UNIVERSIDAD DE PANAMÁ

VICERRECTORÍA DE INVESTIGACIÓN Y POSTGRADO

FACULTDAD DE CIENCIAS NATURALES, EXACTAS Y TECNOLOGÍA

ESCUELA DE ESTADÍSTICA

INFORME DE PROYECTO DE INTERVENCIÓN

ESTIMACIÓN DEL PESO PROMEDIO DE COSECHAS DE *Rachycentron canadum* LINNAEUS (RACHYCENTRIDAE) EN LA REGIÓN ATLÁNTICA DE PANAMÁ.

SUSANA ISABEL KOO CHONG

PRESENTADO COMO UNO DE LOS REQUISITOS PARA OBTENER EL GRADO DE MAESTRO EN ESTADÍSTICA APLICADA

PANAMÁ, REPÚBLICA DE PANAMÁ

2018

DEDICATORIA

Este trabajo que presento hoy en día se lo dedico enteramente a mis padres quienes han sido mis formadores y los que me han impulsado a ser mejor cada día. Me han enseñado a que no hay barreras que no se puedan cruzar y que el único obstáculo es el que nosotros nos idealizamos para excusarnos. Mis pilares, la fuerza motora que me impulsa a ser mejor cada día y querer ser como ellos.

A mi abuela Ana, aquella persona que me tuvo entre sus brazos desde pequeña y ha vivido junto conmigo cada etapa de mi vida. Sus sabios consejos de la importancia del estudio y de seguir adelante son una de las cosas más valiosas que puede permanecer en mí.

A mis 4 bellos hermanos, los cuales admiro muchísimo y siempre han sido fuente de inspiración para mí. Me han hecho superarme académicamente y poder ir de la mano con ellos.

AGRADECIMIENTO

Quiero agradecer a Dios por bendecirme y permitir cumplir una meta más en mi vida.

Agradecida me siento con la empresa Open Blue, S.A., por darme la oportunidad de poder trabajar en conjunto con ellos. Al Sr. Rodger Miranda, el cual estuvo en las primeras etapas, sentando las bases y proporcionándome toda la información necesaria para el desarrollo del proyecto.

A la profesora Mitzi Cubilla quien me introdujo al Sr. Rodger y que me recomendó como estudiante para llevar a cabo esta temática que presento.

A mis tres asesores, la profesora Aurora Mejía, Gonzalo Carrasco y Estelina Ortega, por dedicar su tiempo en la revisión del documento, por su comprensión y apoyo en todo este proceso, especialmente en la etapa final.

A la profesora Clara Ruíz, que a pesar de que oficialmente no fue asignada como asesora, estuvo asesorándome en las primeras etapas de análisis de la data proporcionada por la empresa.

A la profesora María Bustamante, la cual siempre ha sido un apoyo para mi persona y estaré eternamente agradecida.

Al Dr. Medianero, el cual ha sido una gran inspiración para mí y en parte por el cual me inspiré a comenzar este viaje. Sus consejos siempre los tendré presentes.

A mis amigas y amigos, que siempre me han acompañado en las buenas y en las malas, su apoyo moral fue una de las grandes fortalezas para rendirme. Especialmente me gustaría mencionar a amigas Damaris Bernal, Angélica Castro, Maryori Wilson, mi mejor amiga Brigitte Henríquez y mis amigos Ricardo Márquez y Josué Young quien a pesar de que llegó a dormirse en alguna reunión a la cual me acompañó, estuvo siempre al pendiente e impulsando.

Mis dos últimos **especiales agradecimientos** están dirigidos hacia mi colega el Ing. Erick Gordon, el cual no solo es un gran amigo, sino que también fue mi tutor en este viaje que emprendí, cuando pensé que la nave no tenía reparación, él me extendió la mano para guiarme a estudiar otras vías que pudieran servirme para mantenerme en ruta. Sus enseñanzas, consejos, paciencia, dedicación fue lo que me ayudó a poder llegar a la meta. Me enseñó a que hay caminos alternos, que, aunque no las conozca tengo que empezar de cero y aprender de ellas. Su mentoría me orientó a sentar las bases de lo que tenía que ver de ahora en adelante y de cómo era la mejor manera para empezar y darle continuidad. Infinitamente agradecida.

Y mi familia, que es el mejor regalo que Diosito me pudo haber otorgado. Mis padres, siempre digo y siempre diré y no me cansaré de repetirlo que no sé qué haría sin todo el apoyo incondicional que me han brindado y que estaré eternamente agradecida. Su comprensión y tolerancia no tienen precio. A mi abuela Ana y mis hermanos, que forman parte de mi complemento y que siempre han estado para mi en las buenas y malas.

INDICE DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE GRÁFICOS

Resumen

Anualmente la tasa de crecimiento poblacional va en aumento, proporcionalmente la demanda alimenticia. Es por ello que se ha presentado como alternativa la piscicultura, como la rama de la producción de alimentos con un rápido desarrollo para satisfacer las necesidades a nivel mundial. Open Blue Sea Farms, es una empresa estadounidense radicada en Panamá desde el año 2010, dedicada a la maricultura en mar abierto, específicamente del pez *Rachycentron canadum* conocido vernacularmente como cobia. Sus actividades se concentran en la crianza y venta de la cobia, ofreciendo la mejor calidad posible al mercado. Sin embargo, la empresa no cuenta con una herramienta que les permita aumentar el nivel de confianza sobre el peso del producto al momento de la compra. Es decir, que requieren poder optimizar su producto para satisfacer las necesidades del cliente, ya que los mismos cuentan con una serie de especificaciones que requieren que sean cumplidas para poder continuar de manera eficiente la cadena de mercado. El presente proyecto tiene como objetivo proponer un modelo estadístico para la estimación del peso promedio de cosecha de Rachycentron canadum cultiva a mar abierto, utilizando los datos suministrados por la empresa Open Blue Sea Farms.

Abstract

Annually the rate of population growth is increasing, proportionally the food demand. That is why fish farming has been presented as an alternative, as the branch of food production with a rapid development to meet the needs of the world. Open Blue Sea Farms, is an American company based in Panama since 2010, dedicated to mariculture in the open sea, specifically the fish *Rachycentron* Canada known vernacularly as cobia. Its activities are focused on the raising and sale of cobia, offering the best possible quality to the market. However, the company does not have a tool that allows them to increase the level of confidence about the weight of the product at the time of purchase. That is, they need to be able to optimize their product to meet the needs of the client, since they have a series of specifications that require that they be met in order to continue efficiently the market chain. The objective of this project is to propose a statistical model for estimating the average harvest weight of *Rachycentron canadum* grown in the open sea, using the data provided by the company Open Blue Sea Farms.

Introducción

El análisis predictivo es un área de la minería de datos que consiste en la extracción de información existente en los datos y su uso para predecir patrones de comportamiento y tendencias, haciendo uso de datos históricos o actuales y pudiendo aplicarse sobre cualquier evento desconocido, ya sea en el pasado, presente o futuro (Nyce, 2007).

Se fundamenta en la identificación de las relaciones existentes entre las variables y eventos pasados para examinar las relaciones y predecir los posibles resultados en futuras situaciones. Cabe destacar que la precisión de los resultados va a depender de cómo se ha llevado a cabo el análisis de los datos, así como la calidad de las suposiciones (Espino y Martínez, 2017).

La analítica predictiva permite a las empresas, organizaciones, etc., ser proactivas, tener vista en el futuro, anticipando comportamientos y resultados basados en datos y no en especulaciones.

Implica el uso de técnicas de análisis estadístico y algoritmos de aprendizaje automático aplicados a un conjunto de datos para crear modelos predictivos que den como salida una puntuación o valor numérico en la probabilidad de que se dé un evento en particular (Maimon & Rokach, 2010).

Los modelos predictivos aplican resultados conocidos con el fin de entrenar al modelo para predecir valores, con datos diferentes o completamente nuevos, en un proceso repetitivo. El modelado proporciona resultados en forma de predicciones representadas mediante el grado de probabilidad de la variable en estudio, basado en la significación estimada a partir de un conjunto de variables de entrada. La variable en estudio puede ser cuantitativa o cualitativa (Pérez, 2007).

No existe limitación en el uso y aplicación de análisis predictivos, todo dependerá de lo que se quiera obtener. Es aplicable en cualquier campo, ya sea científico, de negocios, marketing, entre otros.

A través de este proyecto de intervención se busca poder identificar las variables asociadas al desarrollo del pez *Rachycentron canadum*. Para identificar las variables, se hizo uso de la técnica de minería de datos: Random forest, considerada como dos de las técnicas más precisas y eficientes que existen actualmente.

El presente documento se ha organizado en 5 capítulos, cuyo contenido se detalla a continuación:

El capítulo I se presenta el planteamiento del problema, justificación, objetivos a alcanzar e hipótesis, útiles para una mejor comprensión del contexto de la investigación.

El capítulo II contiene el marco de referencia del objeto en estudio, donde se presenta una revisión bibliográfica de los antecedentes, distribución y aspectos biológicos relevantes del pez *Rachycentron canadum*.

El capítulo III describe el marco de referencia de los algoritmos Árbol de decisión y Random Forest, el cual surge como mejora de los árboles simples. Se establecen las bases teóricas de estos algoritmos de minería de datos, los antecedentes. Se resaltará las ventajas, desventajas. los parámetros que utiliza y sus métodos de selección de variables importantes que utiliza el algoritmo Random Forest.

El capítulo IV abarca el marco metodológico, desde la captura de datos en campo, hasta el procesamiento de la base de datos siguiendo la metodología de CRISP-DM, identificación de variables importantes, generación y validación del modelo.

En el capítulo V se detallan los resultados obtenidos a partir de la identificación de las variables importantes utilizadas para generar el modelo final con el algoritmo Random Forest.

Por último, se detalla las conclusiones, recomendaciones y limitaciones, anexos y bibliografía de las citas utilizadas para la documentación y revisión del objeto en estudio y del algoritmo utilizado para generar el modelo.

CAPÍTULO I

1. Marco introductorio

1.1. Planteamiento del problema

Open Blue Sea Farms, es una empresa estadounidense radicada en Panamá desde el año 2010, dedicada a la maricultura en mar abierto, específicamente del pez *Rachycentron canadum* conocido vernacularmente como cobia. Sus actividades se concentran en la crianza y venta de la cobia, ofreciendo la mejor calidad posible al mercado. Sin embargo, la empresa no cuenta con una herramienta que les permita aumentar el nivel de confianza sobre el peso del producto al momento de la compra. Es decir, que requieren equipararlo de tal forma que cuando lo vendan u ofrezcan al mercado puedan asegurar el peso promedio de los peces.

Basado en lo establecido en el párrafo anterior, se establece la siguiente pregunta de investigación:

¿Cuáles son las variables no biológicas que influyen en el desarrollo del pez *Rachycentron canadum*?

1.2. Justificación

Como empresa, Open Blue Sea Farms, tienen la necesidad de poder optimizar su producto y así poder satisfacer al cliente, ya que los mismos cuentan con una serie de especificaciones que requieren que sean cumplidas para poder continuar de manera eficiente la cadena de mercado. El producto final va desde productos culinarios hasta procesos de industrialización de enlatado, de esta manera si la empresa mantiene un producto uniforme u homogéneo, le permitirá a su gama de clientes poder determinar el costo de su producto acorde a su tamaño y peso.

1.3. Hipótesis

"Es posible generar un modelo que permita predecir el peso promedio de las cosechas de *Rachycentron canadum* a través de las variables no biológicas presentes en este estudio".

1.4. Objetivos

1.4.1. Objetivo General

- Proponer un modelo estadístico para la estimación del peso promedio de cosecha de *Rachycentron canadum*, utilizando datos suministrados por la empresa Open Blue Sea Farms.

1.4.2. Objetivos Específicos

- Determinar las variables significativas que intervienen para obtener un modelo de predicción.
- Estimar el peso promedio de cosecha de *Rachycentron canadum*.
- Demostrar mediante la validación de la técnica Random Forest, que ésta se ajusta apropiadamente a los requerimientos del problema planteado.

CAPÍTULO II

2. Marco de Referencia: Pez *Rachycentrum canadum*

2.1. Antecedentes

Los primeros registros de investigación acuícola de cobia datan de 1975 con la recolección de huevos de cobia silvestre en las costas de Carolina del Norte. Se describe el desarrollo larvario y tras la culminación de pruebas de cría de 131 días, se concluyó que la cobia mostraba un buen potencial acuícola debido a su rápido crecimiento y buena calidad de la carne. Durante la última parte de la década de 1980 y principios de los 90, tanto en Estados Unidos como en la Provincia China de Taiwán, se continuó la investigación relativa a la cobia.

El primer desove en cautiverio de esta especie se produjo en la Provincia China de Taiwán a principios de los años 90. La investigación continuó su curso y hacia 1997 ya se había desarrollado una base tecnológica para la producción de crías de cobia en grandes cantidades, aportando peces juveniles para su engorda, principalmente en sistemas de jaulas cerca de la costa para evitar pérdidas en la producción e infraestructura por tifones, sin embargo, ha traído como consecuencia un descenso en la calidad del producto, además del incremento de enfermedades por contaminación de aguas (Mendoza et al., 2012).

Por otro lado, la implementación de jaulas sumergidas de tipo octogonal y esférico a representando un importante avance en países como Taiwán, Puerto Rico, Estados Unidos, al estar protegidas contra depredadores, además de reducir el efecto ecológico que implicaría si un escape masivo de esta especie se mezclara con las especies nativas, teniendo en consideración que estamos hablando de individuos con hábitos carnívoros (Mendoza et al., 2012).

Otro de los avances que se han obtenido en cuanto al cultivo de cobia es el sistema de alimentación mediante alimento peletizado, monitoreando su funcionamiento mediante un circuito cerrado de televisión para evitar la pérdida de alimento en las jaulas (Mendoza et al., 2012).

2.2. Taxonomía

Pertenecen al Phylum Chordata, Subphylum Vertebrata, Supraclase Gnathostomata, Clase Actinopterygii, Superorden Acanthopterygii, Orden Perciformes, Suborden Percoidei, Familia Rachycentridae,

La familia Rachycentridae cuenta con una sola especie, *Rachycentron candum*. Descrita originalmente por Linnaeus (1766) como *Gasterosteus canadus* y posteriormente cambiado a *Rachycentron candum* (Linnaeus, 1766). Entre los sinónimos de la cobia se puede mencionar:

Tabla 1. Nombres científicos que ha recibido el pez "Cobia" a lo largo del tiempo.

SINÓNIMO	AUTOR
Apolectus niger	Bloch 1793
Scomber niger	Bloch 1793
Naucrates niger	Bloch 1793
Elacate nigra	Bloch 1793
Centronotus gardenii	Lacepede 1801
Centronotus spinosus	Mitchill 1815
Rachycentron typus	Kaup 1826
Elacate Motta	Cuvier y Valenciennes 1829
Elacate atlantica	Cuvier y Valenciennes 1832
Elacate bivittata	Cuvier y Valenciennes 1832
Elacate malabarica	Cuvier y Valenciennes 1832
Elacate pondiceriana	Cuvier y Valenciennes 1832
Meloderma nigerrima	Swainson 1839
Naucrates niger	Swainson 1839
Elacate falcipinnis	Gosse 1851
Thynnus canadensis	Gronow 1854
Elacate nigra	Gunther 1860
Rachycentron canadus	Jordan y Evermann 1896
Rachycentron pondicerrianum	Jordan 1905

2.3. Características morfológicas distintivas de *Rachycentron canadum*

La siguiente descripción de ***Rachycentron canadum*** es de Collete (1978): cuerpo elongado, subcilíndrico o fusiforme; cabeza ancha y deprimida. **Boca** grande, terminal, con la proyección de la mandíbula inferior; **banda de dientes** viliformes en las mandíbulas y en el cielo de la boca y lengua. **Primera aleta dorsal** con 7 a 9 (normalmente 8) espinas cortas y aisladas, no conectadas por una membrana; **segunda aleta dorsal** larga, radios anteriores algo elevada en los adultos; **aletas pectorales** puntiagudas, tornándose más curvas con la edad; **aleta anal** similar a la aleta dorsal, pero más corta; **aleta caudal** semilunar en adultos, lóbulo superior más largo que corto (aleta caudal redondeada en los jóvenes, los rayos centrales mucho más prolongados). **Escamas pequeñas**, incrustado en la piel gruesa; **línea lateral** ligeramente ondulado en sentido anterior. Color: dorso y costados de color marrón oscuro, con 2 bandas de luz estrechos bien definidos; vientre amarillento.

2.4. Distribución mundial

Posee una amplia distribución en todos las aguas tropicales y templados del mundo, con excepción de la zona este del océano Pacifico. En el océano Atlántico occidental, se distribuye desde Nueva Escocia en Canadá hasta Argentina; en la zona oriental del Atlántico se distribuye desde Marruecos hasta Sudáfrica (Monod, 1973, Smith, 1965); en el Pacifico oeste desde Japón a Australia. En la actualidad se encuentran cultivos de cobia en países como: Australia, China, Estados Unidos, Taiwán, Vietnam y en Latinoamerica (Kaiser y Holt, 2004).

2.5. Desarrollo reproductivo

Las hembras alcanzan la madurez sexual a los 2-3 años de edad, mientras que los machos lo alcanzan antes, al año o a los 2 años de edad. Sin embargo, la tasa de crecimiento es mayor en las hembras, alcanzando hasta 60 kg. Presentan un tipo de reproducción sexual gonocórica (de gonos, sexo y choris, separación) o bisexual, es decir que se reproducen solo como hembras o solo como machos. Poseen un único sistema reproductor, pudiendo tener ovarios o testículos. Este tipo de reproducción se presenta regularmente en especies de ambiente natural y de cultivo.

Durante el apareamiento se da la fertilización externa, que tiene lugar tanto en la línea costera como fuera de ella, en donde la hembra libera desde varios cientos de miles de huevos no fertilizados hasta varios millones de ellos y simultáneamente el macho expulsa a su vez espermas para fertilizar los huevos.

Los huevos no fertilizados presentan tres fases (Richard, 1967): 1) *inmaduro claro*, que consiste en células nucleadas de 0.10-0.30 mm de diámetro, 2) *huevos de maduración*, de 0.36-0.66 mm de diámetro, presentan un aspecto opaco y el glóbulo oleoso apenas distinguible, 3) *huevos maduros claros o transparentes*, presentan un tamaño promedio de 1.20 mm de diámetro y su glóbulo oleoso ≈0.37 mm diámetro.

Los huevos fertilizados son pelágicos, su yema esta segmentada y se distinguen por su largo glóbulo oleoso, el cual junto con el embrión presentan una coloración amarillenta y moteada con pigmento de melanina (Hassler y Rainville, 1975). Durante la fase embrionaria se da un rápido crecimiento del blastodermo –epibolia–. Justo antes del cierre del blastoporo, se da la formación de los lóbulos ópticos, vesícula de Kupffer y somitas.

Las larvas de cobia crecen rápidamente, se liberan aproximadamente 24-36 horas después de la fecundación, alcanzando 3,5 mm longitud total al eclosionar y carentes de pigmentación. A los cinco días de la eclosión, es visible una raya de color amarillo a lo largo del cuerpo y se da inicio al desarrollo de los ojos y boca, permitiéndoles poder alimentarse activamente. Al día 30 el juvenil adquiere la apariencia de un adulto de cobia con dos bandas de color que va desde la cabeza hasta el extremo posterior (Holt, Faulk, y Schiwarz, 2007).

2.6. Bioecología

Los huevos y larvas se encuentran en alta mar y los primeros juveniles reposan cerca de la zona costera, muy cercanos a las bahías, desembocaduras de ríos, islas de barrera, cercanos a las playas (Benson, 1982; Hoese y Moore, 1977; McClane, 1974; Swingle, 1971).

Los adultos de esta especie son principalmente solitarios, raramente se encuentran en grupos de 3-4 especímenes. Se encuentran en bahías, estuarios, pantanos de mangle y muy asociados a los arrecifes a profundidades que rondan en un rango de 0 a 1200 metros (Freeman y Walford, 1976; Relyea, 1981; Goodson, 1985; Vaught, Shaffer y Nakamura, 1989).

Es una especie carnívora, oportunista, se alimenta de diversos peces (arenques, corvinas, lisas, pargos), crustáceos (calamar, camarón y cangrejo) y cefalópodos (Darracott, 1977). Su periodo de vida se estima que es de 15 años aproximadamente, llegando a alcanzar una longitud de 2 metros con un peso máximo de 68 kg (Robins y Ray, 1986).

2.7. Factores fisicoquímicos

Existen muchos factores por los cuales se puede ver limitada la distribución y desarrollo de la cobia en sus diferentes etapas. Además de que son de vital importancia para tomar decisiones adecuadas para su cultivo. En la **tabla 2** se resume los factores físicos-químicos y sus rangos de permisibilidad.

Tabla 2. Factores físicos-químicos que influyen en el desarrollo *de Rachycentron canadum*.

PARÁMETRO	MÍNIMO	MÁXIMO
Temperatura (°C)	16	32
Oxígeno disuelto (mg/l)	6	8
pH	7.6	7.8
Salinidad (ups)	22	44
Amonio (mg/l)	<0.3	

CAPÍTULO III

3. Marco de Referencia: Minería de datos

3.1. Antecedentes

El diagrama de árbol de decisión tuvo sus inicios en 1944 cuando se dio lugar a la Teoría de Juegos diseñada por el matemático John von Neumann y el economista Oskar Morgenstern, quienes utilizaron este tipo de gráfico como herramienta para representar la estructura del análisis del juego de forma extensiva (desde un principio hasta el final) que implican secuencias de movimientos (un movimiento es un binomio decisión-acción) (Mangani, 2015).

Para el año 1971 Sonquist, Baker y Morgan dieron lugar al programa para la Detección de Interacciones Automáticas (AID, siglas en inglés, Automatic Interaction Detection), representando uno de los primeros métodos de ajuste de datos basados en modelos de árboles de clasificación (López, 2015). El mismo surgió para enfrentar los desafíos que presenta la investigación estadística en manejar enormes cantidades de datos que incluye un gran número de variables, requeridas para encontrar patrones, definir tendencias y así poder adquirir información.

Kass introdujo en 1980 un algoritmo recursivo de clasificación no binaria llamada CHAID (Chi-square Automatic Interaction Detection o Detección de Interacciones Automáticas Chi-cuadrado) (Serna, 2009). Hacia el año 1984, se incorporó un nuevo algoritmo para la elaboración de árboles aplicados a problemas de regresión y clasificación, dado por los investigadores R. Olshen, L. Breiman, C. Stone y J. Friedman (Ordoñez, 2010).

Otros métodos más recientes son: FIRM (Formal Inference-based Recursive Modeling) propuesto por Hawkins (Hadidi, 2003); y MARS (Multivariate Adaptive Regression Splines), propuesto por Friedman en el año 1991.

3.2. Árbol de decisión

Un árbol de decisión es un método no-paramétrico de segmentación binaria (es decir, que cuentan con dos opciones, pudiendo existir tres o más) utilizado por diversas áreas desde la inteligencia artificial hasta la Economía, para clasificar las observaciones o realizar predicciones (Rokach y Maimon, 2008).

Recibe este nombre debido a la apariencia que constituye el gráfico, semejándose a un árbol cuyas ramificaciones constituyen un conjunto de decisiones. Estas decisiones se construyen de forma lógica basados en reglas de decisión y sirven para representar y categorizar un conjunto de condiciones que ocurren de forma sucesiva, para la resolución de la interrogante que se pretende solucionar.

Un árbol de decisión posee entradas que pueden ser una situación u objeto definido por un conjunto de atributos, a partir de los cuales se genera una respuesta que resulta ser una decisión tomada a partir de las entradas. Los valores que pueden tomar tanto las entradas como las salidas pueden ser valores discretos o continuos.

El objetivo de este modelo de predicción es la obtención de grupos más homogéneos con respecto a la variable que se pretende discriminar dentro de un subgrupo de individuos y heterogéneos entre los subgrupos.

Entre las ventajas que presenta este método podemos mencionar (Orallo, Ramírez, Ramírez, 2004):

- Son muy fáciles de entender e interpretar.
- Es robusta con respecto a datos atípicos. Además, que se desempeña bien incluso si sus suposiciones son violadas por el verdadero modelo a partir del cual se generaron los datos.
- Reduce el número de variables independientes.
- Pueden ser analizados grandes cantidades de datos.
- Toma en consideración las interacciones que se puedan dar entre los datos.
- Los criterios de construcción del método, árbol y algoritmo es el mismo tanto para los árboles de regresión como los árboles de clasificación.

- Su uso es permisible para cualquier tipo de variables explicativas: nominales, continuas, ordinales.

Y dentro de las desventajas que podemos encontrar, podemos mencionar:

- Dificultad para elegir el árbol *óptimo.*
- Para la construcción de árboles es preferible grandes cantidades de datos para garantizar que las observaciones es significativa.
- Las reglas de asignación son sensibles a leves cambios en los datos.
- Ausencia de una función global de las variables (como pueden ser una ecuación de regresión, función lineal discriminante, etc.), lo que trae como resultado la pérdida de la representación geométrica.

3.2.1. Estructura del árbol de decisión

El árbol de decisión se construye siguiendo un orden cronológico de los acontecimientos y está compuesto por nodos, ramas, líneas y etiquetas los cuales se detallan a continuación (Carmona, 1997):

- ***Nodo de raíz:*** es el nodo principal del cual parten líneas hacia los nodos inferiores, que a su vez pueden hacer de nodo raíz.
- ***Nodo de decisión***: indica el momento en que se debe tomar la decisión. Está representado por un cuadrado (□).
- ***Nodo de probabilidad***: representa un suceso incierto. Está representado por un círculo (o).
- ***Nodo terminal***: es el nodo donde termina el flujo, es decir que no se ramifica y que ya no son raíz de ningún otro nodo. Señala el resultado obtenido a través de la serie de sucesos y decisiones que llegan hasta él. Está representado por un triángulo (Δ).
- ***Rama***: son los arcos conectores por medio de los cuales se conectan los nodos y representan las posibles respuestas que los atributos pueden tomar.

3.2.2. Clasificación del árbol de decisión

Dependiendo del tipo de variable a discriminar se pueden distinguir dos tipos de árboles de decisión (Russel y Norvig, 2004):

a) **Árboles de regresión:** se emplea a variables continuas, en otros términos, el valor predicho es considerado un número real.
b) **Árboles de clasificación:** se aplica cuando se hace uso de variables categóricas, ya sea ordinales o nominales. El resultado será entonces la clase a la cual pertenecen los datos.

Estos dos tipos de árboles de decisión presentan similitudes, pero se distinguen por el procedimiento que utilizan para determinar en qué punto se realiza la división de los datos. A continuación, se menciona algunos de los métodos que construyen más de un árbol de decisión (Breiman et al., 1984):

- **Árboles impulsados:** son utilizados tanto en problemas de clasificación como en problemas de regresión (Hastie, Tibshirani, Friedman, 2001).
- **Bagging:** construye múltiples árboles de decisión efectuando continuamente un remuestreo de los datos de entrenamiento con recambio, y votando los árboles para encontrar la mejor predicción (Breiman, 1996).
- **Rotation Forest:** los árboles de decisión reciben un previo entrenamiento con un análisis de componentes principales (ACP) en un subconjunto aleatorio de las características de entrada (Rodríguez, Kuncheva, Alonso, 2006).
- **Random Forest o Bosques aleatorizados:** con el fin de mejorar la tasa de clasificación, hace uso de un grupo de árboles de decisión.

3.2.3. Elaboración del árbol de decisión

Los árboles de decisión se crean de la heurística denominada "particionamiento recursivo". Lo ideal disponer de datos de entrenamiento y cuantos más datos haya disponibles para aprender y más ricos y completos sea el algoritmo, funcionará mejor, ya que el algoritmo aprende de los datos. Una vez obtenido los datos de entrenamiento, se les serán proporcionados al algoritmo para entrenarlo y que aprenda de relaciones, patrones. Esto lo logra a través de la siguiente ejecución:

A partir del nodo raíz se divide el set de datos en subgrupos homogéneos tomando en consideración atributos de las variables de entrada (**Gráfico 1**)*, las cuales están representadas por los nodos internos y son utilizadas para predecir la variable de destino. De esta manera se va formando el primer conjunto de ramas del árbol, representando los posibles valores que puede adquirir el atributo y continúa subdividiéndose en cada subgrupo hasta alcanzar un tamaño prefijado formando al final las hojas o los nodos de decisión. Esto último se logra cuando un subconjunto en un nodo posee el mismo valor de la variable objetivo o cuando la partición no adiciona valor a las predicciones* (Zuniga y Abgar, 2011).

Una vez obtenido el modelo entrenado, se puede realizar predicciones a partir de las características de los nuevos datos que le serán suministrados, del cual no tiene conocimiento. El modelo será capaz de darnos una predicción en base al conocimiento que extrajo de los datos de entrenamiento.

En el siguiente ejemplo se puede observar que el modelo para predecir el éxito futuro de las películas se puede representar en un árbol simple como se muestra en el **Gráfico 2**. Para evaluar una secuencia de comandos, siga las ramas a través de cada decisión hasta que se haya predicho su éxito o falla. En poco tiempo, podrá clasificar la acumulación de scripts y volver a trabajos más importantes, como escribir un discurso de aceptación de premios (Lantz, 2013).

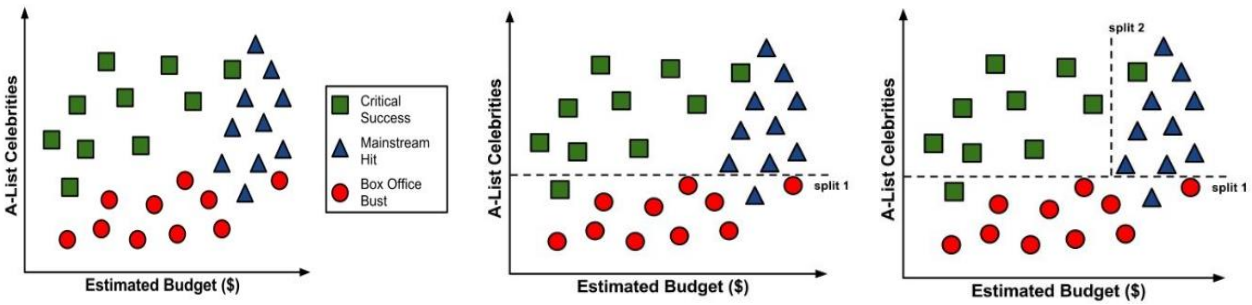

Gráfico 1. División de subgrupos homogéneos por atributos (Lantz, 2013).

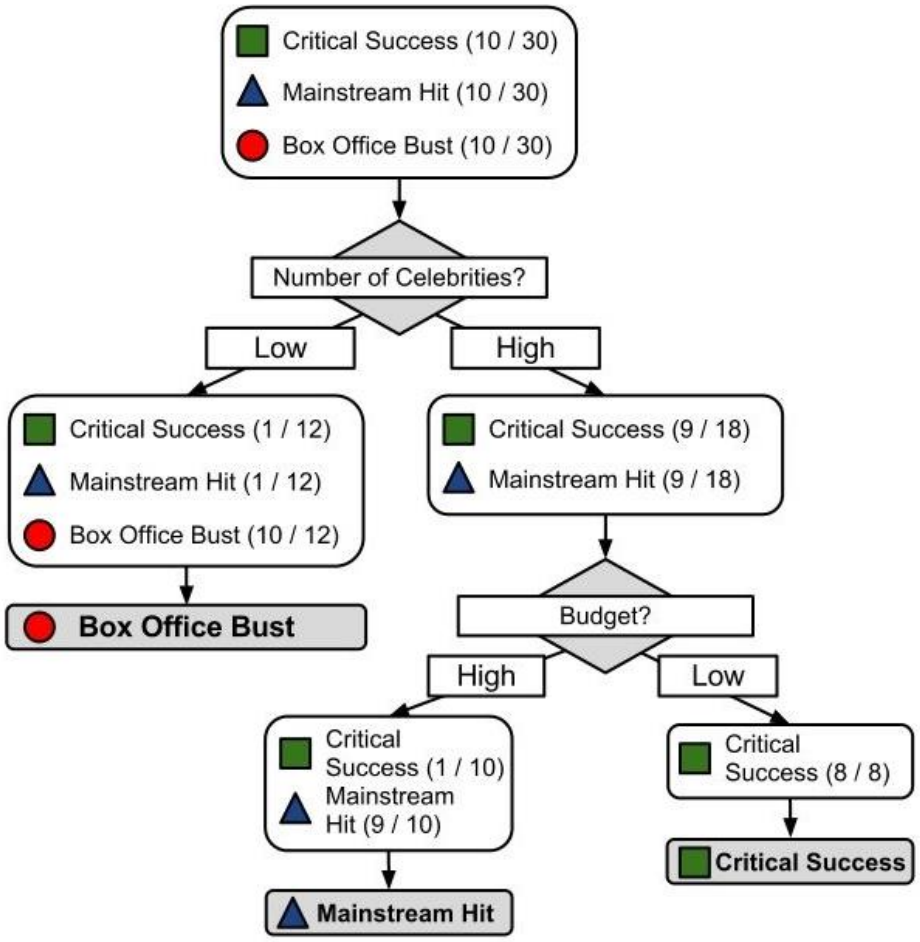

Gráfico 2. Representación gráfica de un árbol de decisión (Lantz, 2013).

En este proceso de división se pueden presentar ciertas situaciones:

a) Si existen objetos de distintas categorías, para separarlos se seleccionaría el que más información tenga.

b) Si los objetos restantes pertenecen a una misma categoría, entonces se clasifican en una sola categoría.

c) Si no hay más atributos, pero aún existen objetos con sus respectivas categorías, entonces nos encontramos con un problema. Esto es indicativo que la descripción de los objetos hasta ese punto es la misma, pero su clasificación es diferente, es decir hay ambigüedad en los datos. También se presenta el caso cuando los atributos no proporcionan la información suficiente para describir la situación.

Existe una serie de algoritmos utilizados en el análisis de árbol de decisión. Los más destacados son (Tarazona, 2016):

- **ID3** (Iterative Dichotomiser 3) (Ramírez, 2013)
- **C4.5** (Sucesor de ID3) (Ramírez, 2013)
- **C5.0** (versión optimizada y completa de la C4.5) ()
- **ACR** (Árboles de Clasificación y Regresión)
- **CHAID (Detector automático de Chi-cuadrado de interacción):** efectúa divisiones de múltiples niveles al calcular árboles de clasificación.
- **MARS:** extiende los árboles de decisión para manejar mejor los datos numéricos.
- **Árboles de Inferencia** Condicional: utiliza pruebas no paramétricas como criterios de división, corregidos para múltiples pruebas para evitar un sobreajuste. Se traduce en la elección de un predictor imparcial y no requiere poda.

3.2.4. Métricas como criterios de selección

Como se pudo observar en el apartado anterior un árbol de decisión se construye de arriba hacia abajo, de forma continua sin retroceder y las decisiones de división se basan en las características o atributos proporcionados en primera instancia. Uno de los retos que se enfrenta al momento de construir un árbol de decisión es determinar ¿cuál es la variable que mejor divide el conjunto de datos?

Para ello cada algoritmo se apoya en métodos matemáticos que determinan y miden la pureza de la variable de destino dentro de los subconjuntos. Estas métricas son aplicadas a cada subconjunto y los valores que se generan se combinan para proporcionar una medida de la calidad de la división (Rokach y Maimon, 2005).

3.2.4.1. Entropía

Utilizado por los algoritmos de generación de árboles ID3, C4.5 y C5.0. La entropía es una medida de la cantidad de incertidumbre o desorden que se puede encontrar en el conjunto de datos y es usado para ayudar a decidir que atributo es el más útil para ser seleccionado en cada paso. En términos generales, la entropía será menor entre mayor sea el número de objetos que discrimine un atributo.

El valor mínimo que puede adquirir la entropía es cero (0) e indica que la muestra es completamente homogénea, es decir, todos se clasifican igual), y si el valor que adquiere es uno (1), la muestra es igualmente distribuida, es decir, que tiene el mismo número de ejemplos de cada posible clasificación e indica la máxima cantidad de desorden. La función matemática que describe el grado de incertidumbre o entropía es la siguiente:

$$\boldsymbol{Entropía}(\boldsymbol{S}) = \sum_{i=1}^{C} -p_i log_2 p_i$$

Donde:

S: colección de objetos

C: número de diferentes clasificaciones

P*i*: probabilidad de los posibles valores

i: las posibles respuestas de los objetos

Por otro lado, también se introduce el concepto "Ganancia de Información", el cual representa una medida de discriminación, un indicador del siguiente atributo a ser seleccionado para que se de continuación al proceso de segmentación, discriminando el atributo previamente seleccionado entre los atributos que aún no han sido clasificados. La Ganancia de información se calcula mediante la siguiente fórmula (Russel y Norvig, 2004):

Gan Inf(S, A) = Entropia(S) - $\sum_{v \epsilon V(A)} \frac{|Sv|}{|S|} Entropia(Sv)$

Donde:

S: colección de objetos

A: son los atributos de los objetos

V(A): conjunto de valores que A puede tomar

3.2.4.2. Impureza de Gini

Es utilizado por el algoritmo de ACR (Árboles de Clasificación y Regresión). Mide la cantidad de veces que un elemento que fue elegido al azar del conjunto de datos sería etiquetado incorrectamente si ya fue etiquetado aleatoriamente de acuerdo a la distribución de las etiquetas en el subconjunto. Su valor mínimo es cero (0) y es indicativo de que todos los casos del nodo corresponden a una sola categoría.

Para el cálculo de la Impureza de Gini, supongamos que i toma valores en {1, 2, 3, …….,n}, y sea f_i la fracción de artículos etiquetados con valor i en el conjunto (Rokach, and Maimon, 2014):

$$\boldsymbol{I_g(f)} = -\sum_{i=1}^{m} f_i\,(1-f_i) = \sum_{i=1}^{m} f_i\,(f_i - f_i^2) = \sum_{i=1}^{m} f_i - \sum_{i=1}^{m} f_i^2 = 1 - \sum_{i=1}^{m} f_i^2$$

3.2.4.3. Reducción de la varianza

Es utilizado por el algoritmo Árboles de Regresión. Mide la reducción en la desviación estándar del valor original a la desviación estándar ponderada después de la división. Está representada por la siguiente fórmula (Quinlan, 1986):

$$\boldsymbol{SDR} = sd\ (T) - \sum_{I} \frac{|T_i|}{|T|} x\ sd\ (T_i)$$

Donde:

sd (T): es la desviación estándar de los valores en el conjunto T.

T1, T2…,Tn: son los conjuntos de valores resultantes de una división en una característica.

|T|: número de observaciones en el conjunto T.

3.3. Random Forest

Conocido en castellano como "bosques aleatorizados", es un método de aprendizaje tanto para regresión como para clasificación, combina la idea de Bagging de Breiman y la selección aleatoria de atributos, introducida independientemente por Ho, Amit y Geman, para construir una colección de árboles de decisión con variación controlada (Amit y Geman, 1997; Ho, 1998). El término aparece de la primera propuesta de **Random decision forests**, hecha por Tin Kam Ho de Bell Labs en 1995.

Consisten en la confección de múltiples árboles de decisión simples, que se utilizan en conjunto para determinar el resultado final. Se les aplica la técnica de Bootstrap Aggregating o Bagging con la finalidad de disminuir el sesgo de las predicciones y la varianza y prevenir el sobreajuste en los casos cuando se trabaja con grandes cantidades de datos (Breiman, 2001).

Cada árbol depende de valores de un vector aleatorio muestreado de forma independiente (con reemplazo) y con la misma distribución para todos los árboles del bosque, que es un subgrupo de los valores de predicción del conjunto de datos original. El tamaño óptimo del subgrupo que conforma las variables predictoras viene dado por $\boldsymbol{log2\ M\ +\ 1}$, donde **M** es el número de entradas.

Para problemas de clasificación, el conjunto de árboles simples opta por la clase más popular. A partir de un conjunto de variables aleatorias de predicción y un conjunto de árboles simples, los bosques aleatorizados definen una función de margen que mide el grado en que el número promedio de votos para la clase elegida supera el voto promedio para cualquier otra clase presente en la variable dependiente.

En los problemas de regresión, las respuestas se promedian para calcular la estimación de la variable dependiente. Están formados por el crecimiento de árboles simples, cado uno es capaz de producir un valor de respuesta numérica. Al igual que en los problemas de clasificación, el conjunto de predictores se selecciona al azar de la misma distribución y para todos los árboles.

Los bosques aleatorizados corrigen el sobreajuste que presentan los árboles de decisión a su conjunto de datos utilizados en entrenamiento. El error cuadrático medio de los bosques aleatorizados está dado por:

$$\boldsymbol{Error\ medio} = (observado - respuesta\ del\ árbol)$$

De forma resumida la confección del algoritmo del random forest sigue el siguiente proceso (Martínez, 2015):

1. Selecciona individuos al azar (usando muestreo con reemplazo) para crear diferentes sets de datos.
2. Crea un árbol de decisión con cada set de datos, obteniendo diferentes árboles, ya que cada set contiene diferentes individuos y diferentes variables.
3. Al crear los árboles se eligen variables al azar en cada nodo del árbol, dejando crecer el árbol hasta un determinado punto. Cabe destacar que de todas las variables originales se selecciona p variables y de estas selecciona la mejor para realizar la partición.
4. Una vez se construya todos los árboles del bosque se realiza la media de todos ellos y esta representará la predicción final del Random Forest (Hastie, Tibshirani, Friedman, 2008).

En random forest no existe una ecuación, esto se debe a que la naturaleza del algoritmo de bosque aleatorio conduce intrínsecamente a la destrucción de cualquier representación matemática simple, ya que el mismo funciona construyendo árboles de decisión y luego agregándolos y, por lo tanto, los valores Beta no tienen contrapartida en el bosque aleatorio. Aunque se obtienen los valores de "Importancia variable / Índice de Gini" para el bosque, que pueden usarse para dar sentido al modelo pero no como un factor de multiplicación.

3.3.1. Ventajas del Random Forest (Sullivan, 2017):

- Puede manejar una base de datos grandes, además de cientos de variables de entrada sin exclusión de alguna e identifica las más significativas entre ellas.
- Algoritmo de bosque aleatorio se puede utilizar en ambos tipos de problemas. Se puede usar tanto en regresión como en clasificación.
- Estima con eficacia los datos faltantes y mantiene la precisión cuando se presenta un gran porcentaje de perdida de datos.
- Presenta estimaciones de cuáles son las variables importantes para la clasificación.
- Muestrea datos de entrada con reemplazo. Proceso conocido como "muestreo bootstrap".
- Presenta varios métodos que se pueden usar para equilibrar errores en el conjunto de datos.
- Las características anteriores también se pueden usar con datos no etiquetados. Por lo tanto, puede
- trabaja sin supervisión.
- Computa los prototipos que dan información sobre la relación entre la clasificación y las variables.
- Computa las proximidades entre los pares de casos que pueden usarse en los grupos, localizando valores atípicos, o (ascendiendo) dando vistas interesantes de los datos.

3.3.2. Desventajas del Random Forest:

- Se puede presentar sobreajuste si los datos de muestra son demasiado ruidosos (Sullivan, 2017).
- Puede actuar como un enfoque de caja negra para modeladores estadísticos ya que no puede controlar el rendimiento del modelo. Solo puedes probar semillas aleatorias y diferentes parámetros (Sullivan, 2017).
- Para los datos que incluyen variables categóricas con diferente número de niveles, el random forests se parcializa a favor de esos atributos con más niveles. Por consiguiente, la posición que marca la variable no es fiable para este tipo de datos. Métodos como las permutaciones parciales se han usado para resolver el problema (Altmann et al., 2010).
- Si los datos contienen grupos de atributos correlacionados con similar relevancia para el rendimiento, entonces los grupos más pequeños están favorecidos sobre los grupos más grandes (Tolosi y Thomas, 2011).

3.3.3. Parámetros del Random Forest

El algoritmo Random Forest utiliza los parámetros: número de variables predictoras a utilizar en cada partición de cada uno de los árboles (mtry) y el número de árboles (ntree) (Hastie, Tibshirani, Friedman, 2008).

- *Parámetro ntree*: representa el número de árboles que se generan en cada Random Forest.
- *Parámetro mtry*: determina el número de variables que cada árbol analiza al momento de realizar el "Split" en cada nodo.

3.3.4. Importancia de la variable

La importancia de una variable está condicionada a su interacción, posiblemente compleja, con otras variables. El Random Forest calcula dos medidas de importancia distintas.

a) Mean Decrease Accuracy (MDA)

Con la finalidad de entrenar cada árbol que se agrega en el bosque, se utiliza el Bootstrap. En dicho proceso se deja aproximadamente un tercio de los casos de la muestra; a los casos que no son considerados para entrenar el árbol se les llama out-of-bag (OOB). Con ellos se puede estimar un error insesgado de clasificación y también se pueden utilizar para hacer una estimación de la importancia de las variables (Medina y Ñique, 2017).

El funcionamiento de esta lógica, es la siguiente: primero se escoge el error de clasificación out-of-bag, después se toma una variable al azar y se permutan sus valores dentro de los datos de entrenamiento, ocasionando que dicha variable escogida descorrelacione lo aprendido por el modelo. Luego se vuelve a calcular el error OOB, para luego compararlo con el error calculado inicialmente. En consecuencia, por lógica, si el error cambia, se afirma que dicha variable es importante. Este proceso se repite con todas las variables y luego estas se ordenan de acuerdo a los cambios que produjeron cada una en los errores OOB (Medina y Ñique, 2017).

b) Mean Decrease Gini (MDG)

Otra manera de estimar la importancia de las variables en el modelo Random Forest es utilizando el índice de Gini. Consiste en seleccionar la variable en cada partición en la construcción de los árboles y que corresponde a una disminución de esta medida. La importancia de una variable en un árbol se mide como la suma de los decrementos atribuidos a esa variable y la importancia final como la media en todos los árboles (González, 2015).

CAPÍTULO IV

4. Marco Metodológico

El presente estudio es abordado desde el paradigma cuantitativo, caracterizado en un conocimiento medible, observable y cuantificable; se sostiene en la base de predecir los fenómenos y como estos se dan en la naturaleza (Usher y Bryant, 1992). Está fundamentado en la filosofía positivista que afirma que el conocimiento auténtico es el conocimiento científico y que dicho conocimiento surge de la afirmación de las hipótesis a través del método científico (Meza, 2015).

4.1. Tipo de Investigación

Según el *estado del conocimiento, la profundidad de los objetivos y el alcance de los resultados* se trata de un estudio explicativo de tipo observacional (en su primera etapa). Es explicativo ya que el fenómeno de estudio busca explicar las variables que mayormente influyan en el crecimiento y desarrollo de los peces y observacional porque el factor de estudio no es controlado por el investigador.

Según las *posibilidades de aplicación de los resultados* es de tipo aplicada, cuyo objetivo es la solución de problemas prácticos en el contexto (Salkind, 1998).

Según el *lugar de desarrollo de la investigación* se trata de una investigación de campo, ya que se recopilan los datos de los peces en una situación natural donde se encontraban en jaulas flotantes a mar abierto en la Costa Atlántica de Colón, sin controles propios.

Según su temporalidad es un *estudio longitudinal*, ya que los datos primarios (variables) se colectan en diferentes tiempos: día, mes, años (Hernández et al, 2010).

Según la cronología de los hechos a investigar el estudio es de *tipo retrospectivo*, ya que los datos se recogieron desde el 2014 al 2016.

4.2. Tipo de Diseño

Es un diseño observacional, analítico, no experimental, longitudinal, retrospectivo, en donde los peces están agrupados acorde a la jaula y fecha en los cuales fueron cosechados.

4.3. Universo

El universo estará constituido por todos los peces ubicados en la Costa Atlántica de la provincia de Colón, Panamá.

4.4. Población de estudio

Para la siguiente investigación se ha delimitado como población de estudio al pez *Rachyentrum canadum* cosechados en diferentes jaulas flotantes ubicadas en la Costa Atlántica de la provincia de Colón, Panamá. Esta actividad es desarrollada por la empresa Open Blue S.A.

4.5. Criterios de inclusión

Serán incluidos dentro de este estudio los siguientes aspectos:

- Que estén ubicados en la Costa Atlántica de la provincia de Colón, Panamá
- La población de peces que sea plenamente perteneciente al grupo conocido científicamente como "*Rachyentrum canadum*" (Linnaeus, 1766) y vernacularmente como cobia. Se caracteriza por poseer cabeza ancha y deprimida, con la mandíbula inferior sobresaliente; primera dorsal con espinas aisladas cortas, no conectadas por membrana; aleta caudal bifurcada en adultos, con lóbulo superior más largo que el inferior; dorsal con 7 a 9 espinas y 31 radios blandos; anal con 2 espinas y 24 radios blandos.
- Los peces cobia recolectados se encuentren en jaulas flotantes pertenecientes a la empresa Open Blue, S.A.

4.6. Instrumento para la recolección de la información

Se hará uso de una hoja de cálculo en el programa informático Excel, en donde se colocará toda la información requerida para llevar un control de los datos. En la misma se detallará aspectos como la fecha de la cosecha (día, mes, año), número de la jaula, número de lote, peso inicial, peso final, cantidad de peces cosechados.

4.7. Variables y su definición conceptual

Las variables que se van a tomar en consideración en este estudio se van a clasificar según su naturaleza:

Tabla 3. Variables del presente estudio

VARIABLE	CÓDIGO	TIPO DE VARIABLE	DEFINICIÓN CONCECPTUAL
Días entre cada cosecha	DBH	Cuantitativa discreta	Periodo, en días, transcurridos entre cada cosecha.
Días entre cada cosecha	DBHa	Cuantitativa discreta	Periodo acumulado, en días, transcurridos entre cada cosecha.
Edad final	FA	Cuantitativa discreta	Cantidad de días que permanecieron los peces de una determinada jaula, desde que fueron colocados en las jaulas hasta su cosecha.
Edad inicio	SA	Cuantitativa discreta	Cantidad de días que duró el desarrollo de un grupo peces, desde la fase de huevo hasta el momento en que fueron colocados en la jaula flotante.
Jaula	PEN	Cualitativa ordinal	Nombre representativo para la identificación de la jaula flotante.
Mes	MO	Cualitativa ordinal	Mes en el que se realizó la cosecha de los peces en una determinada jaula flotante.
Número de la cosecha	HN	Cualitativa ordinal	Orden cronológica de cada cosecha.
Peso ganado	GWh	Cuantitativa continua	Promedio del incremento en kilogramos que obtuvieron los peces cosechados.

Peso inicial	IW	Cuantitativa continua	Promedio de peso en kilogramos de los peces al momento de ser colocados en la jaula flotante.
Peces restantes	RF	Cuantitativa discreta	Cantidad de peces que permanecen en la jaula después de cada cosecha.
Semana	WEEK	Cualitativa ordinal	Semana en el que se realizó la cosecha de los peces en una determinada jaula flotante.
Tasa de crecimiento	SGRh	Cuantitativa continua	Promedio de la tasa a la que aumentó los peces cosechados.
Total peces	TH	Cuantitativa discreta	Número de peces totales por cosecha en una fecha y jaula determinada.

4.8. Procedimiento Metodológico

4.8.1. Fase 1: Observaciones y mediciones

Se realizan las observaciones y mediciones por el personal capacitado de la empresa Open Blue, S.A.

a) Suministro de semilla

Para la producción de semillas, inicialmente se realizó la captura en el medio natural de juveniles o adultos de cobia para cultivarse y seleccionar los pie de cría. Una vez seleccionados los peces de las jaulas de crecimiento fueron transportados a estanques en tierra. Aproximadamente 100 peces en proporción de sexo 1:1 son colocados en estanques de desove, donde de manera espontánea o por inducción mediante la manipulación de fotoperiodo y temperatura producirán huevos durante todo el año.

b) Recepción y eclosión de huevos

Los huevos fertilizados fueron recolectados de los estanques de desove para ser incubados artificialmente. La eclosión tuvo lugar en las 24 horas posteriores al desove. El día posterior a la eclosión se realizó la estimación de larvas en el tanque.

c) Larvicultura

Para esta etapa las larvas han reabsorbido su saco vitelino y han experimentado metamorfosis hasta alcanzar su forma definitiva. El segundo día después de la eclosión se llenan los tanques de larvicultura con agua filtrada y 100 L de *Nannochloropsis oculata* con el fin de simular las condiciones naturales y estimular la producción de enzimas en el tracto digestivo.

A los tres días, aproximadamente, los especímenes en estado larvario son alimentados adecuadamente con rotíferos o nauplios de copépodos. Estos alimentos fueron suministrados los primeros cuatro días, luego los siguientes 25 a 30 días, tras su eclosión, se les proveyó de alimento seco, hasta alcanzar un peso promedio de 150 g.

d) Alvinaje y engorde

Una vez alcanzado el peso promedio de 1.50 kg, los juveniles fueron trasladados a jaulas ubicadas cerca la costa, para que su desarrollo diera continuidad en su hábitat natural, hasta alcanzar un peso promedio de 5.6 kg, aproximadamente. Para ello se le suministró de alimentos peletizados tanto flotantes como sumergibles (con 42-45% de proteína cruda y de 15 a 16% de lípidos), 6 días de la semana, a una tasa de entre 0,5 y 0,7 % del peso corporal/día hacia el final de la fase de engorde.

e) Cosecha

Una vez alcanzado el peso promedio requerido para poder ser cosechados, atendiendo a la demanda de venta, los especímenes fueron extraídos de las jaulas y transportados a la planta de procesamiento, donde se realiza la limpieza y el pesaje de cada uno de los individuos haciendo uso de una balanza digital. Toda esa información fue digitalizada y almacenada en una base de datos.

4.8.2. Fase 2: análisis de datos

Se trabajará con una base de datos proporcionada por la empresa estadounidense Open Blue Sea Farms S.A., cuyas operaciones de cultivo de cobia son realizadas en la costa Atlántica de Panamá, específicamente en Colón. La base de datos cuenta con un registro histórico de las cosechas del pez cobia que fueron realizadas durante los años comprendidos entre 2014-2016.

Para el análisis de los datos se siguió la metodología de "Proceso Estándar de la Industria Cruzada para la Minería de Datos), conocida por sus siglas en inglés como CRISP-DM (Cross Industry Standard Process for Data Mining). La misma está estructurada en seis fases:

1) **Comprensión del negocio o problema:** esta fase comprende el entendimiento desde una perspectiva de negocio, los requerimientos y objetivos que el cliente quiere alcanzar. Para obtener resultados fiables, es necesario comprender el problema que se quiere resolver, lo cual a su vez nos permitirá seleccionar los datos correctos que nos conducirán a la conversión del conocimiento del negocio en objetivos técnicos. En esta fase se pueden describir 3 principales tareas:

- *Identificar los objetivos del negocio*: determinar cuál es el problema que se quiere resolver y los criterios de éxito (de tipo cuantitativo o cualitativo). Para esto último, se requiere involucrar personas que tenga conocimiento del negocio.
- *Valoración de la situación*: se debe realizar una evaluación general antes de dar inicio al desarrollo del análisis. Establecer si son suficientes los datos proporcionados para resolver el problema, determinar todas las exigencias y preguntas del negocio, identificar las expectativas y necesidades del cliente, fijar el conocimiento previo del problema con el que se cuente. En esta fase se constituyen las necesidades del problema en términos del negocio y en términos de Data Mining.
- *Establecer los objetivos del Data Mining*: traducir y convertir los objetivos del negocio que se quiere alcanzar en un sistema operacional de Data Mining. Es decir, el objetivo del negocio es optimizar su venta de pez cobia ofreciéndole un mejor servicio a sus clientes acorde a sus exigencias, la finalidad del Data Mining es determinar el peso promedio por cosecha del pez cobia.

2) **Comprensión de los datos:** esta fase corresponde a la familiarización de los datos con los cuales se cuenta. Comienza con una fase inicial de recopilación de los datos, seguido de una descripción y exploración de los datos, para luego finalizar con la verificación de la calidad de los datos.

3) **Preparación de los datos:** esta fase incluye la tarea de selección de los datos que serán utilizados para aplicarles una técnica de modelado; limpieza de los datos: tratamiento de valores ausentes, normalización de datos, reducción del volumen de los datos, etc.; estructuración de los datos, que incluye: la generación de nuevos atributos a partir de los ya existentes, integración de nuevos registros, transformación de valores para atributos existentes; integración de los datos; formateo de los datos, que consiste en la transformación sintáctica de los datos sin modificarlos, por ejemplo: eliminar tabuladores, caracteres especiales, comas, etc.

4) **Modelado:** en esta fase se realiza una selección de las técnicas de modelados adecuadas acorde a los siguientes razonamientos: cumplimiento de los requisitos, conocimiento de la técnica, apropiada para el problema, datos apropiados. Las principales tareas que presenta esta fase son las siguientes:

- *Selección de la técnica de modelado*: consiste en elegir la técnica de Data Mining que se ajuste a los requerimientos del objetivo del proyecto y datos. Establecer si el problema es de predicción o clasificación.
- *Diseño de evaluación*: generar un procedimiento para probar la calidad y validez del modelo. Para ello se dividen el set de datos en datos de entrenamiento y datos de prueba, para una vez construido el modelo con los datos de entrenamiento, se mide su calidad con los datos de prueba.
- *Construcción del modelo*: ejecutar el algoritmo seleccionado a los datos previamente preparados, para generar los modelos.
- *Evaluación del modelo*

5) **Evaluación:** establecer si los modelos son útiles a las necesidades del negocio.

6) **Despliegue o implementación:** integrar el modelo generado, supervisar las etapas de los procesos para garantizar la precisión y validez de los resultados.

Basado en lo establecido anteriormente, con los datos proporcionados se realizará un análisis descriptivo y un análisis de predicción ejecutando el algoritmo de Random Forest para regresión. Se utilizó el software libre de programación estadística R-Studio versión 1.0.153 - © 2009-2017 RStudio, Inc. para realizar los análisis.

CAPÍTULO V

5. Resultados

Se utilizó una base de datos proporcionada por la empresa norteamericana, Open Blue, S.A., radicada en la costa atlántica de Panamá. La misma se dedica a la cosecha del pez *Rachycentrum canadum*, conocido vernacularmente como Cobia.

Los datos representan el peso promedio de cada uno de los peces que fueron cosechados durante los años 2014 al 2016. Cabe destacar que la información del año 2014 solo está presente los meses de octubre a diciembre y las del año 2016 hasta el mes de julio.

La información proporcionada, correspondiente a cada individuo (en este caso los peces cosechados), se agrupó tomando en consideración el número de jaula donde se encontraba el pez y la fecha en que fue cosechado. De esta forma se calculó el promedio de cada una de las variables, facilitando el análisis e interpretación de los datos.

5.1. Análisis Univariado

Entendiendo que se quiere obtener una aproximación del peso promedio por cada cosecha antes de que se lleve a cabo. Se realizó un análisis descriptivo de cada una de las variables que se tomaron en consideración en el presente estudio.

a) <u>Meses del año</u>

En el siguiente gráfico podemos observar los meses en donde se presentaron mayor número de cosechas a lo largo de los tres años de datos. Presenta una distribución bimodal, lográndose observar dos picos: enero y diciembre. Por otro lado, el mes donde hubo menor actividad fue en agosto (Gráfico 3).

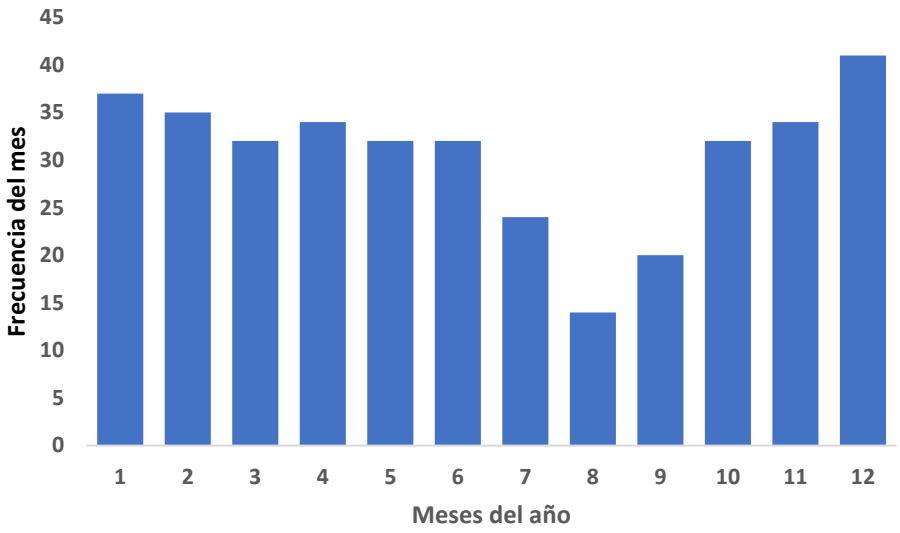

Gráfico 3. Frecuencia de los meses donde se presenta número de cosechas en los años de estudio.

b) <u>Número de cosechas</u>

La cantidad de cosechas, por su parte, que fueron realizadas en total en cada jaula para comenzar un nuevo ciclo desde la cría, vario escalonadamente. El gráfico de barras muestra un sesgo hacia la izquierda, lo que representa que el número óptimo de cosechas para realizar la venta completa de peces en una jaula es de 9-12 cosechas, aproximadamente. Si sobrepasa la cantidad de cosechas para poder terminar la venta de peces de una jaula, podría representar una pérdida para la empresa (Gráfico 4).

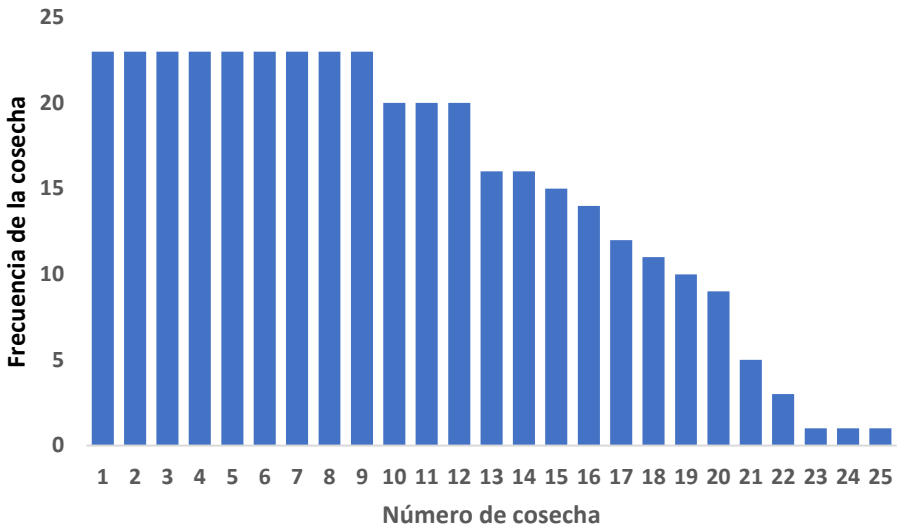

Gráfico 4. Frecuencia de cosechas presente a lo largo de los años de estudio.

c) <u>Número de jaula</u>

Las jaulas que se utilizaron con mayor frecuencia fueron las jaulas 15 y 16. Sin embargo, también se puede tomar en consideración la jaula 12. Contrario a esto la jaula que con menor frecuencia fue utilizada fue la 4. Este hecho se interpreta porque en la gráfica no se refleja una distribución equitativa de las veces que fueron tomadas en cuenta las jaulas para las cosechas. Los clientes siempre solicitarán un peso promedio aproximado de los peces al momento de realizar la compra, acorde a esto se evalúan las jaulas con los peces que poseen dicha especificación (Gráfico 5).

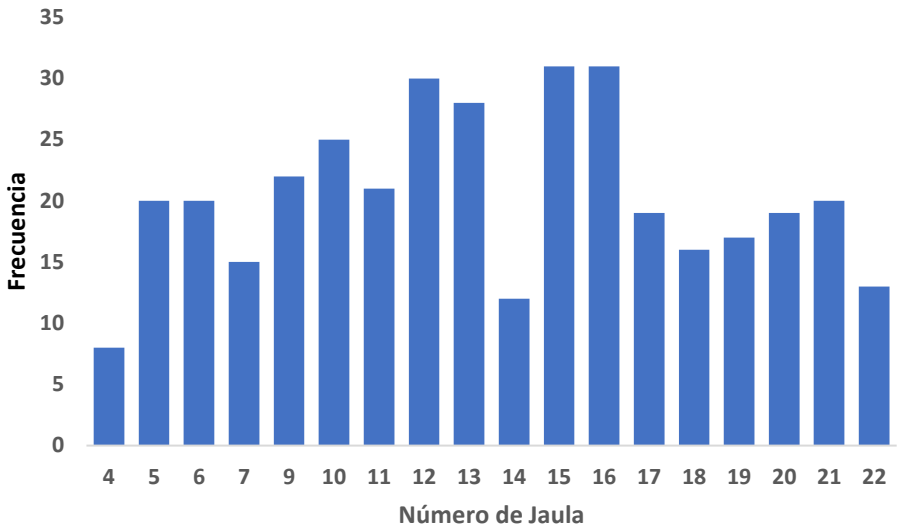

Gráfico 5. Frecuencia de la cantidad de veces en que fueron utilizadas las jaulas para cosechar.

d) <u>Edad de inicio</u>

La edad de inicio promedio que tenían los peces al momento de ser colocados en la jaula para seguir con su desarrollo y dar continuidad a la etapa de engorde, fue de 327 días, con una desviación de 57 días. El 50% de los datos indica que pasado los 331 días y comparándolo con la media, los peces adquieren el peso requerido para ser ubicados en el mar. Esto último lo podemos confirmar con la asimetría, la cual es negativa, lo que indica que los datos se aglomeran más hacia el lado derecho de la media, por ende, la media es menor que la mediana y ambas menores que la moda. La curtosis al ser negativa indica que la distribución es platicúrtica, es decir, que las puntuaciones tienden a dispersarse más que en la distribución normal.

Tabla 4. Estadística descriptiva de la variable "Edad de inicio".

Media	Error típico	Mediana	Moda	Desviación estándar	Curtosis	Asimetría	Máximo	Mínimo
327.43	2.99	331	410	57.44	-1.11	-0.15	410	228

e) Edad final

El promedio de la edad final que tenían los peces al momento de ser cosechados fue de 356 días, con una desviación de 55 días. El 50% de los datos indica que pasado los 357 días y con el valor similar a la media, los peces presentan el peso adecuado para ser cosechado. Esto último lo podemos confirmar con la asimetría, la cual es negativa, lo que indica que los datos se aglomeran más hacia el lado derecho de la media, por ende, la media es menor que la mediana. La curtosis al ser negativa indica que la distribución es platicúrtica, es decir, que las puntuaciones tienden a dispersarse más que en la distribución normal. Los peces pueden llegar hasta un máximo de 458 días en el mar hasta su cosecha

Tabla 5. Estadística descriptiva de la variable "Edad final".

Media	Error típico	Mediana	Moda	Desviación estándar	Curtosis	Asimetría	Máximo	Mínimo
356.50	2.87	357	346	55.02	-0.68	-0.31	458	228

f) Peso inicial

El peso promedio que tenían los peces al momento de ser colocados en las jaulas de engorde, fue de casi 0.1 kg, con una desviación de 0.08 kg, es decir que la talla de los peces no mantuvo una uniformidad en las tallas antes de ser ingresadas al mar. La asimetría es positiva, es decir, que hay una mayor concentración de los datos hacia el lado izquierdo de la media y esta es mayor que la mediana. Esto último nos indica que como el valor de la mediana es de 0.0566 kg., es el peso que en su mayoría los peces tenían al momento de ser ubicados en las jaulas. La curtosis al ser positiva indica que la distribución leptocúrtica, es decir, que sus valores tienen a concentrarse en torno a la media más que en una distribución normal.

Tabla 6. Estadística descriptiva de la variable "Peso inicial".

Media	Error típico	Mediana	Moda	Desviación estándar	Curtosis	Asimetría	Máximo	Mínimo
0.0977	0.0046	0.0566	0.115	0.0897	1.0677	1.4213	0.3681	0.025

g) Días entre cada cosecha

El promedio de días que demoró realizar otra cosecha fue de 3 días, con una desviación de 5 días. La mínima cantidad de días en que se extrajo peces de la jaula para la venta fue de 1 día, sin embargo, el máximo de días fue de 55. Como se podrá observar, estos valores podrían estar interfiriendo con la media y se puede notar con el valor positivo de la asimetría la cual nos está indicando que la carga de los valores se concentra del lado izquierdo del valor de la media. También es observable con el valor de la mediana el cual es de 1 día.

Tabla 7. Estadística descriptiva de la variable "Días entre cada cosecha".

Media	Error típico	Mediana	Moda	Desviación estándar	Curtosis	Asimetría	Máximo	Mínimo
3.30	0.29	1	1	5.73	32.91	5.06	55	1

Peso ganado

Tabla 8. Estadística descriptiva de la variable "Peso ganado".

Media	Error típico	Mediana	Moda	Desviación estándar	Curtosis	Asimetría	Máximo	Mínimo
5.64	0.04	5.66	5.41	0.79	3.24	-0.96	7.41	2.19

5.2. Análisis Bivariado

a) Peso vs Cosecha

El peso promedio de los peces fue muy variante entre cada cosecha. Entre la 1^{era} y 12^{ava} cosecha existe mucha variabilidad en los datos, llegándose a observar una menor dispersión del peso promedio de los peces a partir de la 13^{ava} cosecha. A partir de la 21^{ava} cosecha los peces mantienen un peso similar (sin dispersión de los datos). De forma generalizada a partir de la cosecha 13 ^{ava} se puede asegurar obtener peces con un peso mayor a 5 kilogramos (con algunas pequeñas leves excepciones) y de la 7ma a la 12^{ava} cosecha, peces con pesos promedio mayores a 7 kilogramos. Valores atípicos son observables en las siguientes cosechas: 1 a la 5 (con valores menores a 3 kg) y de la cosecha 15 a la 19 (Gráfico 6).

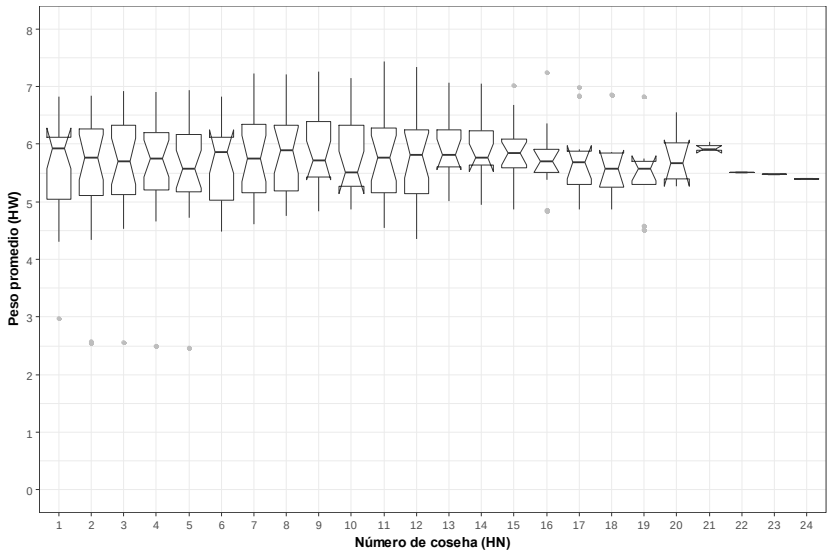

Gráfico 6. Peso promedio de los peces cobia por número de cosecha

b) Peso vs Jaula

El peso promedio de los peces en las diferentes jaulas es muy variable entre ellas, sin embargo, cada jaula no presenta mucha dispersión de sus valores, a excepción de las jaulas 9, 13 y 15 que presentan una mayor variabilidad en sus datos con una distribución positiva. La jaula cuyo peso promedio presentan los menores valores (menos de 4 kilogramos) es la jaula 9, las que tiene valores entre 4 y 5 kilogramos son las jaulas 4, 15, 16, 18 y valores más altos son las jaulas 11 y 22. Se presentan valores atípicos en la mayoría de las jaulas los cuales podrían conducir a una interpretación engañosa. Se puede observar valores atípicos por debajo del primer cuartil en las jaulas 5, 14, 21 y valores por encima del tercer cuartil en las jaulas 4, 18, 20 (Gráfico 7).

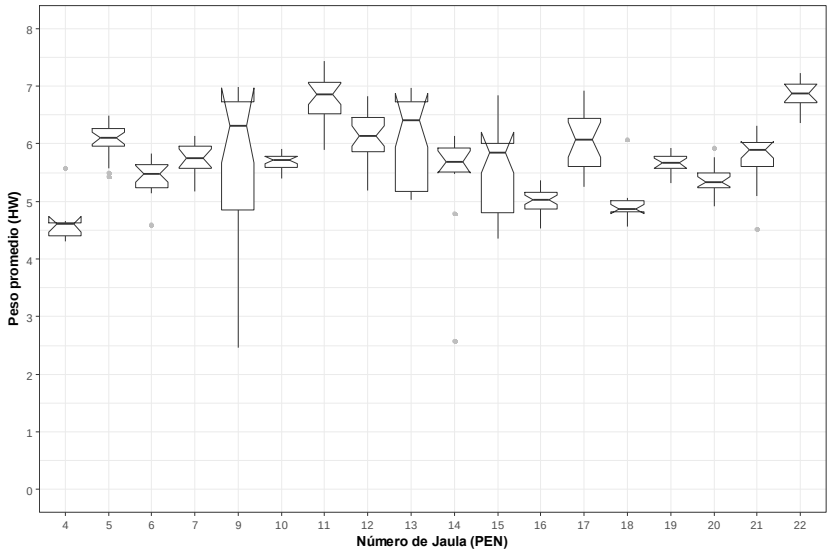

Gráfico 7. Peso promedio de los peces cobia por jaula.

c) Peso vs Número de días en el mar

A medida que se da el aumento de los días que permanecen los peces en el mar, su crecimiento va en descenso (Gráfico 8). El peso óptimo que llegan a tener los peces, en términos generalizados, lo comprenden cuando tienen de 300 a 400 días en el mar, llegando a alcanzar un peso mayor a 5.5 kilogramos.

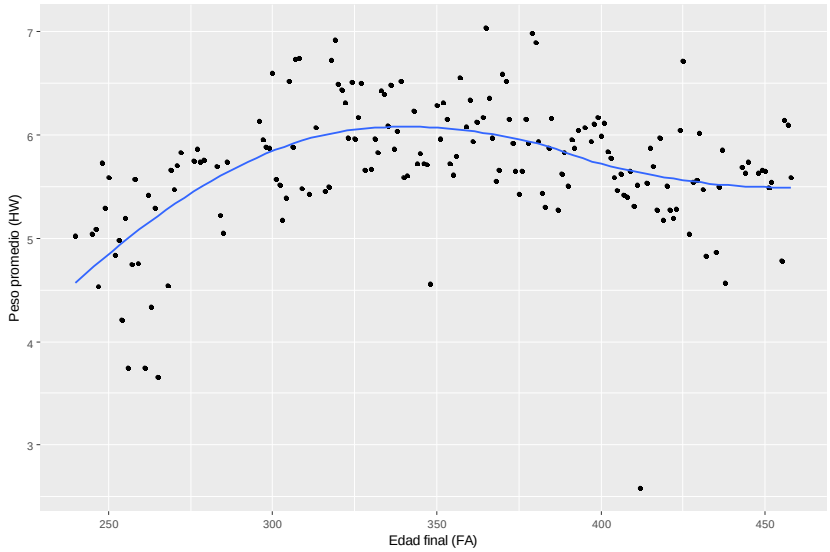

Gráfico 8. Peso promedio de los peces cobia vs edad final.

d) Peso vs Tasa

La tasa de crecimiento representada en el Gráfico 9 está expresada en porcentaje, donde entre 0.9 y 1.5% se concentra la mayor cantidad de valores, representando el aumento en peso que tienen los peces, donde llegan a pesar más de 7 kilogramos. Por otro lado, se observa tasas de crecimiento mayor a 1.6% y en contraparte el peso de los peces no llega a aumentar más, sino que se mantiene en valores estándares >5 y <6 kilogramos, principalmente. Otros valores atípicos que también se presentan están comprendidos entre las tasas de crecimiento >1.3% y <1.8% que se correlacionan con pesos entre 2 y 3 kilogramos.

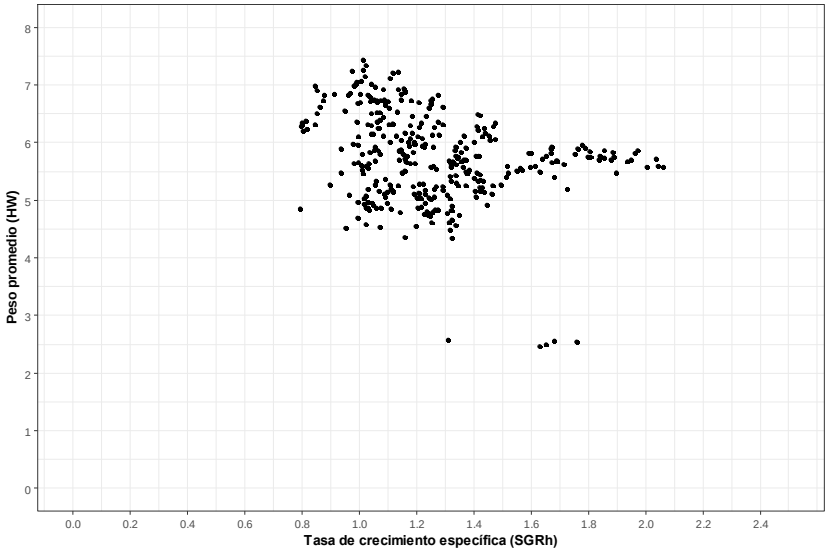

Gráfico 9. Peso promedio de los peces cobia vs su tasa de crecimiento específica

e) Peso vs Mes

En el diagrama de caja (Gráfico 10) podemos observar que el peso promedio de los peces varia a lo largo del año. Durante los primeros 5 meses el peso va en aumento, luego en los dos meses subsiguiente se observa un descenso. Posteriormente el peso incrementa nuevamente, manteniéndose relativamente estable. De forma generalizada podemos indicar que no existe mucha variabilidad en los pesos, sin embargo, los meses que presentan una variabilidad mínima comprende los meses marzo, julio a novimebre y los meses que presentan una mayor variabilidad son enero, febrero, abril a junio y diciembre. Se presenta una distribución simétrica en marzo, agosto, septiembre y noviembre, presentando dispersiones parecidas. Caso contrario a los demás meses que presentan una distribución asimétrica negativa (sesgada a la izquierda) con tendencia a valores por encima de la mediana, excepción de los meses de junio, julio y octubre, cuyas puntuaciones están sesgadas hacia la derecha (asimetría positiva) mayormente orientadas a valores menores y las de mayor puntaje se encuentran más dispersas.

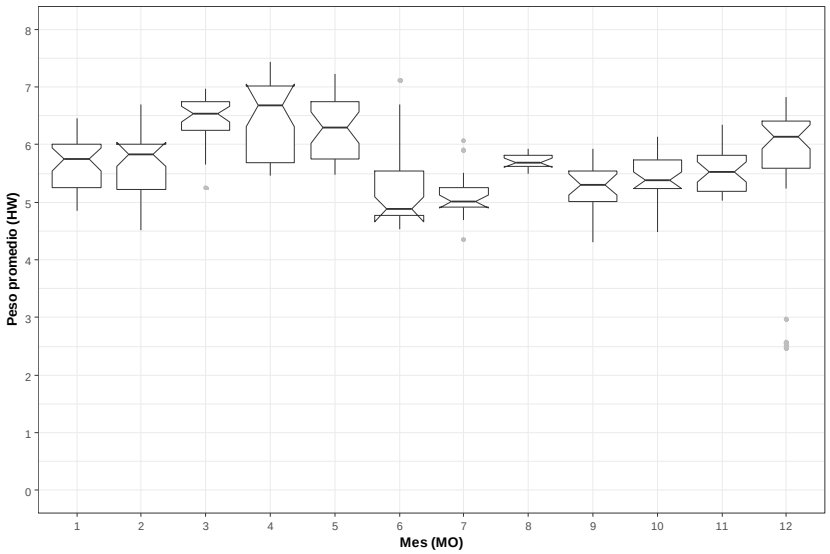

Gráfico 10. Peso promedio de los peces cobia por mes.

5.3. Modelo de predicción estadístico en R

Para la ejecución del presente proyecto se llevaron a cabo distintos modelos de "Random forest" en R-Studio. Se comparó cada uno de los modelos y se seleccionó el que presentaba la mejor predicción, calculado a partir del menor error. De esta manera, con el modelo de predicción seleccionado, la empresa podrá determinar con un error menor al 30% el peso promedio de los peces por cosecha.

5.3.1. Preparación de los datos analizados

Para el desarrollo de los objetivos se trabajó con la fuente de datos que fue mencionada con anterioridad, a la cual se realizó los análisis y se trabajó sobre la misma. Previo a la creación de los modelos, es necesario la correcta preparación de los datos, leyendo y ejecutando desde R-Studio los comandos requeridos.

Los datos están recogidos en un archivo Excel de extensión .csv (archivo de valores separados por comas) llamado **"Modelocobiaoriginaldata2018"**. El archivo está constituido por una hoja, lo que facilita el código en R para su lectura. La información corresponde a distintos aspectos para la empresa en cuestión, sobre la cosecha del pez cobia. Dicha información, representada en columnas con sus respectivos nombres, serán las variables de entrada que se utilizaron.

Por otro lado, la columna que constituye la de mayor interés es: "HW", que será la variable a estudiar en este trabajo. Como se mencionó en el Tabla 3 los valores de esta columna representan el peso promedio de los peces por cosecha en una fecha y jaula determinada.

5.3.2. Códigos en R-Studio

En primera instancia, se cargó los siguiente paquetes y librerías en R-Studio:

- Paquetes: “randomForest” para poder trabajar con múltiples árboles de decisión y usar la función randomforest, "psych", "ranger", "Boruta", "randomForestSRC", "e1071", "caret", "randomForest", "rpart.plot", "rpart", "caTools".
- Librerías: caTools, randomForest, e1071, caret, dplyr, psych, corrplot, gplot2, dplyr, rpart, rpart.plot, randomForest, modelr, purr, ranger, Boruta.

Para determinar el directorio actual sobre el cual se trabajó, se ejecutó el comando **getwd ()** y para establecerlo se utilizó el comando **setwd ()**. Posteriormente se abrió y llevó a cabo la lectura del archivo con el comando **read.csv**.

Para garantizar la calidad en el proceso de predicción de los modelos generados, se dividieron los datos en dos grupos: datos de entrenamiento (datatrain), que fueron utilizados para crear el modelo y datos de prueba (datatest), para realizar comparaciones sobre si la predicción es buena o mala. El primer grupo constituyó el 70% de los datos, mientras que el segundo grupo comprendió el 30% restante.

Durante el proceso de selección del mejor modelo, los modelos se ajustan a los datos de entrenamiento y el error de predicción para dichos modelos es obtenido mediante el uso de los datos de prueba.

La elección de qué datos pertenecen a cada grupo es aleatoria, pero el programa R-Studio es capaz de realizar siempre el mismo “sorteo”. Para ello, se seleccionó el número 1468 que representa el “seed” para que cada vez que se ejecute el código siempre sea seleccionado el mismo conjunto de datos, aunque fuesen escogido de forma aleatoria.

Cabe destacar que se pudo haber seleccionado cualquier número, lo importante es que siempre sea el mismo cada vez que se ejecute los algoritmos de los modelos en estudio. Es de vital importancia, pues se desea comparar la calidad de predicción de cada uno de los modelos ejecutados y para ello es necesario tratar con los mismos conjuntos de datos, evitando las variaciones que surgen a favor o en contra de la predicción por aleatoriedad.

5.3.3. Modelos - Random Forest

Al instalar el paquete library (randomForest) en el programa R-Studio, se implementa el algoritmo de Random Forest (bosque aleatorio) de Breiman para regresión, generando un gran número de árboles de regresión a partir de los datos de partida. Con estos mismos datos se obtiene el error Mean Square Error (MSE) o Error Cuadrático Medio y Variabilidad explicada del modelo.

El Error Cuadrático Medio o Varianza Residual, representa la variación de la variable dependiente debido a causas no controladas por el modelo, por lo que valores altos suponen un modelo no óptimo.

La variabilidad explicada por su parte es el coeficiente de determinación R^2, el cual indica el porcentaje de variabilidad explicada por el modelo. Si su valor es cercano a 1, expresa un poder explicativo alto del modelo, sin embargo, no quiere decir que las predicciones que se vayan a realizar sean buenas, ya que como se presenta en un análisis de regresión, si se trata de predecir fuera del rango de las variables predictoras para las que se ha construido un árbol, su predicción puede estar mal ejecutada.

Posteriormente, cada observación se alimenta en cada uno de los árboles creados y el resultado más común se utiliza como salida final, consiguiendo predecir los valores de la variable en estudio.

En el primer modelo realizado se empleó como posibles variables de entrada todas las proporcionadas en el archivo Excel **"Modelocobiaoriginaldata2018"**. Usando la función para ejecutar el análisis de RandomForest, se presenta los siguientes argumentos para desarrollar el modelo básico:

model = randomForest (formula, data)

Posteriormente, se realizó la predicción objetivo del presente estudio, para ello se ejecutó la función **predict ()**. Esta función calcula los valores predichos para datos nuevos (en este caso los datos de prueba) de un modelo ya ajustado.

Los criterios utilizados para la evaluación de la eficiencia de los modelos fueron determinados calculando las medidas de error o errores de predicción: Mean of squared residuals, Root Mean Square Error (RMSE) y el Mean Absolute Percent Error (MAPE). Cuanto menor sean sus valores mejor será su capacidad de predicción.

Estas métricas de error representan la diferencia entre el valor predicho y el valor verdadero y miden el rendimiento de predicción de un modelo de regresión en cuanto a la desviación media de sus predicciones a partir de los valores reales. Cuanto menor sean los valores de error implican que el modelo es más preciso a la hora de realizar predicciones. Una métrica de error general de "0" es indicativo que el modelo se ajusta perfectamente a los datos.

El primer modelo ejecutado se llevó a cabo integrando todas las variables de entradas y como "seed" se utilizó el número 1468. Los resultados que muestra la consola se presentan a continuación:

MODELO #1

```
> cobiaForest
Call:
randomForest (formula = HW ~ WEEK + FA + RF + MO + PEN + SGRh + SA +
HN + DBH + DBHa + IW + TH + HWh, data = cobiaTrain)

                Type of random forest: regression
                      Number of trees: 500
    No. of variables tried at each split: 5 4

             Mean of squared residuals: 0.1914079 0.1672545
                       % Var explained: 69.67 73.5

#predicción
> PredictcobiaForest = predict (cobiaForest, newdata = cobiaTest)
tree.sse = sum ((PredictcobiaForest - cobiaTest$HW)^2)

> tree.sse
[1] 13.85314 11.2190

#Medida de error: RMSE
> RootMeanSquareError(cobiaTest$HW, PredictcobiaForest)
[1] 0.3598174 0.3280

#Medida de error: MAPE
> MeanAbsolutePercentError(cobiaTest$HW, PredictcobiaForest)
[1] 4.67001 4.4944
```

Se realizó un segundo modelo, esta vez a haciendo uso de los atributos o variables explicativas que son importantes, pero se mantuvo el uso del mismo "seed" (1468). La técnica Random Forest también determina la influencia e importancia que tienen las variables explicativas sobre la variable dependiente. En el presente estudio, como la variable de objeto es cuantitativa (regresión), la importancia es medida por (como se mencionó en el capítulo 3):

- **El incremento del error MSE (IncMSE%)** del modelo cuando la variable en cuestión es permutada. Esto indica un descenso en la precisión de las predicciones cuando la variable es excluida del modelo y sustituida por otra.
- **El incremento de pureza de los nodos (IncNodePurity),** las variables más útiles consiguen mayores aumentos en la pureza del nodo, es decir, encuentran una división que tenga una varianza entre nodos alta y una varianza en el interior del nodo pequeña.

Con **cobiaForest$importance** se obtiene una matriz para cada uno de los criterios de importancia (Anexo 1). Sin embargo, podemos graficar la importancia de las variables ejecutando **varImpPlot()**. En el Gráfico 11 se puede apreciar un diagrama con la importancia de las variables basadas en los criterios previamente mencionados.

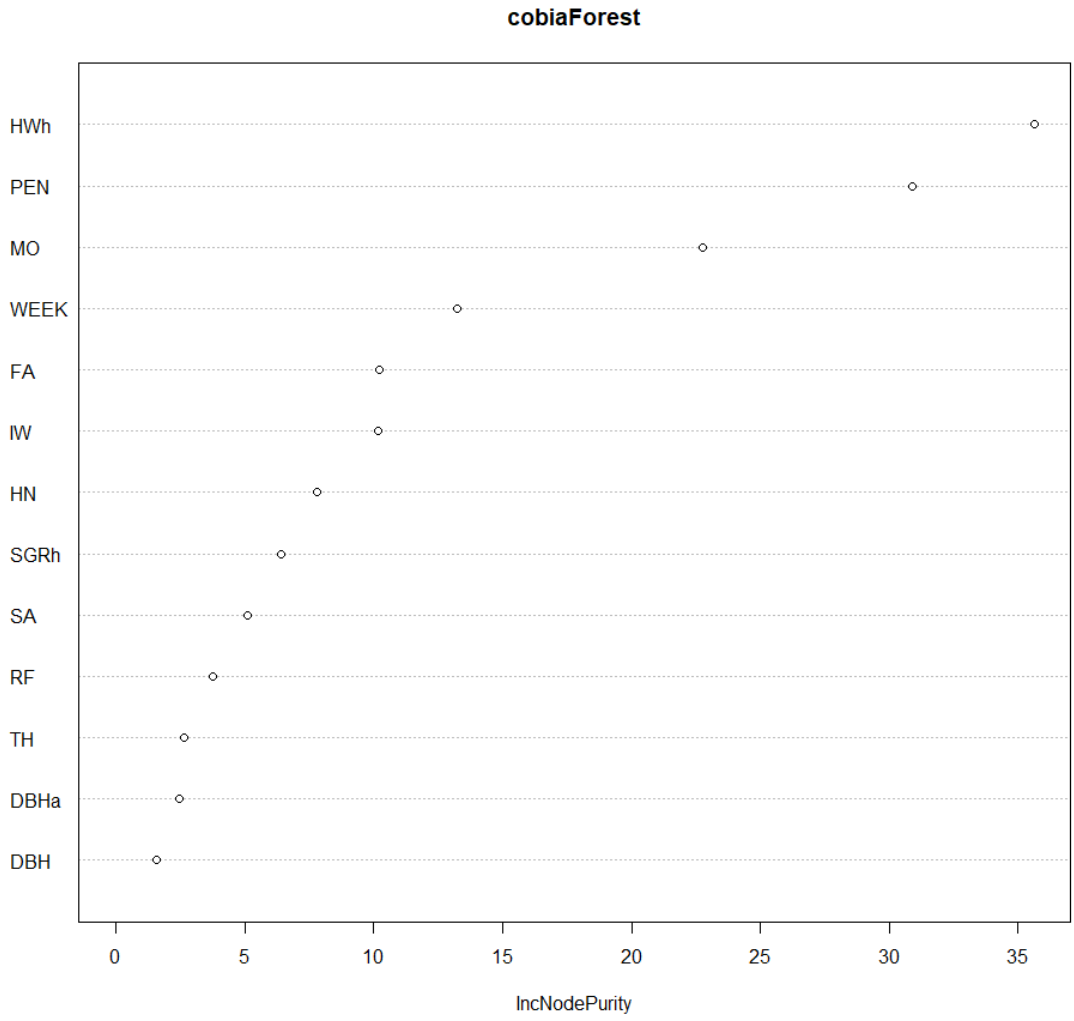

Gráfico. 11. Gráficos de importancia de las variables explicativas.

Las variables que presentan mayor influencia en la predicción del peso promedio de los peces cobia en cada cosecha son: PEN, WEEK, MO, FA, IW, HN, SA SGRh, HWh. Basados en estas variables que el modelo inicial ha deducido como más importantes se obtienen lo siguiente:

#MODELO 2

```
> cobiaForest
Call:
randomForest (formula = HW ~ SA + PEN + WEEK + MO + FA + IW + SGRh + HN + HWh,
data = cobiaTrain)

                  Type of random forest: regression
                        Number of trees: 500
    No. of variables tried at each split: 3


              Mean of squared residuals: 0.1861459 0.1697083
                        % Var explained: 70.05 73.11

#predicción
> PredictcobiaForest = predict (cobiaForest, newdata = cobiaTest)
tree.sse = sum ((PredictcobiaForest - cobiaTest$HW)^2)

> tree.sse
[1] 11.20187

#Medida de error: RMSE
> RootMeanSquareError(cobiaTest$HW, PredictcobiaForest)
[1] 0.323559

#Medida de error: MAPE
> MeanAbsolutePercentError(cobiaTest$HW, PredictcobiaForest)
[1] 4.487986
```

Al ejecutar el nuevo modelo con las variables que presentan mayor importancia, las medidas de error disminuyeron. A partir de esta premisa, se ejecutaron 73 modelos con distintas combinaciones de variables de entrada que presentaban dar mayor aporte al modelo y que en conjunto darían mejor respuesta a la predicción (ver Anexo 2). El mejor modelo que se obtuvo se presenta a continuación:

MODELO #3

```
> cobiaForest
Call:
randomForest (formula = HW ~ PEN + WEEK + SA + IW, data = cobiaTrain)

                  Type of random forest: regression
                        Number of trees: 500
    No. of variables tried at each split: 1


              Mean of squared residuals: 0.15262
                        % Var explained: 75.81

#predicción
> PredictcobiaForest = predict (cobiaForest, newdata = cobiaTest)
tree.sse = sum ((PredictcobiaForest - cobiaTest$HW)^2)

> tree.sse
[1] 7.585491

#Medida de error: RMSE
> RootMeanSquareError(cobiaTest$HW, PredictcobiaForest)
[1] 0.2662563

#Medida de error: MAPE
> MeanAbsolutePercentError(cobiaTest$HW, PredictcobiaForest)
[1] 3.614228
```

La combinación de las variables de entrada PEN, WEEK, SA e IW, permitió un obtener menor error en el modelo de predicción. Para este modelo presentado también se hizo uso del "seed" 1468.

5.3.3.1. Optimización de Modelo #3

A. Parámetros

Random Forest ofrece también la ventaja de poder incluir parámetros al modelo, que son utilizados cuando se quiere optimizar el modelo. Los siguientes parámetros que se tomarán en consideración son los siguientes:

- ***Nodesize*:** constituye el número de nodos terminales.
- ***ntree***: el software por defecto elige un total de 500 árboles. Por recomendación no debería presentarse un número pequeño de árboles para que la calidad de predicción de cada observación al bosque mejore.
- ***mtry***: los valores por defecto que presenta el programa para los casos de los modelos de regresión se calculan por medio del cociente:

$$m/3$$

donde:

m = número de variables inicial

Para la elección de los valores óptimos de los parámetros se aplicó un algoritmo que determina, dado un conjunto de valores para cada parámetro, el valor que mejor se ajuste para la ejecución del modelo. Los valores de los parámetros que ayudan a mejorar el modelo son:

- nodesize = 5
- ntree = 200
- mtry = 3

Aplicando estos valores, se presentan las salidas con los valores de las medidas de error que han sido utilizadas para medir la eficacia que tiene el modelo para predicción:

MODELO #4

```
Parameter tuning of ‘randomForest’:
- sampling method: 10-fold cross validation

- best parameters:
  nodesize mtry ntree
         5    3   200

- best performance: 0.2569471

>   tuning.results$best.model

Call:
best.randomForest (x= formula.new, data = cobiaTrain, nodesize = nodesize.vals,
mtry = mtry.vals, ntree = ntree.vals)

                      Type of random forest: regression
                            Number of trees: 200
        No. of variables tried at each split: 3


                Mean of squared residuals: 0.1433463
                          % Var explained: 77.29


#predicción
> PredictcobiaForest = predict (cobiaForest, newdata = cobiaTest)
tree.sse = sum ((PredictcobiaForest - cobiaTest$HW)^2)

> tree.sse
[1] 7.991468

#Medida de error: RMSE
> RootMeanSquareError(cobiaTest$HW, PredictcobiaForest)
[1] 0.2732885

#Medida de error: MAPE
>MeanAbsolutePercentError(cobiaTest$HW,PredictcobiaForest)
[1] 3.557066
```

B. Seed

Al determinar los valores para cada uno de los parámetros, se optimizó el Modelo #4 ejecutando un ciclo que determina cual es el mejor "seed" (o semilla que influye en la aleatoridad al momento de la construcción de un árbol), a partir de un conjunto de números que se le fue proporcionado.

MODELO #5

```
> cobiaForest

Seed 142
Call:
randomForest (formula = HW ~ PEN + WEEK + SA + IW, data = cobiaTrain, nodesize
     = 5, ntree = 200, mtry =3)

                               Type of random forest: regression
                                     Number of trees: 200
                No. of variables tried at each split: 3

                          Mean of squared residuals: 0.1530324
                                    % Var explained: 77.78


#predicción
> PredictcobiaForest = predict (cobiaForest, newdata = cobiaTest)
tree.sse = sum ((PredictcobiaForest - cobiaTest$HW)^2)

> tree.sse
[1] 4.563912

#Medida de error: RMSE
> RootMeanSquareError(cobiaTest$HW, PredictcobiaForest)
[1] 0.2065269

#Medida de error: MAPE
> MeanAbsolutePercentError(cobiaTest$HW,PredictcobiaForest)
[1] 2.906536
```

5.3.3.2. Evaluación del modelo: cross-validation (validación cruzada)

En el último modelo que fue generado (modelo #5), se logró optimizarlo y obtener menores valores en las medidas de error. Por lo que ha sido considerado el mejor modelo que ha generado hasta el momento.

La evaluación del rendimiento de un modelo constituye una de las fases principales en el proceso, lo cual es indicativo del nivel de acierto de las predicciones de un conjunto de datos a través de un modelo entrenado.

Sin embargo, cuando generamos un modelo para realizar predicciones, no es suficiente con su creación, optimización y evaluación por el método tradicional (dividiendo el conjunto de datos en datos de entrenamiento y datos de prueba), ya que podemos entrar en el efecto de sobreentrenar el modelo con datos para los que se conoce el resultado deseado.

Es decir, el modelo debe ser capaz de predecir el resultado de otro conjunto de datos (que desconoce) a partir de lo aprendido con los datos de entrenamiento, generalizando las reglas para predecir los datos que no ha visto. Cuando se sobreentrena el modelo, el mismo puede quedar ajustado a unas características específicas de los datos de entrenamiento o ha memorizado los datos que ha visto. Es por ello, para conocer la precisión que puede tener un modelo y reducir el sobreajuste que pueda tener el modelo, utilizamos la técnica de validación cruzada.

En el método de validación cruzada de k iteraciones o k-fold cross-validation, los datos de entrada se dividen en k subconjuntos, reservando una parte para realizar las pruebas y las otras 9 para llevar a cabo el entrenamiento. Este proceso es repetido k veces y el resultado final es la media aritmética de los valores obtenidos para las diferentes divisiones. Esto ayuda a determinar el nivel al que un modelo se podría generalizar para nuevos conjuntos de datos.

A partir de lo establecido con anterioridad, se realizó una validación cruzada para determinar qué tan buena será la predicción del modelo cuando se le proporcione futuros datos. Para ello, se dividió la data en 10 iteraciones y el proceso se repitió 10 veces. Se utilizó un “seed” número 100. Los resultados se presentan a continuación:

MODELO #6

```
 MSEcobia
# A tibble: 10 x 5
```

	Run	MSE	RMSE	MAPE	R2
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	0.120	0.347	4.48	0.766
2	10	0.0513	0.226	3.05	0.820
3	2	0.118	0.343	3.19	0.720
4	3	0.0760	0.276	3.63	0.890
5	4	0.285	0.534	6.78	0.704
6	5	0.0333	0.182	2.63	0.942
7	6	0.0886	0.298	3.55	0.863
8	7	0.115	0.340	4.41	0.849
9	8	0.0490	0.221	2.73	0.934
10	9	0.0638	0.253	3.46	0.860

```
> mean(MSEcobia$MSE)
[1] 0.1000695

> mean(MSEcobia$RMSE)
[1] 0.3020002

> mean(MSEcobia$MAPE)
[1] 3.790912

> mean(MSEcobia$R2)
[1] 0.8347004
```

En los resultados obtenidos de las métricas de error podemos observar que el modelo anterior posiblemente presentaba un sobreajuste y al realizar la evaluación mediante la validación cruzada se logra comprobar que los modelos anteriores no presentaban una predicción más realista como el del modelo #6. Por otro lado, este modelo presentó una mayor varianza.

6. Conclusión

La ejecución del presente proyecto tenía como finalidad, desde el punto de vista técnico, establecer un modelo que con un error no mayor al 30% pudiera estimar el peso promedio de los peces *Rachycentrum canadum* al momento de ser cosechados. El cual fue posible llevar a cabo ejecutando en el software R-Studio el algoritmo de Random Forest con la base de datos que fue proporcionada por la empresa Open Blue, S.A.

- El algoritmo de Random Forest es una técnica muy utilizada, debido a su simplicidad y al hecho de que se puede usar para tareas de clasificación y regresión. Evita el sobreajuste la mayor parte del tiempo, creando subconjuntos aleatorios de las características y construyendo árboles más pequeños usando estos subconjuntos. Luego, combina los subárboles. Hay que tomar en consideración que el hecho que evita el sobreajuste no funciona todas las veces y que también hace que el cálculo sea más lento, dependiendo de cuántos árboles genere su bosque aleatorio.
- Las variables más importantes que influyeron de forma más significativa en la variabilidad de respuesta y que fueron las que mejor explicaron el modelo para estimar el peso promedio por cosecha del pez cobia son: semana (WEEK), peso inicial (IW), edad de inicio (SA), jaula (PEN).
    - Para la muestra de datos, es posible obtener una combinación de parámetros de funcionamiento de Random Forest, arrojando resultados más precisos: a) **ntree:** 300, b) **mtry:** 3, c) **nodesize:** 5.
- Se obtuvo un ajuste a la predicción evaluado mediante la validación cruzada obteniendo valores con las siguientes medidas de error:
    - **RMSE:** 0.3020002
    - **MAPE:** 3.790912
- Con el modelo final se obtuvo un 83% de la varianza explicada.

Desde el punto de vista práctico, su finalidad, una vez obtenido el mejor modelo de predicción, se proporcionará a la empresa Open Blue, S.A. las directrices para poner en marcha las herramientas que consistirán en un programa en lenguaje R, los cuales ejecutarán cada vez que requiera realizar la predicción del peso de los peces para la toma de decisiones.

7. Limitaciones y Recomendaciones

a) Limitaciones

- No se tuvo acceso a otras variables de tipo biológicas, climáticas, físico-químicas que pueden influir en el desarrollo del pez cobia. Las variables que fueron incorporadas en el estudio fueron determinadas en conjunto con los representantes de la empresa, los cuales conocen sobre la materia en estudio.
- Son pocas las referencias bibliográficas referentes a trabajos similares de predicción asociados a organismos acuáticos.

b) Recomendaciones

- Enriquecer la base de datos para que haya más datos disponibles para entrenar el modelo, de esta forma asegurar un mejor funcionamiento del algoritmo. Si introducimos basura, obtendremos basura ("Garbage in- Garbage out").
- Realizar los análisis incorporando un número mayor de variables, especialmente las de tipo biológica o que estén más relacionadas con el crecimiento y desarrollo del pez cobia.
- Utilizar otro tipo de algoritmos, con el objetivo de ver que técnica es más apropiada para el ajuste del modelo y comparar mediante una prueba de t-student si existen diferencias significativas de uno con respecto al otro.

8. Referencias Bibliográficas

1. Altmann, A., Tolosi, L., Sander, O. and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics 26*, 1340-1347.

2. Amit, Y. and Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural Computation 9* (7), 1545–1588.

3. Benson, N. G. (1982). Life history requirements of selected finfish and shellfish in Mississippi Sound and adjacent areas. U.S. Fish and Wildlife Service, Office of Biological Services, Washington, D.C. FWS/OBS-81/51. 97pp.

4. Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

5. Breiman, Leo. (1996). Bagging Predictors. Machine Learning, 24. 123-140 pp.

6. Breiman, Leo. (2001). Random Forest. University of California, Bekerley.

7. Carmona, Antonio. (1997). Toma de decisiones: análisis y entorno organizativo. Universidad Politécnica de Catalunya. España. 150 pp.

8. Collette, B. B. (1978). Rachycentridae. In Fischer, W. (ed.), FAO species identification sheets for fishery purposes, western central Atlantic (Fishing area 31), vol. 4. FAO, Rome, unpaginated.

9. Darracott, A. (1977). Availability, morphometrics, feeding and breeding activity in a multi-species, demersal fish stock of the western Indian Ocean. *J. Fish. Biol. 10* (1), 1-16.

10. Espino, C. y Martínez, X. (2017). Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo - herramientas Open Source que permiten su uso. Tesis de Grado. 61 pp.

11. Freeman, B. L. and Walford, L. A. (1976). Angler´s guide to the United States Atlantic coast: Fish, fishing grounds & fishing facilities. Sect. VII: Altamaha Sound, Georgia to Fort Pierce Inlet, Florida. Natl. Mar. Fish. Serv., NOAA, Seatle, WA 98115-0070. 21 pp.

12. González, Nuria V. (2015). Técnicas de Machine Learning para el Post-Proceso de la predicción de la Irradiancia. Tesis de Maestría. Universidad de Granada. 60 pp.

13. Goodson, G. (1985). Fishes of the Atlantic coast. Stanford Univ. Press, Stanford, CA. 204p.

14. Hadidi, N. (2003). "Classification ratemaking using decision trees". CAS Forum.

15. Hastie, T., Tibshirani, R. and Friedman, J. (2001). The elements of statistical learning: Data mining, inference, and prediction. Springer Verlag. 737 pp.

16. Hastie, T., Tibshirani, R. and Friedman, J. (2008). The Elements of Statistical Learning – Data Mining, Inference, and Prediction. Springer. Stanford, USA. pp

17. Hassler, W. W. and Rainville, R. P. (1975). Techniques for hatching and rearing cobia, *Rachycentron canadum*, through larval and juvenile stages. Publ. UNC-SG-75-30, Univ. N.C. Sea Grant Coll. Prog., Raleigh, NC 27650-8605. 26 pp.

18. Ho, Tin Kam. (1998). The Random Subspace Method for Constructing Decision Forests. IEEE Transactions on Pattern Analysis and Machine Intelligence. 20: 832-844.

19. Hoese, H. D. and Moore, R. H. (1977). Fishes of the Gulf of Mexico; Texas, Louisiana, and adjacent waters. Texas A&M Univ. Press, College Station, TX. 327 pp.

20. Holt, G. J., Faulk, C. K. and Schiwarz, M. H. (2007). A review of the Larviculture of cobia. *Rachycentron canadum*, a warm water marine fish. *Aquaculture, 268,* 181- 187.

21. Kaiser, J. B. and Holt, G. J. (2004). Cobia: a new species for aquaculture in the US. World Aquaculture, 35, 12-14.

22. Lantz, Brett. (2013). Machine Learning with R. Packt Publishing. 396 pp.

23. López, Elena. (2015). Desarrollo de un modelo de indicadores de mejora de la accesibilidad en establecimientos de pequeño tamaño susceptibles de ajustes razonables. Tesis Doctoral. Universidad Politécnica de Madrid. España. 270 pp.

24. Maimon, O. and Rokach L. (2010). Data Mining and Knowledge Discovery Handbook. Second Edition. Editorial Springer Science & Business Media. 1285 pp.

25. Mangani, Felipe. (2015). Teoría de la Decisión – El árbol de decisión. Universidad de Buenos Aires, Argentina. 25 pp.

26. Martínez, Guillermo. (2015). Metodología de Minería de Datos para el estudio de tablas de siniestralidad vial. Tesis de Maestría. Universidad Complutense, Madrid. 91 pp.

27. McClane, A. J. (1974). McClane´s new estándar fishing encyclopedia and international angling guide. Holt, Rinehart & Winston, NY, 1156 p.

28. Medina, R. F. y Ñique, C. I. (2017). Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python. Interfases 10: 165-189.

29. Mendoza–Rivera, M. S., Martínez–Pardo X., Sierra de la Rosa, J. F., and Suárez, C. A. (2012). Manual para el levante de alevinos de cobia. Bogotá. 61 pp.

30. Meza, L. G. (2015). El paradigma positivista y la concepción dialéctica del conocimiento. *Revista Virtual, Matemática Educación e Internet*.

31. Monod, T. (1973). Rachycentridae. In Hureau, J. C., and Monod, T. (eds.), Checklist of the fishes of the northern-eastern Atlantic and of the Mediterranean, UNESCO, Paris. vol, 1, 371-372 pp.

32. Nyce, Charles (2007), *Predictive Analytics White Paper*, American Institute for Chartered Property Casualty Underwriters/Insurance Institute of America, 15 p.

33. Orallo, H, Ramírez, J. y Ramírez, C. F. (2004). Introducción a la Minería de Datos. Pearson Prentice Hall. 651 pp.

34. Ordoñez, O. (2010). Generalización del algoritmo de boosting binario para más de dos clases. Tesis de grado. Universidad Rey Juan Carlos. 58 pp.

35. Pérez, César. (2007). Minería de datos: técnicas y herramientas. Editorial Paraninfo. Madrid, España. 789 pp.

36. Quinlan, J. R. (1986). Induction of decision trees. Machine Learning 1: 81-106.

37. Ramírez, J. L. (2013). Análisis comparativo DBDT vs otros Algoritmos para el manejo de datos no escalares. Tesis de Máster. Universidad Politécnica de Valencia. 70 pp.

38. Relyea, K. (1981). Inshore fishes of the Arabian Gulf. George Allen & Unwin, London. 149 pp.

39. Richards, C. E. (1967). Age, growth and fecundity of the cobia, Rachycentron canadum, from Cheaspeake bay and adjacent mid-Atlantic waters. *Trans. Am. Fish. Soc. 96*(3), 343-350.

40. Robins, C. R. and Ray, G. C. (1986). A field guide to Atlantic coast fishes of North America. Houghton Mifflin Company, Boston, U.S.A. 354 pp.

41. Rodriguez, J. J., Kuncheva, L. I. y Alonso, C. J. (2006). Rotation forest: A new classifier ensemble method. *Journal IEEE Transactions on Pattern Analysis and Machine Intelligence, 28* (10), 1619-1630.

42. Rokach, Lior and Maimon, Oded. (2005). Top-down induction of decision trees classifiers - a survey. *Journal IEEE Transactions on Systems, Man, and Cybernetic, Part: C 35* (4), 476-487.

43. Rokach, L. and Maimon, O. (2008). Data mining with decision trees: theory and applications. World Scientific. 244 pp.

44. Rokach, L. and Maimon, O. (2014). Data mining with decision trees: theory and applications. 2nd Edition. World Scientific. 328 pp.

45. Russel, S. y Norvig, P. (2004). Inteligencia Artificial – Un Enfoque Moderno. Segunda Edición. Peason Prentice Hall. 1241 pp.

46. Serna, S. (2009). Comparación de Árboles de Regresión y Clasificación y Regresión Logística. Tesis de Maestría. Universidad Nacional de Colombia, Colombia. 60 pp.

47. Smith, J. W. (1995). Life history of cobia, *Rachycentron canadum*, (Osteichthyes: Rachycentridae), in North Carolina waters. Brimleyana. 23:1-23 pp.

48. Sullivan, W. (2017). Machine Learning for Beginners Guide Algorithms: Supervised & Unsupervised Learning Decision Tree & Random Forest Introduction. Editorial Createspace. 166 pp.

49. Swingle, H. A. (1971). Biology of Alabama estuarine areas - Cooperative Gulf of Mexico Estuarine Inventory. Ala. Mar. Resour. Bull. 5, 123 pp.

50. Tarazona, S. (2016). Identification of Main Factors and Variables Describing the Quantity and Distribution of Fatal Vehicular Accidents in Metropolitan City of Lima Using data Mining Techniques: Random Forests, Boosting, Decision Trees. Tesis de grado. Universidad Nacional de Ingeniería. Lima, Perú. 66 pp.

51. Tolosi, L. and Lengauer, T. (2011). Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics 27*, 1986-1994

52. Vaught Shaffer, R. and Nakamura, E. L. (1989). Synopsis of Biological Data on the Cobia *Rachycentron canadum* (Pisces: Rachycentridae). NOAA Technical Report NMFS 82, 21.

53. Zuniga, C. y Abgar, N. (2011). Breve aproximación a la técnica de árboles de decisiones. 11 pp.

9. Anexos

9.1. Anexo 1

Tabla 9. Criterios de importancia de las variables.

cobiaForest$importance

	IncNodePurity
WEEK	13.233117
FA	10.232998
RF	3.741697
MO	22.781625
PEN	30.911560
SGR	6.421241
SA	5.113515
HN	7.807326
DBH	1.591106
DBHa	2.460498
IW	10.160989
TH	2.647917
HWh	35.611282

9.2. Anexo 2

Tabla 10. Modelos realizados con el algoritmo Random Forest.

SEED	%VAR. EXPLAINED	MSR	SSE	RMSE	MAPE	PARÁMETROS			VARIABLES												
						Nodesize	ntree	mtry	PEN	WEEK	MO	SA	FA	HN	DBH	DBHa	IW	TH	RF	SGRh	HWh
1468	69.73	0.191	13.62	0.3568	4.598	-	-	-	*	*	*		*				*			*	*
1468	67.41	0.2057	14.29	0.3654	4.763	-	-	-	*	*	*		*	*			*				*
1468	70.6	0.1855	12.69	0.3443	4.517	-	-	-	*	*	*		*				*				*
1468	69.81	0.1905	13.6	0.3565	4.604	-	-	-	*	*	*		*								*
1468	68.58	0.1983	14.42	0.3671	4.636	-	-	-	*	*	*						*				*
1468	70.16	0.1883	13.1	0.3499	4.614	-	-	-	*		*		*				*				*
1468	71.82	0.1778	11.32	0.3252	4.369	-	-	-	*	*	*		*				*				*
1468	72.7	0.1723	12.57	0.3427	4.484	-	-	-	*	*	*		*				*				*
1468	69.89	0.19	12.25	0.3383	4.394	-	-	-	*	*	*		*				*				
1468	70.72	0.1848	13	0.3485	4.598	-	-	-	*	*			*				*				*
1468	67.18	0.2071	14.13	0.3634	4.718	-	-	-		*	*		*				*				*
1468	68.92	0.1961	13.93	0.3609	4.678	-	-	-		*	*		*				*			*	*
1468	71.67	0.1788	13.55	0.3558	4.589	-	-	-	*	*			*				*			*	*
1468	71.08	0.1825	13.32	0.3528	4.612	-	-	-	*		*		*				*			*	*
1468	70.79	0.1843	14.76	0.3714	4.656	-	-	-	*	*	*						*			*	*
1468	70.57	0.1857	14.52	0.3684	4.665	-	-	-	*	*	*		*							*	*
1468	72.02	0.1766	11.73	0.3312	4.394	-	-	-	*	*	*		*				*			*	*
1468	73.59	0.1667	12.45	0.3412	4.437	-	-	-	*	*	*		*				*			*	*
1468	70.57	0.1857	11.79	0.3319	4.336	-	-	-	*	*	*		*				*			*	
1468	69.96	0.1895	11.46	0.3273	4.333	-	-	-	*	*	*		*				*				
1468	70.3	0.1874	12.31	0.3392	4.454	-	-	-	*	*	*		*								*

SEED	%VAR. EXPLAINED	MSR	SSE	RMSE	MAPE	PARÁMETROS			VARIABLES												
						Nodesize	ntree	mtry	PEN	WEEK	MO	SA	FA	HN	DBH	DBHa	IW	TH	RF	SGRh	HWh
1468	72.56	0.1731	12.02	0.3351	4.399	-	-	-	*	*	*		*							*	*
1468	71.32	0.181	11.08	0.3218	4.394	-	-	-	*	*			*				*				*
1468	75.47	0.1548	10.16	0.3081	4.214	-	-	-	*	*	*		*				*				*
1468	72.01	0.1766	9.774	0.3022	4.041	-	-	-	*	*	*		*				*				
1468	75.49	0.1547	9.705	0.3012	4.17	-	-	-	*	*			*				*				*
1468	71.56	0.1795	10.23	0.3093	4.271	-	-	-	*	*			*				*				
1468	71.87	0.1775	11.27	0.3245	4.425	-	-	-	*		*		*				*				*
1468	76.83	0.1462	9.591	0.2994	3.958	-	-	-	*	*	*		*				*				
1468	67.68	0.204	12.39	0.3402	4.49	-	-	-	*	*	*						*				
1468	71.31	0.1811	11.38	0.3261	4.387	-	-	-	*	*	*		*								
1468	71.32	0.181	10.23	0.3091	4.159	-	-	-	*		*		*				*				
1468	71.25	0.1815	12.7	0.3445	4.576	-	-	-		*	*		*				*				
1468	66.84	0.2093	16.09	0.3878	4.795	-	-	-	*	*	*										*
1468	70.47	0.1864	15.04	0.3749	4.671	-	-	-	*	*	*										*
1468	66	0.2146	17.91	0.4091	5.001	-	-	-	*		*										*
1468	69	0.1956	15.87	0.3852	4.803	-	-	-	*		*										*
1468	68.07	0.2015	14.45	0.3675	4.758	-	-	-	*	*	*		*	*			*			*	*
1468	67.33	0.2062	14.61	0.3695	4.828	-	-	-	*	*	*		*	*			*				*
1468	69.49	0.1925	14.05	0.3624	4.617	-	-	-	*	*	*		*				*			*	*
1468	73.19	0.1692	12.51	0.3419	4.485	-	-	-	*	*	*		*				*			*	*
1468	75.64	0.1537	11.74	0.3313	4.382	-	-	-	*	*	*		*							*	*
1468	75.76	0.153	10.41	0.312	4.241	-	-	-	*	*	*		*				*			*	*
1468	77.4	0.1426	9.561	0.2989	3.945	-	-	-	*	*	*		*				*			*	
1468	76.05	0.1511	11.08	0.3217	4.295	-	-	-	*	*	*		*							*	
1468	76.03	0.1513	10.35	0.311	4.053	-	-	-	*	*	*		*								
1468	75.26	0.1561	11.88	0.3333	4.355	-	-	-	*	*	*						*			*	
1468	74.55	0.1606	10.25	0.3096	4.184	-	-	-	*	*	*						*				

SEED	%VAR. EXPLAINED	MSR	SSE	RMSE	MAPE	PARÁMETROS			VARIABLES												
						Nodesize	ntree	mtry	PEN	WEEK	MO	SA	FA	HN	DBH	DBHa	IW	TH	RF	SGRh	HWh
1468	75.97	0.1516	8.726	0.2856	3.847	-	-	-	*	*	*	*					*				
1468	68.57	0.1984	13.68	0.3575	4.666	-	-	-	*	*	*						*	*			
1468	71.07	0.1826	13.76	0.3586	4.888	-	-	-	*	*	*				*		*				
1468	75.8	0.1527	11.49	0.3277	4.243	-	-	-	*	*	*					*	*				
1468	71.59	0.1793	12.09	0.3362	4.379	-	-	-	*	*	*						*		*		
1468	77.41	0.1425	9.239	0.2938	3.754	-	-	-	*	*	*	*				*	*				
1468	77.39	0.1427	9.419	0.2967	3.794	-	-	-	*	*	*	*	*				*				
1468	76.83	0.1462	9.591	0.2994	3.958	-	-	-	*	*	*		*				*				
1468	77.87	0.1396	9.866	0.3037	3.846	-	-	-	*	*	*		*			*	*				
1468	78.23	0.1374	9.498	0.2979	3.734	-	-	-	*	*	*	*	*			*	*				
1468	75.69	0.1534	9.331	0.2953	3.968	-	-	-	*	*	*	*	*			*	*				*
1468	76.88	0.1459	9.214	0.2934	4.006	-	-	-	*	*	*	*	*				*				*
1468	76.78	0.1465	9.427	0.2968	4.063	-	-	-	*	*	*	*	*				*		*		*
1468	77.17	0.1441	9.258	0.2942	3.83	-	-	-	*	*	*	*	*				*			*	
1468	77.41	0.1425	9.239	0.2938	3.754	-	-	-	*	*	*	*				*	*				
1468	77.8	0.1401	9.487	0.2978	3.821	-	-	-	*	*	*	*					*		*		
1468	76.59	0.1477	9.658	0.3004	3.891	-	-	-	*	*	*	*					*	*			
1468	76.28	0.1497	8.653	0.2844	3.774	-	-	-	*	*	*	*									
1468	74.14	0.1632	8.651	0.2843	3.714	-	-	-	*		*	*					*				
1468	75.73	0.1532	7.416	0.2633	3.593	-	-	-	*	*		*					*				
1468	75.92	0.152	8.75	0.286	3.829	-	-	-	*	*			*				*				
1468	76.94	0.1455	8.101	0.2752	3.655	-	-	-	*	*		*				*	*				
1468	76.25	0.1499	7.972	0.273	3.722	-	-	-	*	*		*					*			*	
1468	77.95	0.1392	8.912	0.8861	3.751	-	-	-	*	*		*				*	*			*	
1468	76.95	0.1455	8.34	0.2792	3.689	-	-	-	*	*		*	*				*				