random Forest y XGBoost se determinó que el mejor modelo que se ajusta a los valores reales de los pesos promedio final de los peces es el modelo Árbol de decisión con un error cuadrático medio de 188, un error porcentual absoluto medio de 5.2, y un error absoluto medio de 75.39. Adicionalmente se

calculó la importancia de variables, determinando que las variables que más influyen en el peso promedio final de los peces es el alimento usado en kg, peso promedio inicial, temperatura y oxígeno, por lo que se debería enfocar más en estas variables si se quiere optimizar el peso de los peces. Para desarrollar los modelos, se obtuvo una muestra aleatoria como parte de la prueba, en la que se vio que el peso promedio final de las especies marinas tuvo un crecimiento elevado en el 2015.

En base a los indicadores hallado para el modelo de predicción de pesos de la especie marina se elegirá realizar Árbol de decisión ya que es la que más se ajusta a los datos reales. Adicionalmente se concluye que la variable que más afecta al peso promedio final de las especies es el peso promedio inicial y el alimento en gramos. Asimismo, preliminarmente se está considerando que los modelos de crecimiento de Gompertz para los dos modelos de predicción (árbol de decisión y random forest) y el los pesos reales tienen un valor adecuado de R2 sin embargo posteriormente se revisará otros indicadores como AIC, BIC.

6. BIBLIOGRAFÍA

Arys Carrasquilla-Batista, A. C.-R. (2016). Regresión lineal simple y múltiple: aplicación en la predicción de variables naturales relacionadas con el crecimiento microalgal.

Gloria A Casas 1*, M. M., Daniel Rodríguez1, Z., & Germán Afanador Téllez1, M. M. (2010). Propiedades matemáticas del modelo de Gompertz y su aplicación al crecimiento de los cerdos. Universidad de Antioquia, 3.

Kathleen M. C. Tjørve, E. T. (2017). The use of Gompertz models in growth analyses, and new Gompertz-model approach: An addition to the Unified-Richards family. PLOS ONE.

La O Arias, M. A., Guevara, F., Fonseca, N., Rodríguez, L., Pinto, R., Gómez, H., & Medina, F. J. (2013). Aplicación de los modelos logístico y Gompertz al análisis de curvas de peso vivo en cabritos criollos. Revista Cubana de Ciencia Agrícola, 5.

María E. Cayré (1), G. M. (2007). Selección de un Modelo Primario para Describir la Curva de Crecimiento de Bacterias Lácticas y Brochothrix thermosphacta sobre Emulsiones Cárnicas Cocidas. SCielo.

Osborne, J. &. (2002). Four assumptions of multiple regression that researchers should always test.

Tabachnick, B. G. (1996). Using Multivariate Statistics (3rd ed.). New York.

Williams, M. G. (2013). Assumptions of multiple regression: correcting two misconceptions.

7. APENDICE

1.Estandares:

Cross Industry Standard Process for Data Mining

Es una metodología internacional utilizada para los procesos que sigue la minería de datos, que se encarga del descubrimiento de información valiosa para una organización, que se encuentra oculta dentro de las bases de datos. Para ello utiliza los métodos y algoritmos de estadística, aprendizaje automático e inteligencia artificial como lo son los modelos de Random Forest. Consta de las siguientes etapas:

Entendimiento del negocio: Esta fase inicial se centra en la comprensión de los objetivos y requerimientos del proyecto desde una perspectiva empresarial, convirtiendo luego este conocimiento en una definición de problema de DM y un plan preliminar diseñado para lograr los objetivos.

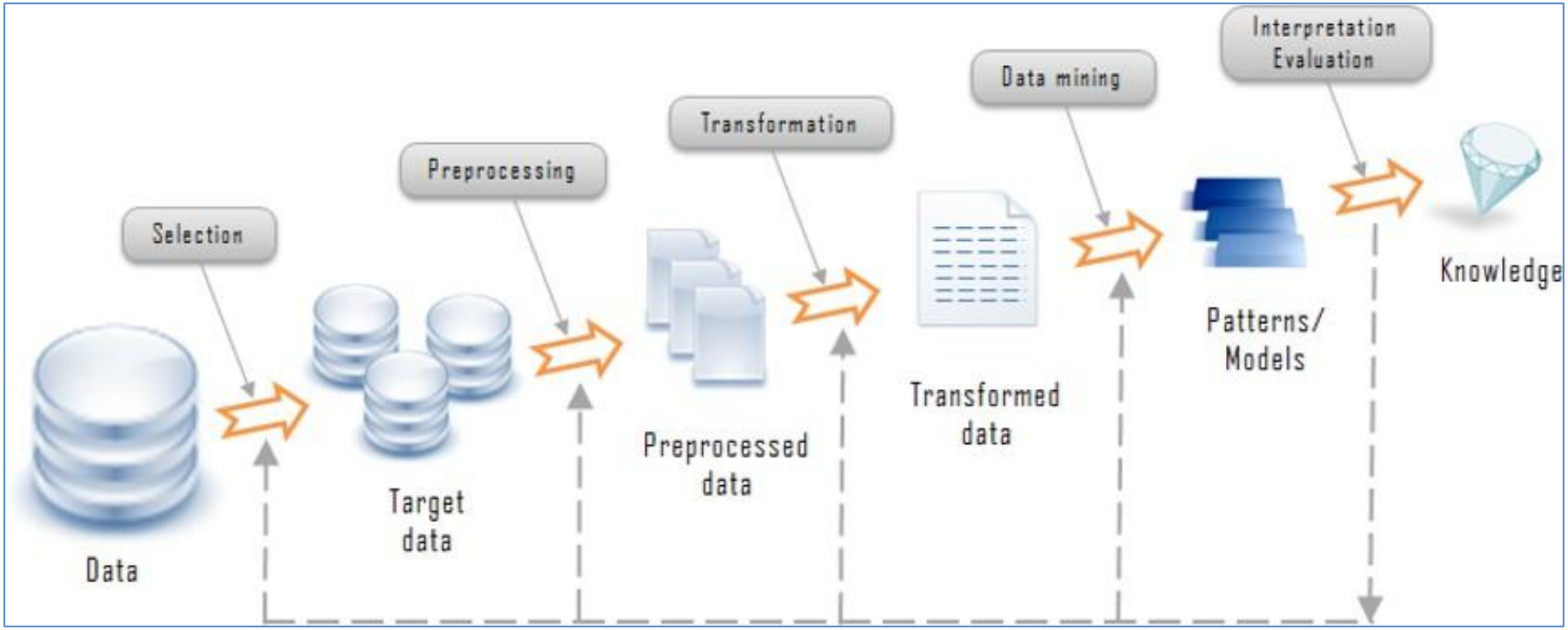
Comprensión de los datos: La fase de comprensión de los datos comienza con una recopilación inicial de datos y prosigue con las actividades para familiarizarse con los datos, identificar problemas de calidad de los datos, descubrir las primeras percepciones de los datos o detectar subconjuntos interesantes para formar hipótesis de información oculta .

Preparación de datos: La fase de preparación de datos cubre todas las actividades para construir el conjunto de datos final a partir de los datos iniciales sin procesar; Modelado: en esta fase, se seleccionan y aplican diversas técnicas de modelado y se calibran sus parámetros a valores óptimos.

Evaluación: En esta etapa el modelo o modelos obtenidos son evaluados más a fondo y los pasos ejecutados para construir el modelo son revisados para asegurarse de que alcanza adecuadamente los objetivos de negocio.

Despliegue: La creación del modelo generalmente no es el final del proyecto. Incluso si el propósito del modelo es aumentar el conocimiento de los datos, el conocimiento adquerido tendrá que ser organizado y presentado de manera que el cliente pueda utilizarlo.

Esquema de etapas de CRISP DM



Esquema del proceso de minería de datos bajo la metodología de CRISP DM

UNITED NATIONS: Methodologies and Procedures

- GPS – Guide to Presenting Statistics
  Este estándar nos indica cómo hacer comprensibles nuestros datos y resultados.
  Usamos este estándar, al momento de presentar nuestros resultados finales y realizar la comparación correspondiente, ya que estos deben ser comprensibles por cualquier lector.
- DMDRP – Data and Meta Data Reporting and Presentation
  Hicimos uso de este estándar al momento de presentar las descripciones de cada variable de la data que nos fue compartida, asimismo se presentaron tablas como la matriz de operacionalización para resumir a las variables, tablas para la comparación de resultados obtenidos, etc.

DATA BASE STANDARDS

- UNE Referential Data Base
  Se tomó en cuenta al momento de definir nuestras variables que nos ayudaban a juntar las bases de datos que nos fueron proporcionadas

ISO-19115

el objetivo de este estándar internacional es proporcionar un procedimiento claro para la descripción del conjunto de datos geográficos, de forma que los usuarios puedan determinar si los datos les serán útiles y como acceder a ellos.

Mediante el establecimiento de un conjunto común de terminología de metadatos y definiciones, este estándar promueve el uso apropiado y el intercambio efectivo de la información geográfica.

Este estándar tiene otros beneficios, dado que facilita la organización y mantenimiento de los datos y proporciona información sobre el conjunto de datos de una organización hacia las otras.

La ISO 19115 se complementa por la ISO 19115-2, la cual define los elementos de los metadatos destinados a describir imágenes y datos raster.

2.Limitaciones, ventajas y desventajas:

2.1. Limitaciones:

Para realizar este proyecto, se tuvo como principal limitación la base de datos que nos proporcionó la empresa Alicorp. Esto significó no contar con mayor información sobre la especie, genética de los peces a los cuales Vitapro proporciona sus alimentos balanceados, asimismo no se pudo conocer los establecimientos donde se llevan a cabo este proceso de alimentación a los peces, debido a que se encuentra fuera de Lima.

Otra limitación que se tuvo fue que la empresa no nos brindó información acerca de los estándares que manejan para realiza el proceso de alimentación a los peces, ya que existen ciertas normas que la FAO, organismo de la ONU, exige que las empresas que se encargan

en la producción de alimentos balanceados a animales, así como también normas que exigen la buena práctica de la acuicultura donde se incluye el proceso de alimentación.

2.1. Ventajas y desventajas:

En este proyecto se tuvieron desventajas tales como: comprensión de los modelos de crecimiento, ya que existen diversas formas de expresar la curva de crecimiento de Gompertz, ya que se aplica a muchos otros campos como medicina, biología, microbiología, etc. Asimismo, al intentar explicar nuestro modelo de regresión solo se pudo utilizar las variables numéricas, ya que, si usábamos nuestras variables categóricas, habiendo realizado el tratamiento de dummies, nuestros resultados obtenidos no eran concluyentes según nuestros análisis estadísticos.

La principal ventaja que se tuvo fue la aplicación de modelo Machine Learning como el Árbol de Decisión, Random Forest y el XGBoost. Estos modelos fueron ejecutados en el Software Python, el cual hoy en día es muy usado para realizar modelos estadísticos de predicción, esto es por su agradable interfaz y fácil comprensión de resultados.

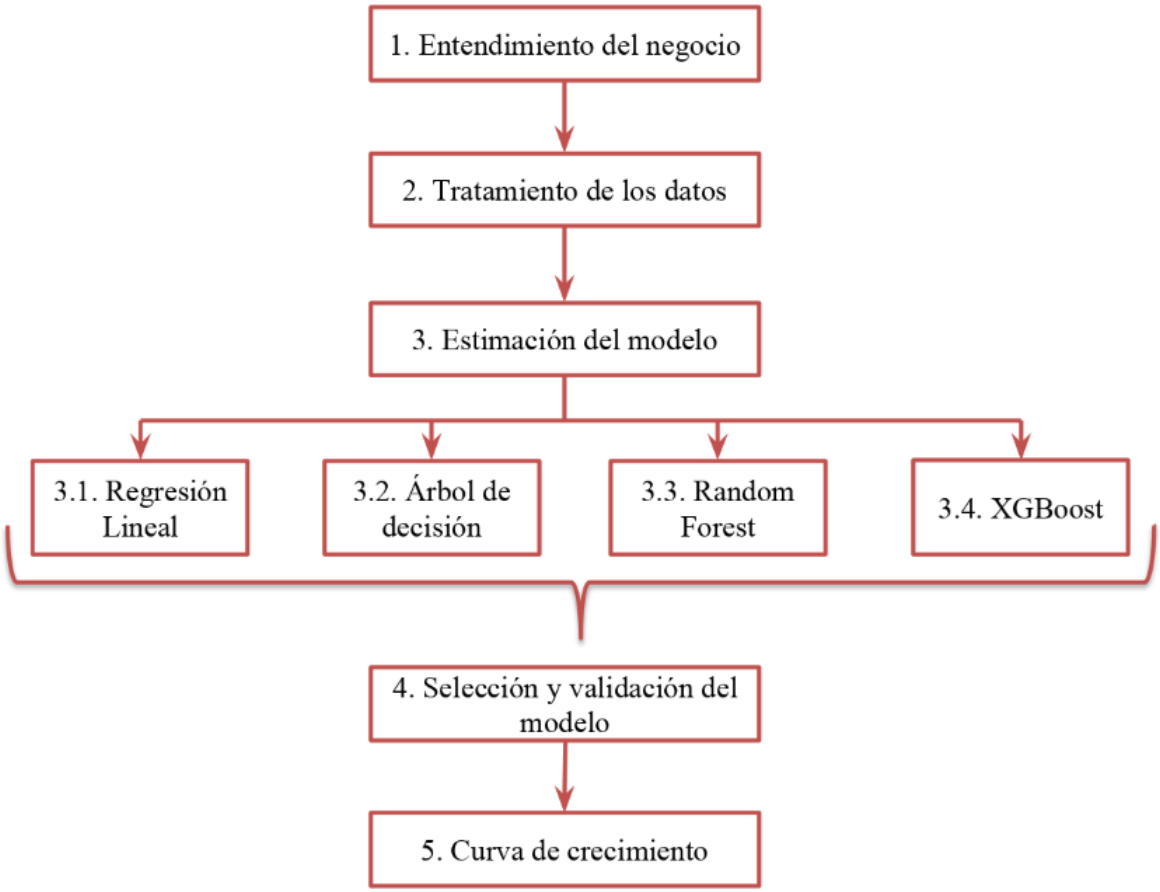
Por otro lado, las ventajas que se pueden obtener luego de seleccionar el mejor modelo de predicción de pesos promedio de peces y el mejor modelo que se ajusta a la curva de crecimiento de Gompertz son:

- Conocer que variables son más influyentes en el peso promedio final de los peces.
- Explicar mediante la curva de crecimiento de Gompertz el aumento de peso en los peces de manera mensual.
- Según los datos reales de los pesos promedio final de los peces, conocer el mejor modelo de predicción de los pesos promedio final de los peces.
- Implementar estrategias comerciales para mejora su curva de crecimiento de peces.

Las desventajas serían:

- El posible alto costo que generaría realizar las mejoras a los procesos en los cuales están involucradas las variables más influyentes en el peso promedio final de los peces.

4.Diagrama de flujo (Descripción de la solución):



5.Programación en Python:

1.1.Importación e construcción de data:

Se importo la data con información de factores ambientales:

dfa=pd.read_csv("PROM_AMBIENTAL.csv", sep=";")

dfa.head(5)

	COD_CENTRAL	PROM_TEMPERATURA	PROM_OXIGENO
0	CENTRO01_201304	11.026314	82.057855
1	CENTRO01_201305	11.068411	82.009585
2	CENTRO01_201306	10.947576	81.856395
3	CENTRO01_201307	10.505184	81.518301
4	CENTRO01_201308	9.936811	81.058755

Se importo la data con información de factores productivos:

dfp=pd.read_csv("BBDD_PRODUCTIVA.csv", sep=";", encoding='latin1')

dfp.head(5)

	NOM_EMPRESA	NOM_CENTRO	NOM_JAULA	AÑO_INICIO_CICLO	MES_INFORMACION	PERIODO	COD_CENTRAL	N_MORTALIDAD
0	CLIENTE01	CENTRO01	JAULA01	2013	Set-13	201309	CENTRO01_201309	159.0
1	CLIENTE01	CENTRO01	JAULA01	2013	Oct-13	201310	CENTRO01_201310	117.0
2	CLIENTE01	CENTRO01	JAULA01	2013	Ene-13	201301	CENTRO01_201301	93.0
3	CLIENTE01	CENTRO01	JAULA01	2013	Feb-13	201302	CENTRO01_201302	57.0
4	CLIENTE01	CENTRO01	JAULA01	2013	Mar-13	201303	CENTRO01_201303	66.0

5 rows × 23 columns

#uniendo las bases de factores productivos y ambientales en función del cod_central creado #para facilitar la unión de ambas bases:

df_alicorp=pd.merge(dfp, dfa, on='COD_CENTRAL', how='left')

df_alicorp.head(10)

	NOM_EMPRESA	NOM_CENTRO	NOM_JAULA	AÑO_INICIO_CICLO	MES_INFORMACION	PERIODO	COD_CENTRAL	N_MORTALIDAD
0	CLIENTE01	CENTRO01	JAULA01	2013	Set-13	201309	CENTRO01_201309	159.0
1	CLIENTE01	CENTRO01	JAULA01	2013	Oct-13	201310	CENTRO01_201310	117.0
2	CLIENTE01	CENTRO01	JAULA01	2013	Ene-13	201301	CENTRO01_201301	93.0
3	CLIENTE01	CENTRO01	JAULA01	2013	Feb-13	201302	CENTRO01_201302	57.0
4	CLIENTE01	CENTRO01	JAULA01	2013	Mar-13	201303	CENTRO01_201303	66.0
5	CLIENTE01	CENTRO01	JAULA01	2013	Abr-13	201304	CENTRO01_201304	65.0
6	CLIENTE01	CENTRO01	JAULA01	2013	May-13	201305	CENTRO01_201305	86.0
7	CLIENTE01	CENTRO01	JAULA01	2013	Jun-13	201306	CENTRO01_201306	127.0
8	CLIENTE01	CENTRO01	JAULA01	2013	Jul-13	201307	CENTRO01_201307	202.0
9	CLIENTE01	CENTRO01	JAULA01	2013	Ago-13	201308	CENTRO01_201308	214.0

10 rows × 25 columns

#Se transformará algunas variables numéricas a categórica y se eliminará las innecesarias:

df_alicorp['PERIODO']=df_alicorp['PERIODO'].astype('object')

df_alicorp['AÑO_INICIO_CICLO']=df_alicorp['AÑO_INICIO_CICLO'].astype('object')

df_alicorp.drop(['MES_INFORMACION'], axis=1,inplace=True)

1.2.Análisis Exploratorio:

1.2.1. Distribución por variables (numéricas y categóricas):

#Creamos la variable df_num que excluye las variables categóricas del df:

df_alicorp_num = df_alicorp.select_dtypes(exclude = 'O')

df_alicorp_num.head(5)

df_alicorp_num.shape

##Para las variables numéricas

```
for i in df_alicorp_num.columns:
    if df_alicorp_num[i].dtypes != 'O':
        print(i)
        plt.hist(df_alicorp_num[i], bins = 50, edgecolor = 'k')
```

```
        plt.show()
    else:
            print(i)
            (pd.DataFrame(df_alicorp_num[i].value_counts())).plot.bar(y = i)
            plt.show()

#Creamos la variable df_cat que excluye las variables categóricas del df:
df_alicorp_cat = df_alicorp.select_dtypes(include = 'O')
df_alicorp_cat.head(5)
df_alicorp_cat.shape #13 variables
##Para las variables categoricas
for i in df_alicorp_cat.columns:
    if df_alicorp_cat[i].dtypes != 'O':
        print(i)
        plt.hist(df_alicorp_cat[i], bins = 50, edgecolor = 'k')
        plt.show()
    else:
        print(i)
        (pd.DataFrame(df_alicorp_cat[i].value_counts())).plot.bar(y = i)
        plt.show()
```

1.2.2. Correlación de las variables:

```
#Para visualizar mejor las correlaciones de las variables, pintamos los valores mas representativos y damos formato a las anotaciones.
df_alicorp_corr = df_alicorp.corr()
plt.figure(figsize=(10,10))
sns.heatmap(df_alicorp_corr, annot=True, fmt='.2f', cmap='YlGnBu')
plt.show()
```

1.3. Tratamiento de datos perdidos y outliers:

1.3.1. Datos perdidos:

```
df_alicorp.isnull()

serie_nulos=df_alicorp.isnull().sum()

df_nulos=pd.DataFrame(serie_nulos, columns=['CantNulos'])

nCol=df_alicorp.shape[1]

nRows=df_alicorp.shape[0]

df_nulos['PorcNulos']=round(df_nulos['CantNulos']/nRows,4)*100

df_nulos

df_nulos['Porcentaje'] = round(df_nulos['CantNulos']*100 / df_alicorp.shape[0], 2)

df_nulos[df_nulos['CantNulos'] > 0]  #No se elimina ninguna variable cat o num pq tienen menor de 14% datos perdidos

#9 variables numericas y 5 categoricas

### Tratamiento de missings (VAR CATEGORICAS) - se imputará con la moda

df_alicorp["GENETICA"] = df_alicorp["GENETICA"].fillna(df_alicorp["GENETICA"].mode()[0])

df_alicorp["SISTEMA_OXIGENO"] = df_alicorp["SISTEMA_OXIGENO"].fillna(df_alicorp["SISTEMA_OXIGENO"].mode()[0])

df_alicorp["TIPO_ALIMENTACION"] = df_alicorp["TIPO_ALIMENTACION"].fillna(df_alicorp["TIPO_ALIMENTACION"].mode()[0])

df_alicorp["TIPO_JAULA"] = df_alicorp["TIPO_JAULA"].fillna(df_alicorp["TIPO_JAULA"].mode()[0])

df_alicorp["FOTOPERIODO"] = df_alicorp["FOTOPERIODO"].fillna(df_alicorp["FOTOPERIODO"].mode()[0])

serie_nulos=df_alicorp.isnull().sum()

df_nulos=pd.DataFrame(serie_nulos, columns=['CantNulos'])

df_nulos[df_nulos['CantNulos'] > 0]


### Tratamiento de missings (VAR NUMERICAS) - se imputará con la media

df_alicorp["N_MORTALIDAD"] = df_alicorp["N_MORTALIDAD"].fillna(df_alicorp["N_MORTALIDAD"].mean())
```

```
df_alicorp["N_INICIAL"] = df_alicorp["N_INICIAL"].fillna(df_alicorp["N_INICIAL"].mean())

df_alicorp["N_FINAL"] = df_alicorp["N_FINAL"].fillna(df_alicorp["N_FINAL"].mean())

df_alicorp["PESO_PROMEDIO_INICIAL"] = df_alicorp["PESO_PROMEDIO_INICIAL"].fillna(df_alicorp["PESO_PROMEDIO_INICIAL"].mean())

df_alicorp["ALIMENTO_USADO_KG"] = df_alicorp["ALIMENTO_USADO_KG"].fillna(df_alicorp["ALIMENTO_USADO_KG"].mean())

df_alicorp["DIAS_ALIMENTADOS"] = df_alicorp["DIAS_ALIMENTADOS"].fillna(df_alicorp["DIAS_ALIMENTADOS"].mean())

df_alicorp["PROFUNDIDAD"] = df_alicorp["PROFUNDIDAD"].fillna(df_alicorp["PROFUNDIDAD"].mean())

df_alicorp["PROM_TEMPERATURA"] = df_alicorp["PROM_TEMPERATURA"].fillna(df_alicorp["PROM_TEMPERATURA"].mean())

df_alicorp["PROM_OXIGENO"] = df_alicorp["PROM_OXIGENO"].fillna(df_alicorp["PROM_OXIGENO"].mean())

serie_nulos=df_alicorp.isnull().sum()

df_nulos=pd.DataFrame(serie_nulos, columns=['CantNulos'])

df_nulos[df_nulos['CantNulos'] > 0] #YA NO HAY DATOS PERDIDOS
```

1.3.2. Datos outliers:

```
##Para las variables numéricas(DIAGRAMA DE CAJAS)

df_alicorp_num = df_alicorp.select_dtypes(exclude = 'O')

for i in df_alicorp_num.columns:

    sns.boxplot (x = df_alicorp_num[i], orient='v')

    print(i)

    plt.show()

### Analizando cuanto porcentaje de outliers hay

def ident_outliers(df, columns):

    lista_vars= []

    lista_outlier_sup= []
```

```
    lista_outlier_inf= []
    lista_pct= []
    lista_vs= []
    lista_vi= []
    for w in columns:
        q1, q3 = np.percentile(df[w],[25,75])
        iqr = q3 - q1
        vs = q3 + 1.5*iqr
        vi = q1 - 1.5*iqr
        outlier_sup = df[df[w]>vs]. shape [0]
        outlier_inf = df[df[w]<vi]. Shape [0]
        lista_vars.append(w)
        lista_outlier_sup.append(outlier_sup)
        lista_outlier_inf.append(outlier_inf)
        lista_vs.append(vs)
        lista_vi.append(vi)
        lista_pct.append(round(((outlier_sup+outlier_inf)/df.shape[0])*100,2))

df_outlier=pd.DataFrame({'Variable':lista_vars,'Outlier_sup':lista_outlier_sup,'Outlier_inf':
lista_outlier_inf,  'VS' :lista_vs, 'VI':lista_vi, 'Porcentaje':lista_pct})
df_outlier=df_outlier.sort_values(by=['Porcentaje'],ascending=False).reset_index(drop=Tr
ue)
 return df_outlier
columnsNumeric=list(df_alicorp.select_dtypes(["int64","float64"]).columns)
outlier=ident_outliers(df_alicorp,columnsNumeric)
outlier
#Imputamos los outliers por metodo TUKEY
def tukey_outliers(df, column):
    q1, q3 = np.percentile(df[column],[25,75])
    iqr = q3 - q1
    vs = q3 + 1.5*iqr
    vi = q1 - 1.5*iqr
```

```
    df.loc[df[column] > vs, column] = vs
    df.loc[df[column] < vi, column] = vi
    return df
for column in df_alicorp_num.columns:
    df_alicorp = tukey_outliers(df_alicorp, column)
```

1.4.Construcción de nuevas variables:

```
df_prom=df_alicorp.groupby('PROFUNDIDAD')['HORAS_LUZ'].mean()
df_prom=pd.DataFrame(df_prom)
df_prom.rename(columns={'HORAS_LUZ': 'PROM_HL'}, inplace=True)
df_prom
##uniendo las bases de factores productivos y ambientales
df_alicorp=pd.merge(df_alicorp, df_prom, on='PROFUNDIDAD', how='left')
df_alicorp['PROF_HOR']=0.000000
for i in df_alicorp.index:
    if df_alicorp['PROFUNDIDAD'][i]==df_alicorp['HORAS_LUZ'][i]:
        df_alicorp['PROF_HOR'][i]=df_alicorp['HORAS_LUZ'][i]
    else:
        df_alicorp['PROF_HOR'][i]=df_alicorp['PROM_HL'][i]
df_alicorp['INI_FIN']=df_alicorp['N_INICIAL']-df_alicorp['N_FINAL']
df_alicorp.drop(['N_INICIAL'],axis=1,inplace=True)
df_alicorp.drop(['N_FINAL'],axis=1,inplace=True)
df_alicorp.drop(['PROFUNDIDAD'],axis=1,inplace=True)
df_alicorp.drop(['HORAS_LUZ'],axis=1,inplace=True)
df_alicorp.drop(['PROM_HL'],axis=1,inplace=True)
```

1.5. Dividiendo en data de entrenamiento y prueba:

```
df_alicorp_dummy = pd.get_dummies(df_alicorp, drop_first = True)
df_alicorp_dummy['LOG_PESO_PROMEDIO_FINAL'] =
np.log(df_alicorp_dummy['PESO_PROMEDIO_FINAL'])
df_alicorp_dummy.drop(['PESO_PROMEDIO_FINAL'],axis=1,inplace=True)
```

```
#Separamos nuestros features del data frame para trabajarlos de manera independiente.

X = df_alicorp_dummy.drop(['LOG_PESO_PROMEDIO_FINAL'], axis = 1)

Y = df_alicorp_dummy['LOG_PESO_PROMEDIO_FINAL']

#Se utiliza el MinMaxScaler para estandarizar el dataset en caso algún algoritmo lo requiera

scaler = MinMaxScaler()

df_alicorp_scaled = X.copy()

for column in X.columns:

    df_alicorp_scaled[column] = scaler.fit_transform(X[[column]])

#La división la hacemos de manera aleatoria, repartiendo un 80% del data frame como data para entrenar.

X_train, X_test, Y_train, Y_test = train_test_split(df_alicorp_scaled, Y, test_size = 0.2, random_state = 42)
```

1.6.Modelamiento:

1.6.1.Modelo de Regresión Lineal:

```
#1. linear regression

from sklearn import linear_model

lr = linear_model.LinearRegression()

model = lr.fit(X_train, Y_train)

print ('RMSE=',(sum((np.exp(preds)-np.exp(Y_test))**2)/Y_test.count())**0.5)

print ('MAPE=',(sum(abs((np.exp(Y_test)-np.exp(preds))/np.exp(Y_test)))/Y_test.count())*100)

print ('MAE=',(sum(abs(np.exp(Y_test)-np.exp(preds)))/Y_test.count()))
```

1.6.2.Arbol de decisión:

```
# create a regressor object

regressor = DecisionTreeRegressor(random_state = 0)

# fit the regressor with X and Y data

regressor.fit(X_train, Y_train)

preds = regressor.predict(X_test)

print ('RMSE=',(sum((np.exp(preds)-np.exp(Y_test))**2)/Y_test.count())**0.5)

print ('MAPE=',(sum(abs((np.exp(Y_test)-np.exp(preds))/np.exp(Y_test)))/Y_test.count())*100)
```

```
print ('MAE=',(sum(abs(np.exp(Y_test)-np.exp(preds)))/Y_test.count()))
```

1.6.3.Random Forest:

 # Instantiate a RandomRegressor object

MAXDEPTH = 50

```
regr = RandomForestRegressor(n_estimators=1000,       # No of trees in forest
                        criterion = "mse",       # Can also be mae
                        max_features = "sqrt",  # no of features to consider for the best split
                        max_depth= MAXDEPTH,    #  maximum depth of the tree
                        min_samples_split= 2,   # minimum number of samples required to split an internal node
                        min_impurity_decrease=0, # Split node if impurity decreases greater than this value.
                        oob_score = True,        # whether to use out-of-bag samples to estimate error on unseen data.
                        n_jobs = -1,           #  No of jobs to run in parallel
                        random_state=0,
                        verbose = 10           # Controls verbosity of process
                        )
def Regression(regr,X_test,Y_test):
    start = time.time()
    regr.fit(X_train,Y_train)
    end = time.time()
    rf_model_time=(end-start)/60.0
    print("Time taken to model: ", rf_model_time , " minutes" )
regr.fit(X_train,Y_train)
preds=regr.predict(X_test)
print ('RMSE=',(sum((np.exp(preds)-np.exp(Y_test))**2)/Y_test.count())**0.5)
print ('MAPE=',(sum(abs((np.exp(Y_test)-np.exp(preds))/np.exp(Y_test)))/Y_test.count())*100)
print ('MAE=',(sum(abs(np.exp(Y_test)-np.exp(preds)))/Y_test.count()))
```

1.6.4.XGBoost:

```
dtrain = xgb.DMatrix(X_train, label=Y_train)
dvalid = xgb.DMatrix(X_test, label=Y_test)
watchlist = [(dtrain, 'train'), (dvalid, 'valid')]
def rmsle(y, y0):
    y0 = y0.get_label()
    assert len(y) == len(y0)
    return 'error', np.sqrt(np.mean(np.power(np.log1p(y)-np.log1p(y0), 2)))
xgb_pars = {'min_child_weight': 30, 'eta': 0.1, 'colsample_bytree': 0.3,
          'max_depth': 8,
'subsample': 0.9, 'lambda': 1., 'nthread': -1, 'boosting_type': 'gbdt', 'silent': 1,
'eval_metric': 'rmse', 'objective': 'reg:linear'}
model = xgb.train(xgb_pars, dtrain, 200, watchlist, early_stopping_rounds=50,feval=rmsle,
      maximize=False, verbose_eval=20)
print('Modeling RMSLE %.5f' % model.best_score)
columnsVars=list(X_test.columns)
dtest = xgb.DMatrix(X_test[columnsVars])
preds = model.predict(dtest)
print ('RMSE=',(sum((np.exp(preds)-np.exp(Y_test))**2)/Y_test.count())**0.5)
print ('MAPE=',(sum(abs((np.exp(Y_test)-np.exp(preds))/np.exp(Y_test)))/Y_test.count())*100)
print ('MAE=',(sum(abs(np.exp(Y_test)-np.exp(preds)))/Y_test.count()))
```