ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

“MODELO DE APRENDIZAJE SUPERVISADO PARA LA PROYECCIÓN DEL VOLUMEN DE COSECHA DE CAMARÓN EN EMPRESA CAMARONERA DEL CANTÓN DURÁN”

TRABAJO DE TITULACIÓN

Previo a la obtención del título de:

Magíster en Ciencias de Datos

Presentado por:

YAMBAY PINARGOTE CARLOS ANDRES

GUAYAQUIL – ECUADOR

AÑO: 2023

DEDICATORIA

A Dios por ser motor principal de vida y fuente de todas las ciencias.

Carlos Yambay P.

AGRADECIMIENTO

A mi familia por el esfuerzo, sacrificio y apoyo que me han sabido brindado a lo largo de mi vida.

A mis padres por inculcarme el valor del sacrificio, resiliencia y perseverancia, los cuales me han permitido conseguir mis objetivos.

Carlos Yambay P.

TRIBUNAL DE SUSTENTACIÓN

Dr. Christian Galarza Morales	**Dr. Sergio Baúz**
PROFESOR TUTOR	PROFESOR EVALUADOR

DECLARACIÓN EXPRESA

“Los derechos de titularidad y explotación, nos corresponde conforme al reglamento de propiedad intelectual de la institución; Carlos Andrés Yambay Pinargote doy el consentimiento para que la ESPOL realice la comunicación pública de la obra por cualquier medio con el fin de promover la consulta, difusión y uso público de la producción intelectual"

Yambay Pinargote Carlos Andrés

RESUMEN

Para el cumplimiento de las órdenes de compra la empresa debe de adquirir materia prima (camarón) a proveedores terceros ya que la producción de las piscinas propias no abastece para el cumplimiento de las órdenes. Por lo cual, el presente proyecto busca brindar una herramienta que le permita a la empresa poder conocer con anticipación el volumen de cosecha de camarón que tendrá en una corrida de siembra de las piscinas propias para de esta forma estimar el volumen de camarón que deben de comprar a proveedores terceros para satisfacer las ordenes de producción en curso.

Para el presente proyecto se analizaron un total de 455 registro que comprenden a los ciclos de cosecha de los últimos 13 años de las diversas piscinas pertenecientes a la empresa objeto de estudio, con estos datos se realizó una evaluación de diversos modelos de machine learning según los criterios de evaluación MAE, MSE, RMSE, R^2 , teniendo como modelos de mejor rendimiento a Light Gradient Boosting con un R^2 0.839, *MAPE* 0.30, a Random Forest Regressor con un R^2 0.842, *MAPE* 0.25 y a Extra Trees Regressor con un R^2 0.854, *MAPE* 0.22.

Palabras claves: Cosecha, Camarón, Aprendizaje supervisado.

ABSTRACT

In order to fulfill the purchase orders, the company must purchase raw material (shrimp) from third party suppliers, since the production of the company's own ponds does not supply enough to fulfill the orders. Therefore, this project seeks to provide a tool that allows the company to know in advance the volume of shrimp harvest it will have in a planting run of its own ponds in order to estimate the volume of shrimp that must be purchased from third party suppliers to satisfy the current production orders.

For the present project, a total of 455 records were analyzed, comprising the harvest cycles of the last 13 years of the various pools belonging to the company under study. With this data, an evaluation of various machine learning models was carried out according to the evaluation criteria MAE, MSE, RMSE, R^2 , with the best performing models being Light Gradient Boosting with an R^2 0. 839, *MAPE* 0.30, Random Forest Regressor with an R^2 0.842, *MAPE* 0.25 and Extra Trees Regressor with an R^2 0.854, *MAPE* 0.22.

Keywords: Harvest, Shrimp, Supervised learning.

Contenido

DEDICATORIA II
AGRADECIMIENTO III
TRIBUNAL DE SUSTENTACIÓN IV
DECLARACIÓN EXPRESA V
RESUMEN I
ABSTRACT II
ÍNDICE DE FIGURAS V
ÍNDICE DE TABLAS VI
CAPÍTULO 1 1
1. PLANTEAMIENTO DEL PROBLEMA 1
1.1. Descripción del Problema 1
1.2. Justificación 2
1.3. Objetivos (General y Específico) 3
1.4. Metodología 3
1.5. Resultados Esperados 5
1.6. Dataset 5
CAPÍTULO 2 6
2. ESTADO DEL ARTE 6
2.1 Marco referencial 6
2.2 Machine Learning 8
2.2.1 Tipos de aprendizajes en Machine Learning 8
2.2.1.1 Aprendizaje supervisado 8
2.2.1.2 Aprendizaje no supervisado 10
CAPÍTULO 3 11
3 DISEÑO E IMPLEMENTACIÓN 11
3.1 Exploración y validación de datos 11
3.1.1 Análisis de características 14
3.1.2 Ingeniería de variables 17
3.2 Prototipo de algoritmo y modelos 19
3.3 Plataforma y prototipo 20
3.4 Métricas y comunicación de resultados 21
CAPÍTULO 4 22
4 ANÁLISIS DE RESULTADOS 22
4.1 Evaluación de métricas para la selección de modelos 22
4.2 Evaluación de la capacidad predictiva en test de los modelos seleccionados 24

4.3 Evaluación del poder predictivo 26
CONCLUSIONES Y RECOMENDACIONES 27
CONCLUSIONES 27
RECOMENDACIONES 28
REFERENCIAS BIBLIOGRÁFICAS 29
ANEXOS 31
5 Anexos 31

ÍNDICE DE FIGURAS

Figura 2.1 Interacción de un aprendizaje supervisado 9
Figura 2.2 Interacción de un aprendizaje no supervisado 10
Figura 3.1 Correlación de variables con libras cosechadas 16

ÍNDICE DE TABLAS

Tabla 1.1 Variables del modelo 5
Tabla 3.1 Número de entidades a usar en el modelo 11
Tabla 3.2 Variables excluidas 14
Tabla 3.3 Variables Iniciales para exploración 15
Tabla 3.4 Variables agregadas al dataset 17
Tabla 3.5 Tipos de datos final de las variables 18
Tabla 3.6 Principales modelos de regresión tradicionales 19
Tabla 3.7 Variables a ingresar en el sistema de cálculo del volumen de cosecha 20
Tabla 3.8 Indicadores para evaluación del modelo 21
Tabla 4.1 Evaluación de modelos 22
Tabla 4.2 Pipeline Modelos seleccionados 23
Tabla 4.3 Valores de predicción de modelos evaluados 24
Tabla 4.4 Predicciones Reales 26

CAPÍTULO 1

1. PLANTEAMIENTO DEL PROBLEMA

1.1. Descripción del Problema

Durante el primer trimestre del año 2022, Ecuador se convirtió en el productor de camarón número uno a escala mundial de con exportaciones que superan las 1600 libras exportadas según datos proporcionados por el Ministerio de Producción.

Por lo anterior, las exportadoras de camarón deben de realizar tareas y esfuerzos extraordinarios para dar cumplimiento a sus órdenes de producción que nacen de acuerdos comerciales con sus clientes, debido a la necesidad de altos volúmenes de producto (camarón) requerido para satisfacer las ordenes de producción en curso, se hace uso de cosecha de camarón propia de la exportadora originadas desde sus piscinas, sin embargo, este porcentaje no logra satisfacer el 100% del producto requerido para satisfacer la orden, por lo cual, el porcentaje faltante es adquirida mediante acuerdos comerciales con camaroneras terceras y pequeños productores.

En el sector camaronero la compra de camarones a productores terceros es un proceso demandante y competitivo, ya que existe una gran cantidad de plantas procesadoras que requieren de materia prima para satisfacer su producción. Por lo cual, una inadecuada planificación de compras de camarones a terceros puede tener un impacto económico considerable en la organización ya que no se cubren las necesidades de las ordenes de producción y producen retrasos, lo cual se traduce en daños de productos, multas aduaneras, costos adicionales de producción, costos

logísticos e impacto en volúmenes de ventas sin mencionar el daño de imagen corporativa.

Por lo anterior, se requiere como medida preventiva a este evento tener una claridad de los volúmenes de cosecha propia de camarón que le permita realizar una adecuada planificación y pactar acuerdos comerciales previos para la adquisición de compras a terceros y de esta forma evitar los impactos económicos consecuentes.

En la actualidad, la organización no cuenta con un sistema para estimar sus volúmenes de cosecha propia que le permitan tener este nivel de planificación de comprar a terceros, por lo cual, se ve afectada en sus procesos productivos teniendo un impacto en el cumplimiento de sus acuerdos comerciales y de órdenes de producción en curso.

1.2. Justificación

El presente proyecto busca desarrollar un modelo para predecir el volumen de cosecha de camarón que le permita a la empresa tener una aproximación de la producción de materia prima (camarón) de sus propias piscinas y de esta forma poder planificar con antelación la compra de camarón faltante a camaroneras terceras que le permita satisfacer los volúmenes de producto ordenes de producción en curso.

Esta planificación le permitirá a la empresa tener una reducción considerable de costos y gastos operativos generados debido a una inadecuada planificación y las pérdidas económicas que nacen a partir de la falta de cumplimiento de los requerimientos de productos de sus clientes.

1.3. Objetivos (General y Específico)

1.3.1 Objetivo general

Diseñar un modelo de aprendizaje supervisado para la proyección del volumen de cosecha de camarón

1.3.2 Objetivo específico

- Extraer variables y características meteorológicas, demográficas y productivas de las fases de siembra
- Identificar características significativas para la predicción de la cosecha de camarón.
- Desarrollar un modelo predictivo haciendo uso de técnicas de Machine Learning para predecir el volumen de cosecha de camarón.

1.4. Metodología

En el presente proyecto se hace uso de técnicas supervisadas de Estadística, Machine Learning para la evaluación de parámetros y selección de modelos predictivos que se ajusten a la naturaleza de nuestro dataset con las variables previamente evaluadas y seleccionadas, para de esta forma obtener resultados adecuados que satisfagan la necesidad de la investigación.

La primera fase del proyecto consiste en hacer un análisis de la literatura actual referente a problemas similares que sirvan de base para seleccionar la tecnología, metodología y prácticas modernas que permitan dar solución a la problemática inicialmente planteada.

La segunda fase consiste en la recolectar los datos históricos de las corridas y siembras de camarones que se han dado dentro de la exportadora en los últimos 2 años para obtener patrones de comportamiento del proceso productivo, además, permite la evaluación de las diferentes características meteorológicas, de suelo, calidad de agua, alimentación y otras variables productivas para crear el dataset que será usado en la presente investigación.

La tercera fase consiste en el tratamiento y análisis de los datos obtenidos mediante la limpieza de datos eliminando los datos nulos, faltantes, outliers, además de normalizar los datos y analizar las características obtenidas mediante la evaluación de la correlación de las diferentes variables y su incidencia en la predicción de nuestra variable dependiente haciendo uso de análisis descriptivo.

La cuarta fase consiste en evaluar modelos supervisados los cuales serán entrenados mediante el uso de datos históricos de ciclos de siembras y volúmenes de cosechas de los 2 años anteriores. Dentro de los modelos a ajustar se encuentran modelos clásicos de pronóstico como el análisis de series temporales, además de modelos de vanguardia como lo son las redes neuronales recurrentes.

En particular, los algoritmos a evaluar son los modelos ARIMA usando criterios de evaluación como la raíz del Error Cuadrático Medio y otros más robustos como el Error Porcentual Absoluto Medio, estos son el RMSE y MAPE, respectivamente, por sus siglas en inglés.

1.5. Resultados Esperados

El presente proyecto tiene como finalidad brindar una predicción del volumen de cosecha de camarón de la piscina de producción propia con un accuracy superior al 85% en la predicción.

1.6. Dataset

Para la realización del presente proyecto se considerarán los datos de los volúmenes de producción históricos de las piscinas 16 piscinas pertenecientes a la camaronera objeto de estudio, para lo cual se hace uso de los datos de cosecha de los periodos comprendidos entre el 2008-2021 distribuidas en un total de 455 registros con las variables expuestas en la tabla 1.1.

Tabla 1.1 Variables del modelo

N	Variable	Descripción
1	**CodPiscina**	Corresponde al número de la piscina
2	**has**	Corresponde a la dimensión de la piscina dada en hectáreas
3	**Ciclo**	Es la secuencia del número de cosechas realizadas
4	**FechaInicio**	Fecha de siembra de las larvas en las piscinas
5	**CodTransferencia**	Tipo de transferencia realizada en el cultivo de las larvas en la piscina
6	**CodProcedencia**	Código identificador proveedor de la larva
7	**CodCaliLarva**	clasificación de la calidad de la larva
8	**Peso_siembra**	Peso promedio del total de larvas sembradas
9	**Cantidad_siembra**	Total de libras sembradas
10	**FechaCosecha**	Fecha de la cosecha
11	**LibrasCos**	Total de libras cosechadas al final del ciclo.

Fuente: elaboración propia

CAPÍTULO 2

2. ESTADO DEL ARTE

2.1 Marco referencial

En la actualidad, existe un número significativo de estudios relacionados al análisis predictivo en los procesos productivos. Por ejemplo, en relación a las investigaciones más recientes sobre las predicciones en procesos de cosecha de camarón haciendo uso de modelos de Deep learning encontramos la investigación titulada “Predicción del nivel de cosecha de camarón blanco: el caso de una pequeña camaronera en la parroquia Tenguel del cantón Guayaquil, Ecuador” (Cevallos Valdiviezo, Rodríguez Cristiansen, Valdiviezo Valenzuela, Arévalo Avecillas, & Padilla Lozano, 2020) podemos observar el uso de técnicas para modelos supervisados y la evaluación de diferentes modelos clásicos y de machine learning como lo son Regresión Lineal Múltiple (RLM) por mínimos cuadrados, Árbol de Regresión CART, Bosques Aleatorios, Splines de regresión adaptativa multivariante (MARS) y Máquinas de Soporte Vectorial (SVM) los cuales fueron evaluados bajo Validación Cruzada identificando como mejores modelo predictor según la selección previa de variables a el modelo Multivariate Adaptive Regression Splines (MARS) sin iteracciones, tanto en las validaciones cruzadas como en las predicciones realizadas en los dos siguientes ciclos.

Por otra parte, en la investigación titulada “Application of Artificial Neural Networks to forecast Litopenaeus vannamei and Penaeus monodon harvests in Indramayu Regency, Indonesia” (Pamungka, y otros, 2020) se hace uso de técnicas de Deep Learning para desarrollar un modelo con el fin de pronosticar la producción total de camarón para Litopenaeus vannamei y Penaeus Monodon en una región de

Indonesia utilizando redes neuronales de retro- propagación (BPNN) haciendo uso de los criterios de evaluación RMSE y MAPE para la selección del modelo. Sus resultados apuntan a un buen rendimiento en el pronóstico en fase de entrenamiento, sin embargo, según lo indicado en la investigación la calidad, integridad y falta de datos obtenidos no permitieron tener un óptimo rendimiento durante su fase de producción, obteniendo resultados deficientes en sus pronósticos reales.

Por lo anterior, se entiende que la calidad e integridad de los datos recolectados son un pilar fundamental a la hora de construir el modelo predictivo ya que por su naturaleza supervisada requiere de datos históricos previamente etiquetados para identificar patrones y correlaciones para la selección adecuada de las variables y así obtener un desempeño óptimo en las predicciones del modelo e las fases de producción.

Además, se puede inferir que los resultados durante la fase de entrenamiento y testing fueron adecuados debido a la poca cantidad de datos; según indica la investigación, esto pudo ocasionar que el modelo se sobre ajuste a los datos de entrenamiento y no aprendió a generalizar los patrones.

De las anteriores investigaciones citadas se puede concluir que el uso de técnicas vanguardistas puede aportar significativamente a la mejora de los procesos productivos mediante el uso de un modelo predictivo optimo.

2.2 Machine Learning

Según (Valdez Pino, 2021) indica que “El machine learning es disciplina del ámbito de la Inteligencia Artificial que tiene por finalidad crear sistemas que puedas aprender automáticamente. Con esto se tiene que la máquina va aprendiendo de algunos ejemplos de cómo hacer o no una tarea”. (pág. 40)

Por lo anterior, el machine learning es interdisciplinario debido a que realiza la abstracción de conceptos y técnicas de otras disciplinas como las matemáticas, física y estadística para desarrollar modelos que permitan dar soluciones a problemas de clasificación y predicción de variables sujetas a estudio mediante el uso del poder computacional.

El Machine Learning se clasifican principalmente en tres tipos según su forma de aprendizaje, estos pueden ser:

1) Aprendizaje supervisado;
2) Aprendizaje no supervisado;
3) Aprendizaje por reforzamiento.

2.2.1 Tipos de aprendizajes en Machine Learning

2.2.1.1 Aprendizaje supervisado

Las técnicas de aprendizajes supervisados hacen referencia a métodos que usan datos previamente etiquetados para ser usados como datos de entradas y

mediante la aplicación del procesamiento computacional extraer características y reconocer de patrones que permitan entender el fenómeno a estudiar.

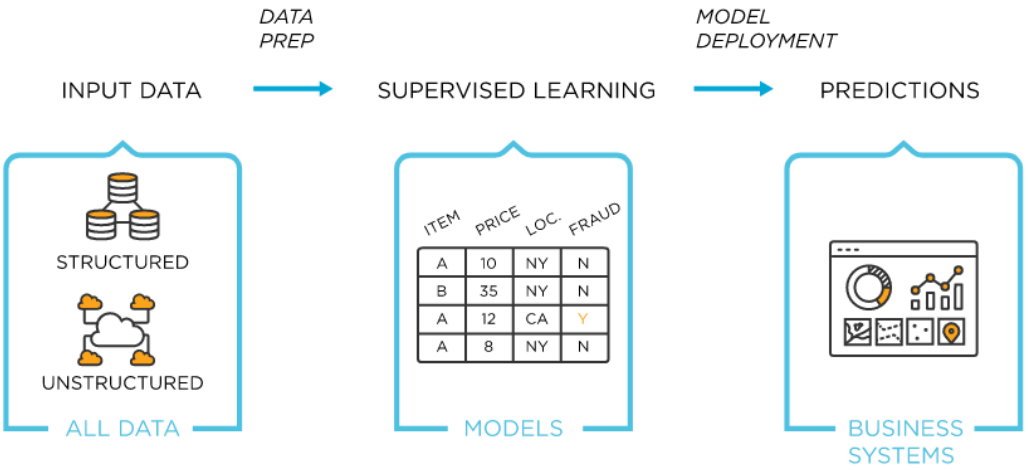


Figura 2.1 Interacción de un aprendizaje supervisado

Fuente: (TIBCO Software, s.f)

Los principales usos de las técnicas de aprendizaje supervisado es dar solución a problemas de:

1. Clasificación: división de objetos por clases;
2. Regresión: predicción de valores.

Mientras que los principales algoritmos de aprendizaje supervisados son:

1. Árboles de decisión;
1. Clasificación de Naïve Bayes;
2. Regresión por mínimos cuadrados;
3. Regresión Logística;
4. Support Vector Machines (SVM);
5. Métodos “Ensemble” (Conjuntos de clasificadores).

2.2.1.2 Aprendizaje no supervisado

Las técnicas de aprendizaje no supervisado hacen uso de datos no etiquetados que usan métodos matemáticos y estadísticos que permite reconocer patrones de forma independiente para crear las etiquetas de los datos a partir de tendencias comprendidas.

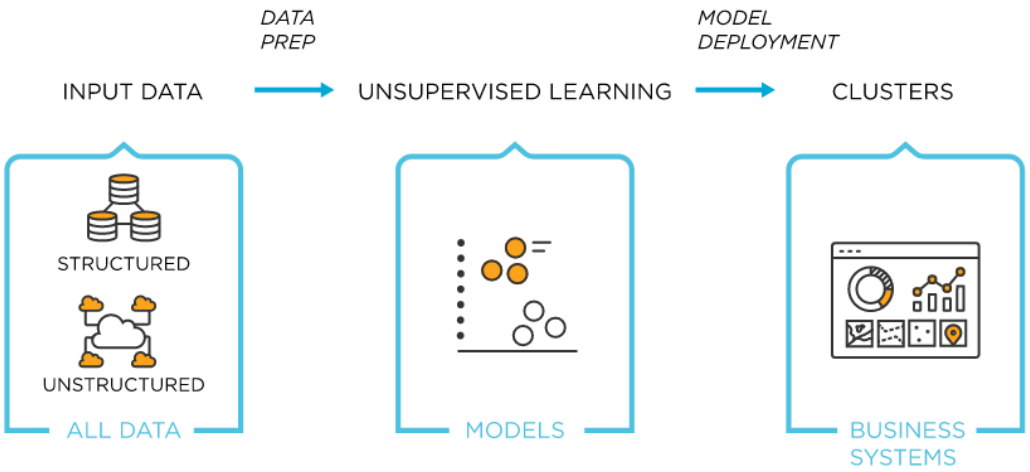


Figura 2.2 Interacción de un aprendizaje no supervisado

Fuente: (TIBCO Software, s.f)

Los principales usos de las técnicas de aprendizaje no supervisado es dar solución a problemas de:

1. Clustering: Clasificación de grupos por características detectadas;
2. Asociación: Relaciones entre grupos por coincidencias entre conjuntos.

Los principales algoritmos de aprendizaje no supervisados son

1. PCA (Análisis de Componentes Principales);
2. K-Modes;
3. K-Means.

CAPÍTULO 3

3 DISEÑO E IMPLEMENTACIÓN

3.1 Exploración y validación de datos

Para el presente proyecto inicialmente se obtuvo el registro de datos recolectados de los ciclos de cosecha de los últimos 28 años comprendidos desde 1994 hasta el año 2021.

Sin embargo, debido a que el registro de los datos fue realizado en su mayoría de forma manual, estando sujetos a errores, se decidió trabajar únicamente con los datos provenientes de los mecanismos IOT como lo son sensores de temperaturas y, alimentadores automáticos, los que datan de los últimos 13 años; así, se obtiene un total de 455 registros.

Tabla 3.1 Número de entidades a usar en el modelo

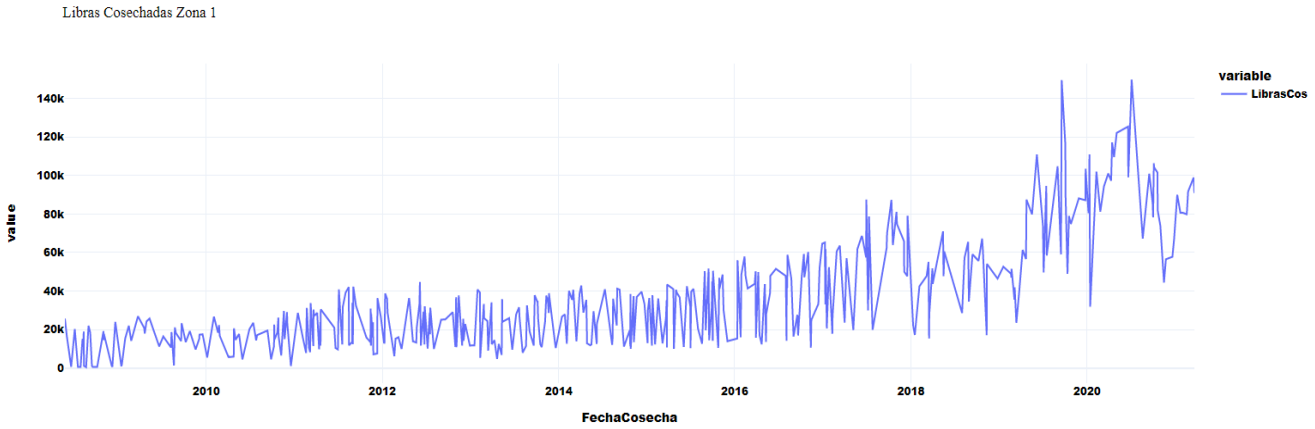
Entidad	Número
Años	13
Piscinas	16
Registros	455
Características	12

Fuente: Elaboración propia

Conforme la Tabla 3.1, los datos obtenidos son provenientes de piscinas ubicadas en el mismo sector geográfico y compartiendo características similares. El

número de registros es 455, originados de los periodos de cosecha de 16 piscinas pertenecientes a la Zona 1 de la camaronera observada en el presente estudio.

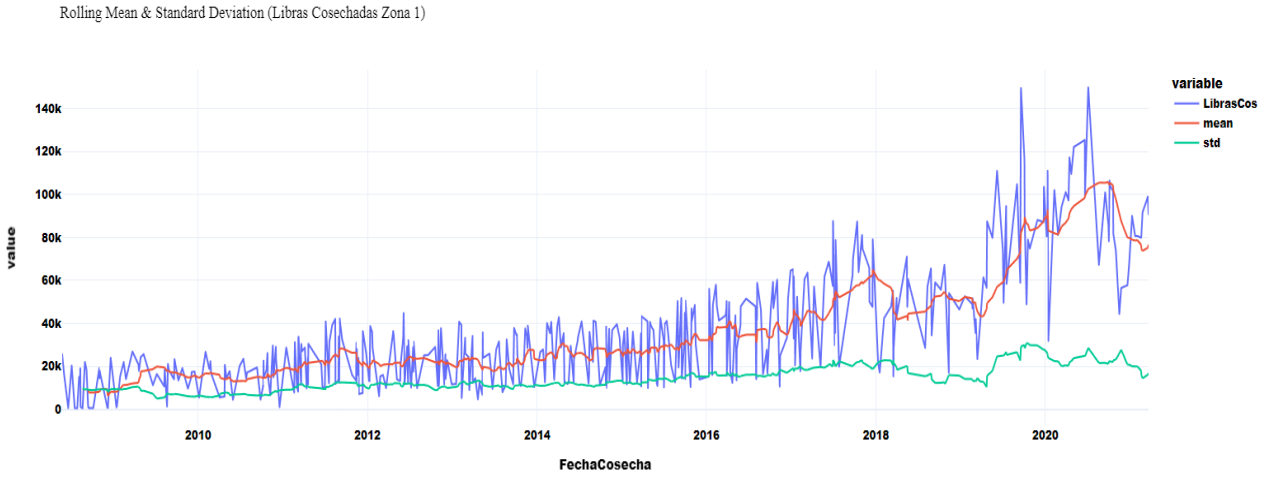
Gráfico 3.1 Libras cosechadas en Zona 1



Fuente: Elaboración propia

Como podemos observar en la grafico 3.1 los volúmenes de producción han tenido un incremento significativo durante los últimos años a tal punto de duplicar la producción.

Gráfico 3.2 Rolling Mean & Standard Deviation (Libras Cosechadas Zona 1)



Fuente: Elaboración propia

En el gráfico 3.2 podemos observar la representación gráfica del total del volumen de producción a través del tiempo representada en la variable a predecir denominada Libras cosechadas, junto con su promedio de cosecha y desviación estándar Según se observa en el gráfico la serie no es estacionaria, ya que la desviación y la media están sujetas a la variación en el tiempo.

Para la confirmación de lo anterior se procede a realizar la una prueba Dickey--Fuller aumentada, que indica que si el valor p <0.05 la serie es estacionaria lo que significa que su media y varianza no cambian en el transcurso del tiempo lo que impacta de forma significativa en la predicción de la variable Y (libras cosechadas).

```
Augmented Dickey-Fuller Test:
ADF test statistic       -1.091160
p-value                   0.718584
# lags used               7.000000
# observations          447.000000
critical value (1%)      -3.445064
critical value (5%)      -2.868028
critical value (10%)     -2.570226
Weak evidence against the null hypothesis
Fail to reject the null hypothesis
Data has a unit root and is non-stationary
```

Ilustración 3.1 Prueba Dickey-Fuller

Fuente: Elaboración propia

El resultado obtenido en la ilustración 3.1 se observa que el valor p no es menor que 0.05, por lo que no se rechaza la hipótesis nula que indica que la serie tiene una raíz única, por lo cual se infiere que el comportamiento de las libras cosechadas a través del tiempo tiene una naturaleza no estacionaria.

3.1.1 Análisis de características

Durante la recolección de los datos se obtuvo una cantidad considerable de variables pertenecientes a todos los datos del proceso productivos, de las cuales después de un primer análisis algunas fueron descartadas debido a su naturaleza de variables de seguimiento que no las tendremos desde un estado inicial. Esto es debido a que son variables que se van obteniendo a medida que avanza el ciclo de la cosecha para las cuales se debía de establecer supuestos que no entran en el contexto de la actual investigación.

Tabla 3.2 Variables excluidas

Variables	Motivos de exclusión
Total de alimento consumido	El total de alimentos consumidos es una variable de seguimiento el cual se encuentra establecido el alimento a entregar según el porcentaje de camarones vivos en el muestreo de supervivencia.
Oxígeno disuelto promedio del estanque	La cantidad de oxígeno es dependiente del porcentaje que supervivencia según el muestreo
Supervivencia estimada	Es una estimación definida en una tabla entregada por el proveedor de alimento

Fuente: Elaboración propia

Por lo anterior, las variables iniciales tomadas en consideración para el análisis exploratorio son las mostradas en la tabla 3.3. Los tipos de variables también se presentan en la misma.

Tabla 3.3 Variables Iniciales para exploración

Variables	Tipo de dato
Piscina (CodPiscina)	Categórica
Hectáreas (has)	Numérica
Ciclo	Numérica
FechaInicio	Numérica
CodTransferencia	Numérica
CodProcedencia	Numérica
Calidad de larva (CodCaliLarva)	Numérica
Peso_siembra	Numérica
Cantidad_siembra	Numérica
FechaCosecha	Numérica
Libras cosechadas (LibrasCos)	Numérica

Fuente: Elaboración propia

Como podemos observar en la tabla 3.3 tenemos inicialmente un total de 13 variables de las cuales podemos dividir en 1 variable categórica que es la que divide por piscina y 7 variables numéricas, estas variables pasaran por un proceso de análisis y asignación de tipo de variable según sea la necesidad de la presente investigación y la naturaleza de los datos

Una vez seleccionadas las variables a trabajar se procede con el análisis descriptivo para ir identificando asociaciones, reconociendo patrones e inferir el grado de importancia dentro de los modelos.

Se seleccionan las variables predictoras (X) representativas que expliquen de mejor manera la variable a predecir en este caso libras Cosechadas (Y). Esa selección se la realiza haciendo uso del criterio de selección de la correlación de Pearson donde $\rho > 0.5$ se considera una correlación moderada o alta, caso contrario diremos que la variable tiene poco peso al momento de predecir los valores de (Y) usando una estructura lineal.

LibrasCos	1.000000
Ciclo	0.799856
Anio_Siembra	0.758263
Anio_Cosecha	0.751595
Cantidad_siembra	0.735238
has	0.596415
Peso_siembra	0.443166
PRECIPITACION	0.052343
TEMPERATURA	0.018642
Dia_Siembra	0.008708
Dia_Cosecha	0.006599
Mes_Siembra	-0.018736
MES	-0.018736
Mes_Cosecha	-0.024686
Numero_Dias_Cosecha	-0.191556
CodProcedencia	-0.502738

Figura 3.1 Correlación de variables con libras cosechadas

Fuente: Elaboración propia

Según lo que podemos observar en la figura 3.1 Las variables que más presentan un comportamiento lineal con las libras cosechadas son las variables Ciclo, Año de siembra, Año de cosecha, cantidad sembrada, hectáreas.

3.1.2 Ingeniería de variables

El adecuado tratamiento de las variables es una parte fundamental para la construcción del modelo que permita un óptimo funcionamiento, por lo cual se procede con la creación y adaptación de nuevas variables mediante el análisis de los datos que nos permitan generalizar la compresión del modelo y mejorar su desempeño.

Por lo anterior, se procede con la creación de variables que pueden resultar significativas para el modelo, sea porque resumen en mejor forma las variables originales, o porque pueden ser obtenidas sin necesidad de ser observadas. Estas se presentan en la tabla 3.4.

Tabla 3.4 Variables agregadas al dataset

Variables iniciales	Nueva variable
Fecha de siembra	Mes de siembra
	Año de siembra
Fecha de siembra – Fecha de cosecha	Días de siembra
Cantidad sembrada / hectáreas	Ratio de siembra
Fuente externa	Temperatura promedio Mensual
Fuente externa	Precipitación promedio mensual

Fuente: Elaboración propia

Una vez realizada la inserción de las variables y la creación de Features, se reasigna el tipo de datos que nos permitan una correcta predicción de los volúmenes de cosecha de producción.

Tabla 3.5 Tipos de datos final de las variables

Variable	Tipo de datos
CodPiscina	Categórica
Hectárea (has)	Numérica
Anio_Siembra	Numérica
Mes_Siembra	Categórica
Cantidad_siembra	Numérica
Ratio_siembra	Numérica
Anio_Cosecha	Numérica
Mes_Cosecha	Categórica
Numero_Dias_Cosecha	Numérica
Temperatura	Numérica
Precipitacion	Numérica
LibrasCos	Target

Fuente: Elaboración propia

Como se puede observar en la tabla 3.5 los tipos de datos con los que se trabajará son 3 categóricas, 7 numéricas, 1 a predecir dando un total de 11 variables que ingresaran a los modelos para la predicción del volumen de cosecha de camarón.

Las variables categóricas pasaran por una transformación a variables dummies debido a su naturaleza categórica, esto creará una columna adicional en el dataset por cada categoría de la variable.

3.2 Prototipo de algoritmo y modelos

En la presente propuesta se plantea comparar diferentes modelos de aprendizaje supervisado para la selección del mejor modelo, este es el que tenga el mejor rendimiento en términos de una métrica de interés, en nuestro caso, que permita la predicción del volumen de cosecha de camarón en la zona 1 de la camaronera con la mayor precisión.

Tabla 3.6 Principales modelos de regresión tradicionales

	Modelo
Arima	Arima
ightgbm	Light Gradient Boosting Machine
rf	Random Forest Regressor
et	Extra Trees Regressor
gbr	Gradient Boosting Regressor
ada	AdaBoost Regressor
en	Elastic Net
ridge	Ridge Regression
lr	Linear Regression
lasso	Lasso Regression
dt	Decision Tree Regressor
lar	Least Angle Regression

Fuente: Elaboración propia

En la tabla 3.6 podemos observar los modelos que serán evaluados basados en los criterios de evaluación anteriormente descrito. Dentro de los modelos clásicos estadísticos, usaremos Random Forest Regressor, Extra Trees Regressor, Light Gradient Boosting Machine, además de un modelo, ARIMA con variables exógenas, esto es, un modelo de regresión lineal múltiple cuyo término de error tenga una estructura

de autocorrelación y que evaluar (de ser necesario) la significancia de las variables predictoras o independientes a través del tiempo.

3.3 Plataforma y prototipo

El producto resultante de la presente investigación será un sistema de cálculo que permita estimar el volumen de cosecha de camarón mediante el ingreso de los parámetros expuestos en la siguiente tabla.

Tabla 3.7 Variables a ingresar en el sistema de cálculo del volumen de cosecha

Código de la piscina
Fecha de siembra
Cantidad sembrada
Fecha de cosecha o número de días del ciclo de cosecha

Fuente: Elaboración propia

Con el ingreso de las variables indicadas en la tabla 3.6 el sistema internamente calculará la distancia entre la fecha de inicio de siembra y la fecha de cosecha dando como resultado el número de días del ciclo de cosecha; dividirá las variables fechas en unidades de día, mes y, año. Finalmente, se agregarán variables atmosféricas como temperatura y precipitación media únicamente conociendo el mes de siembra.

Por lo anterior, las variables que se agregarían al dataset inicial serían: 1) año de siembra; 2) mes de siembra, 3) ratio de siembra; 4) temperatura promedio mensual; 5) precipitación promedio mensual.

Estas variables fueron evidenciadas durante el análisis descriptivo exploratorio inicial como variables de importancia al momento de explicar las libras de cosecha, esperándose que estas conduzcan a un poder predictivo adecuado conforme el cumplimiento de los objetivos de la presente investigación.

3.4 Métricas y comunicación de resultados

Las métricas de evaluación son establecidas por el negocio en su proceso de producción mediante los cuales se estima el desempeño productivo; adicionalmente, se establecen métricas específicas para la evaluación del aporte del proyecto a los objetivos del negocio como los expuestos en la tabla 3.7.

Tabla 3.8 Indicadores para evaluación del modelo

Indicador	Fórmula
Ajuste de pronóstico de modelo supervisado	$\frac{Libras\ cosechadas\ pronosticadas}{Libras\ cosechadas\ reales}$
Número de acuerdos de compras a proveedores terceros basados en proyecciones del modelo supervisado.	$Número\ de\ acuerdo\ de\ compras\ de\ camarón\ a\ proveedores\ terceros$
Efectividad en el cumplimento de ordenes de producción.	$\frac{Total\ de\ libras\ adquiridas\ basados\ en\ el\ modelo\ predictivo + Libras\ cosechadas}{Total\ de\ libras\ requeridas\ para\ cumplimiento\ de\ ordenes\ de\ producción}$

Fuente: Elaboración propia

CAPÍTULO 4

4 ANÁLISIS DE RESULTADOS

A continuación, se presentan los resultados obtenidos durante la ejecución de los diferentes modelos sujetos a estudio para la predicción del volumen de cosecha de camarón en las piscinas de la zona 1 de la camaronera sujeta de estudio.

4.1 Evaluación de métricas para la selección de modelos

Por los resultados obtenidos durante el análisis descriptivo realizado en el capítulo 3, y debido a la naturaleza de los objetivos de la presente investigación, se realizó la evaluación de diferentes modelos para la predicción del volumen de cosecha de camarón haciendo uso del criterio de evaluación RMSE para medir el error que existe entre el valor predicho y el valor observado (libras cosechadas) obteniendo los resultados expuestos en la tabla.

Tabla 4.1 Evaluación de modelos

	Model	MAE	MSE	RMSE	R^2	RMSLE	MAPE
lightgbm	Light Gradient Boosting Machine	6133.1845	104944295	9925.0177	0.8392	0.337	0.3054
rf	Random Forest Regressor	6068.0552	104337784	9985.7751	0.8427	0.2941	0.2594
et	Extra Trees Regressor	6063.3246	106306368	10010.9526	0.8549	0.2719	0.2246

Fuente: Elaboración propia

En la tabla 4.1 podemos observar los tres modelos con mejor rendimiento en las diferentes validaciones; en este caso, ordenados de mejor a peor según el criterio

de evaluación RMSE; sin embargo, se puede observar rendimientos similares para los otros criterios evaluados como lo son el MAE, MSE, R2.

En la tabla 4.2 se aprecia la configuración de cada uno de los modelos seleccionados y sus diferentes ajustes de hiperparámetros para la predicción.

Tabla 4.2 Pipeline Modelos seleccionados

Modelo	LGBMRegressor
LGBMRegressor(bagging_fraction=0.9, bagging_freq=3, boosting_type='gbdt', class_weight=None, colsample_bytree=1.0, feature_fraction=0.5, importance_type='split', learning_rate=0.4, max_depth=-1, min_child_samples=6, min_child_weight=0.001, min_split_gain=0.3, n_estimators=20, n_jobs=-1, num_leaves=150, objective=None, random_state=123, reg_alpha=0.005, reg_lambda=0.0005, silent='warn', subsample=1.0, subsample_for_bin=200000, subsample_freq=0)	
Modelo	RandomForestRegressor
RandomForestRegressor(bootstrap=True, ccp_alpha=0.0, criterion='mse', max_depth=9, max_features=1.0, max_leaf_nodes=None, max_samples=None, min_impurity_decrease=0.1, min_impurity_split=None, min_samples_leaf=4, min_samples_split=7, min_weight_fraction_leaf=0.0, n_estimators=100, n_jobs=-1, oob_score=False, random_state=123, verbose=0, warm_start=False)	
Modelo	ExtraTreesRegressor
ExtraTreesRegressor(bootstrap=True, ccp_alpha=0.0, criterion='mse', ax_depth=9, max_features=1.0, max_leaf_nodes=None, max_samples=None, min_impurity_decrease=0.1, min_impurity_split=None, min_samples_leaf=4, min_samples_split=7, min_weight_fraction_leaf=0.0, n_estimators=100, n_jobs=-1, oob_score=False, random_state=123, verbose=0, warm_start=False)	

Fuente: Elaboración propia

4.2 Evaluación de la capacidad predictiva en test de los modelos seleccionados

La ejecución de la predicción se la realizó con 90 observaciones seleccionadas (representando el 20% de los datos) para el test de los diferentes modelos a evaluar. Los resultados obtenidos de las predicciones para los 5 últimos ciclos de cosecha se presentan en la Tabla 4.3.

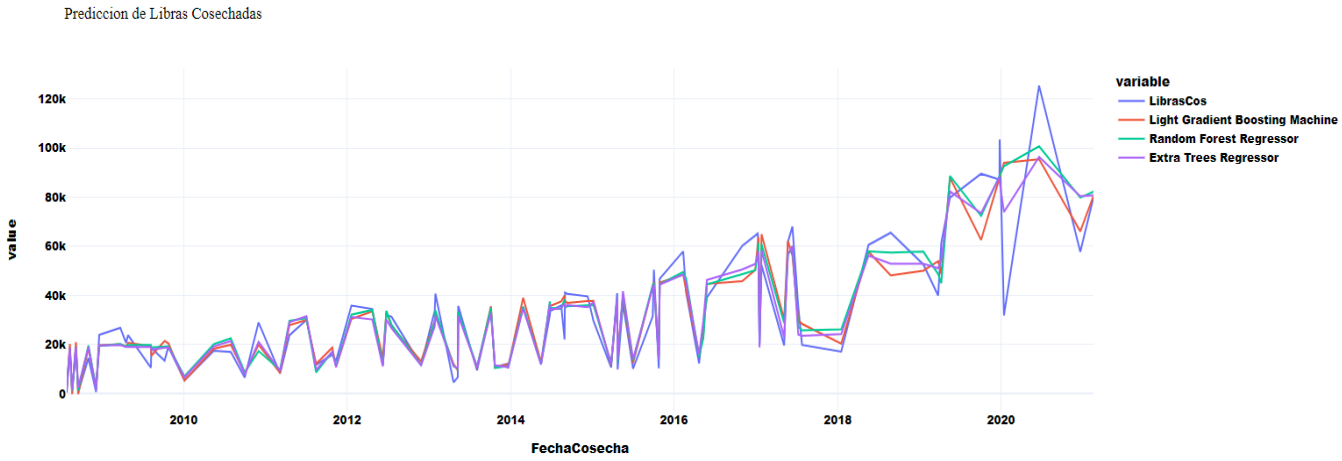
En la tabla 4.3 los valores reales de cosecha se presentan en la primera columna LibrasCos. En general el ajuste es bueno, con diferencias en los valores predichos para las observaciones 2 y 5, donde el modelo tiende a subestimar y sobrestimar, respectivamente. Aun así, el comportamiento para la serie en general es bastante adecuado lo que se puede ver mejor representado a través de un gráfico de líneas.

Tabla 4.3 Valores de predicción de modelos evaluados

	LibrasCos	Light Gradient Boosting Machine	Random Forest Regressor	Extra Trees Regressor
1	19110	21371.161	19635.35	20990.7
2	29010	19691.919	18045	18560.8
3	20760	19180.643	18474.6	18638
4	12180	12758.77	12289.1	11769.6
5	31860	36140.505	37863.05	39334.45

Fuente: Elaboración propia

Gráfico 4.1 Predicciones en test de los diferentes modelos



Fuente: Elaboración propia

El gráfico 4.1 presenta las series temporales correspondiente a los valores reales (en color celeste), y los tres modelos de predicción utilizados presentando en diferentes colores. Como se puede apreciar, las series ajustadas tienen un muy buen performance para modelar los datos reales, presentándose una mayor divergencia en los valores finales debido al aumento del volumen de cosecha que responden a cambios en procesos de siembras y tecnificación de las piscinas haciendo usos de sensores y equipos de IOT que permiten la automatización de la alimentación y mejoras en volúmenes de cosecha en respuesta al aumento significativo de la demanda por parte de los mercados asiáticos.

4.3 Evaluación del poder predictivo

Se realizaron evaluaciones de la capacidad predictiva de los modelos con datos de validación correspondiente a los últimos ciclos de siembra del año 2021 e inicio del año 2022 previo a conocer los volúmenes de cosechas reales.

Tabla 4.4 Predicciones Reales

Piscina	Fecha Siembra	Fecha Cosecha	Libras Cosechadas	Predicciones		
				LGBM	RF	ET
3	2021-08-25	2021-12-10	89250	82702.74	77788.95	67627.99
6	2021-09-23	2022-01-11	93380	84009.28	84501.59	79171.76
5	2021-09-29	2022-01-22	108570	82874.53	86342.04	79088.86
8	2021-10-10	2022-02-10	117600	82296.03	87439.51	84149.57
7	2021-11-15	2022-03-06	81800	85692.44	89224.38	91501.19

Fuente: Elaboración

Como podemos observar en la tabla 4.4 los ejercicios de validación se realizaron con diferentes piscinas seleccionadas aleatoriamente realizando el ingreso de los siguientes datos proporcionados por el usuario: fecha de siembra, fecha de cosecha, piscina y cantidad sembrada.

En general, las predicciones obtenidas por los modelos han brindado un aporte significativo para el negocio, ya que le permite generar una adecuada planificación logística para la cosecha y una correcta planificación para el abastecimiento y compra de camarones para satisfacer las ordenes de producción en curso.

CONCLUSIONES Y RECOMENDACIONES

CONCLUSIONES

Una vez realizado el análisis comparativo de los resultados de los 3 modelos de aprendizajes supervisados para la predicción del volumen de cosecha de camarón del cual se obtuvieron las siguientes conclusiones.

- La correcta selección de variables y adecuada extracción de características permitió enriquecer el dataset y entender de mejor forma el fenómeno a estudiar ya que nos brindó una visión holística y a la vez detallada de los factores que indicen en la obtención de los volúmenes de cosecha de camarón.

- Dentro de las características estudiadas se identificó que los meses cálidos son propicios para la cosecha de camarón ya que durante esos meses los volúmenes de producción eran más altos.

- En el marco del análisis comparativo los 3 modelos a evaluar tuvieron rendimientos similares durante la fase de test y validación en la cual se obtuvo métricas de buen desempeño; sin embargo, dentro de los márgenes evaluados el Random Forest es el modelo con un desempeño ligeramente superior durante la validación obteniendo predicciones más cercanas a los valores reales.

RECOMENDACIONES

Una vez concluido el análisis de los resultados y seleccionado el modelo con mejor rendimiento para la predicción del volumen de cosecha de camarón se procede con las siguientes recomendaciones.

- Se recomienda agregar los datos de temperatura y presión atmosférica a los datos que se recolectan periódicamente mediante uso de sensores de temperatura y humedad, además de agregar datos relacionados al agua como el nivel de acidez pH, temperatura del agua, acidez del suelo de la piscina, entre otras que permitan mejorar el poder predictivo del modelo

- Se recomienda la implementación de un sistema para la obtención de datos online que permita el aprendizaje por reforzamiento y el ajuste del modelo de forma autónoma para tener rendimientos acordes a las situaciones actuales.

- Se recomienda la continua evaluación del modelo y su posterior implementación en un sistema informático propio de la organización que permita usar la información de las predicciones para la toma de decisiones referentes a la planificación de compras de productos a camaroneras terceras y la planificación de la producción.

REFERENCIAS BIBLIOGRÁFICAS

Arana, C. (2021). Redes neuronales recurrentes: Análisis de los modelos especializados en datos secuenciales. *Serie Documentos de Trabajo.*(797). Obtenido de https://www.econstor.eu/bitstream/10419/238422/1/797.pdf

Barbier, M. (2021). *SERIES DE TIEMPO (CADENA DE SUMINISTRO)*. Obtenido de LOKAD: https://www.lokad.com/es/series-de-tiempo-en-cadena-de-suministro

Cevallos Valdiviezo, H., Rodríguez Cristiansen, A., Valdiviezo Valenzuela, P., Arévalo Avecillas, D., & Padilla Lozano, C. (2020). Predicción del nivel de cosecha de camarón blanco: el caso de una pequeña camaronera en la parroquia Tenguel del cantón Guayaquil, Ecuador. *Revista de Métodos Cuantitativos para la Economía y la Empresa*, 227-257. doi:10.46661/revmetodoscuanteconempresa.3791

González, A. (s.f). *¿Qué es Machine Learning?* Obtenido de Cleverdata: https://cleverdata.io/que-es-machine-learning-big-data/

Mauricio, J. A. (2007). Análisis de Series Temporales. *Análisis de Series Temporales.* Universidad Complutense de Madrid, Madrid. Recuperado el 07 de 11 de 2022, de https://www.ucm.es/data/cont/docs/518-2013-11-11-JAM-IAST-Libro.pdf

Montoya, N. P. (2005). HERRAMIENTAS PARA INVESTIGAR ¿Qué es el estado del arte? *Ciencia y Tecnología para la salud Visual y Ocular*(5), 73-75. Recuperado el 09 de 10 de 2022, de https://dialnet.unirioja.es/servlet/articulo?codigo=5599263

Pamungka, A., Zulkarnain, R., Adiyana, K., Waryanto, Nugroho, H., & Saragih, A. S. (2020). Application of Artificial Neural Networks to forecast Litopenaeus vannamei and Penaeus monodon harvests in Indramayu Regency, Indonesia.

IOP Conference Series: Earth and Environmental Science, 521, 1. doi:10.1088/1755-1315/521/1/012018

Shi, J., Jain, M., & Narasimhan, G. (2022). Time Series Forecasting (TSF) Using Various Deep Learning Models. *World Academy of Science, Engineering and Technology International Journal of Computer and Systems Engineering, 16*(6), 224 - 232. doi:10.48550/ARXIV.2204.11115

TIBCO Software. (s.f). *¿Qué es el aprendizaje supervisado?* Recuperado el 27 de 11 de 2022, de TIBCO Software: https://www.tibco.com/es/reference-center/what-is-supervised-learning

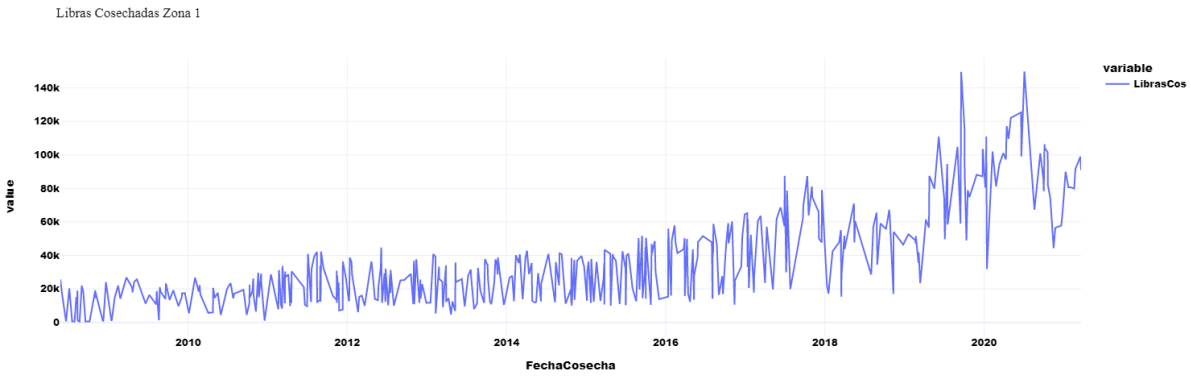
Valdez Pino, E. O. (2021). Desarrollo de un sistema de proyección de costos y costeo unitario de importaciones con métodos predictivos basados en machine learning. *Para optar el título profesional de Ingeniero de Sistemas.* UNIVERSIDAD PERUANA DE CIENCIAS APLICADAS, ima. Obtenido de https://repositorioacademico.upc.edu.pe/bitstream/handle/10757/657810/Valdez_PE.pdf?sequence=3&isAllowed=y

Vanegas López, J. J., & Vásquez Vergara, F. (2017). Multivariate Adaptative Regression Splines (MARS), una alternativa para el análisis de series de tiempo. *Gaceta sanitaria: Órgano oficial de la Sociedad Española de Salud Pública y Administración Sanitaria, 31*(3), 235-237. doi:https://doi.org/10.1016/j.gaceta.2016.10.003

ANEXOS

5 Anexos

Gráfico 5.1 Total de libras cosechadas en el tiempo



Fuente: Elaboración propia

Gráfico 5.2 Desviación y media de las libras cosechadas a través del tiempo

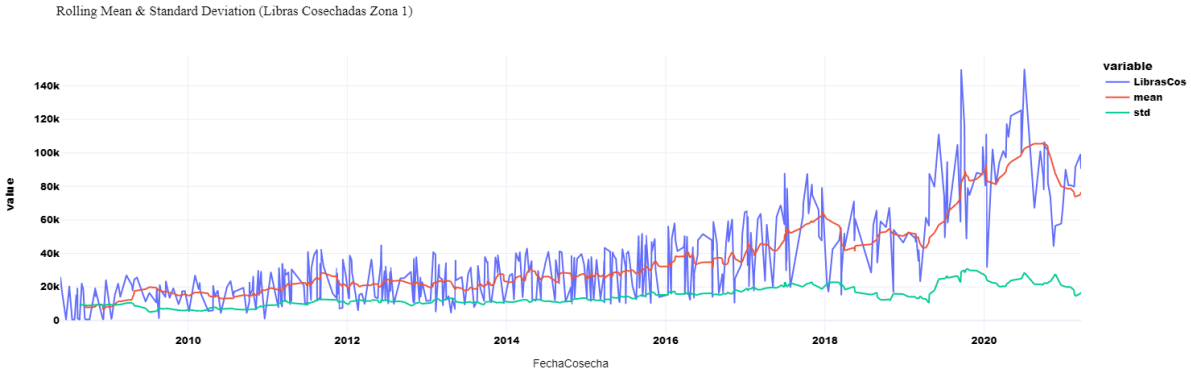
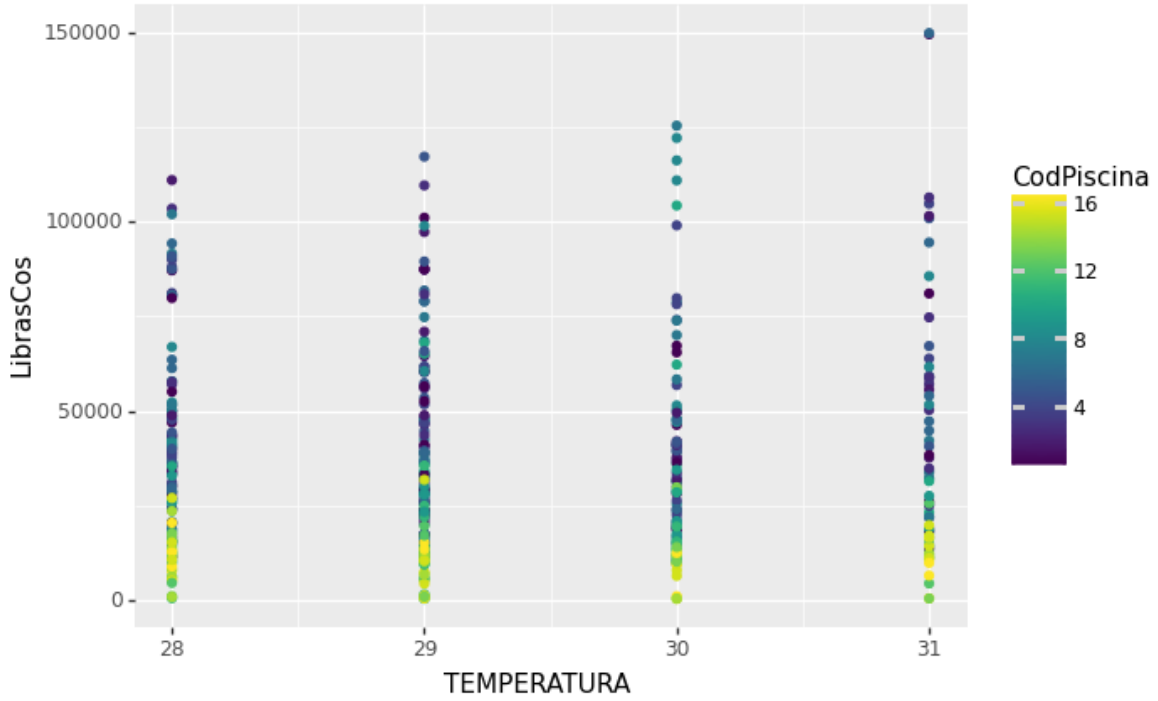
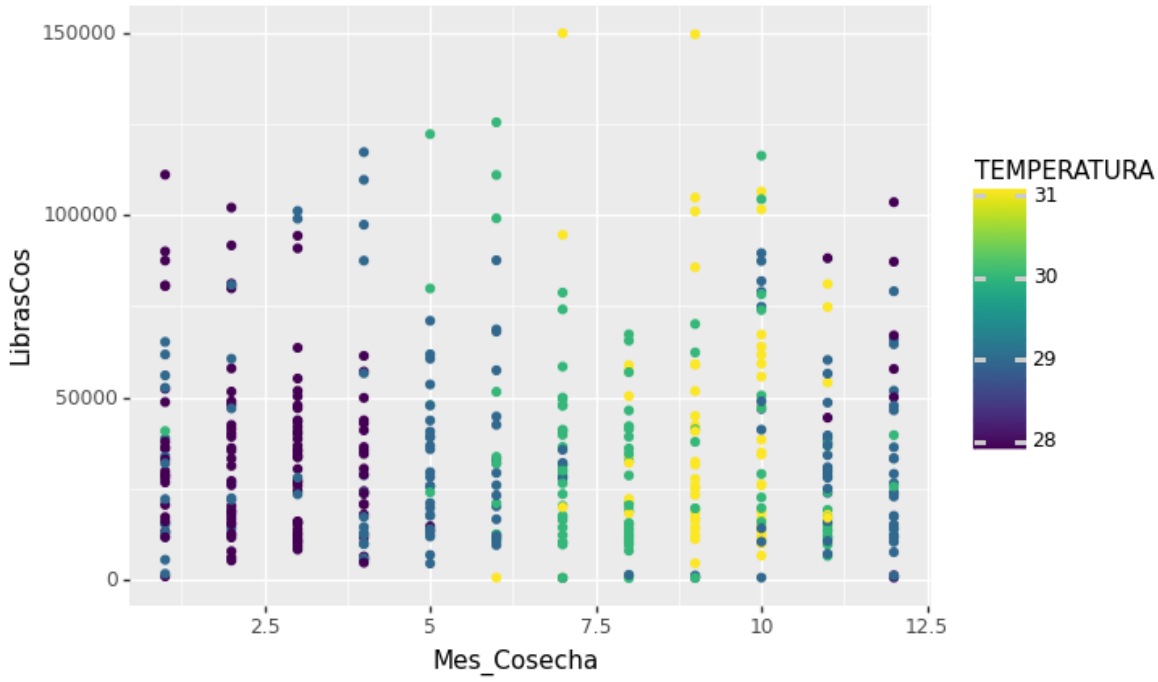


Gráfico 5.3 Total de Libras cosechadas en diferentes temperaturas promedio



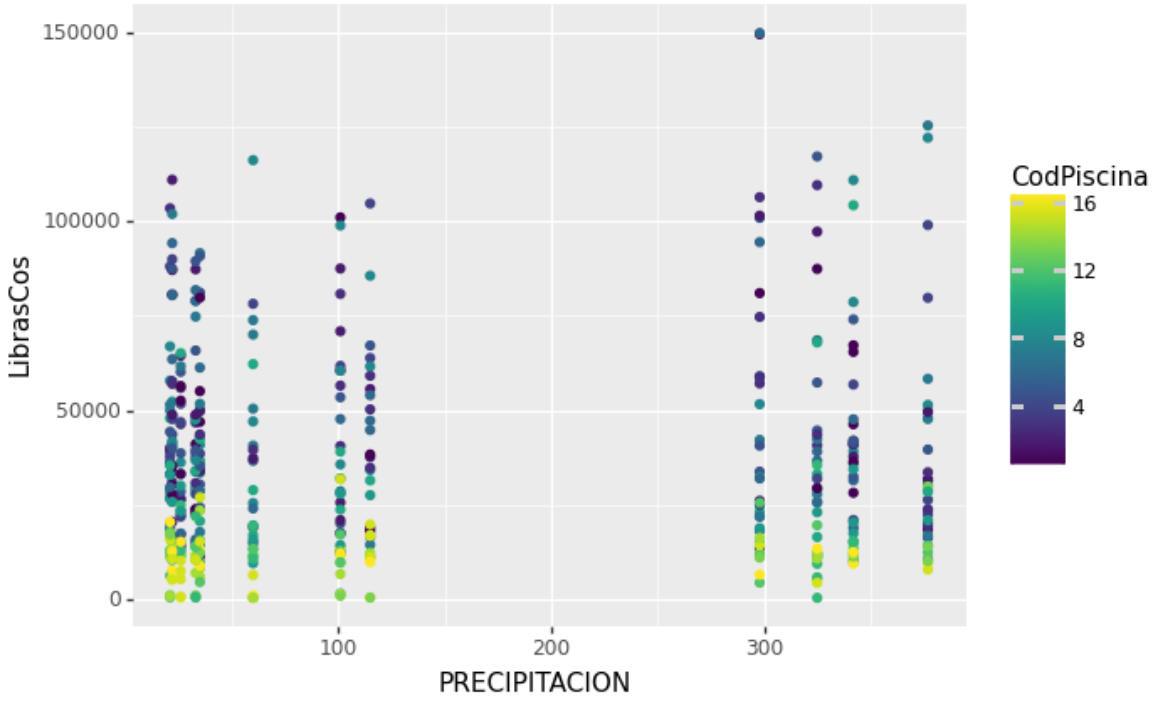
Fuente: Elaboración propia

Gráfico 5.4 Libras cosechadas en los diferentes meses del año



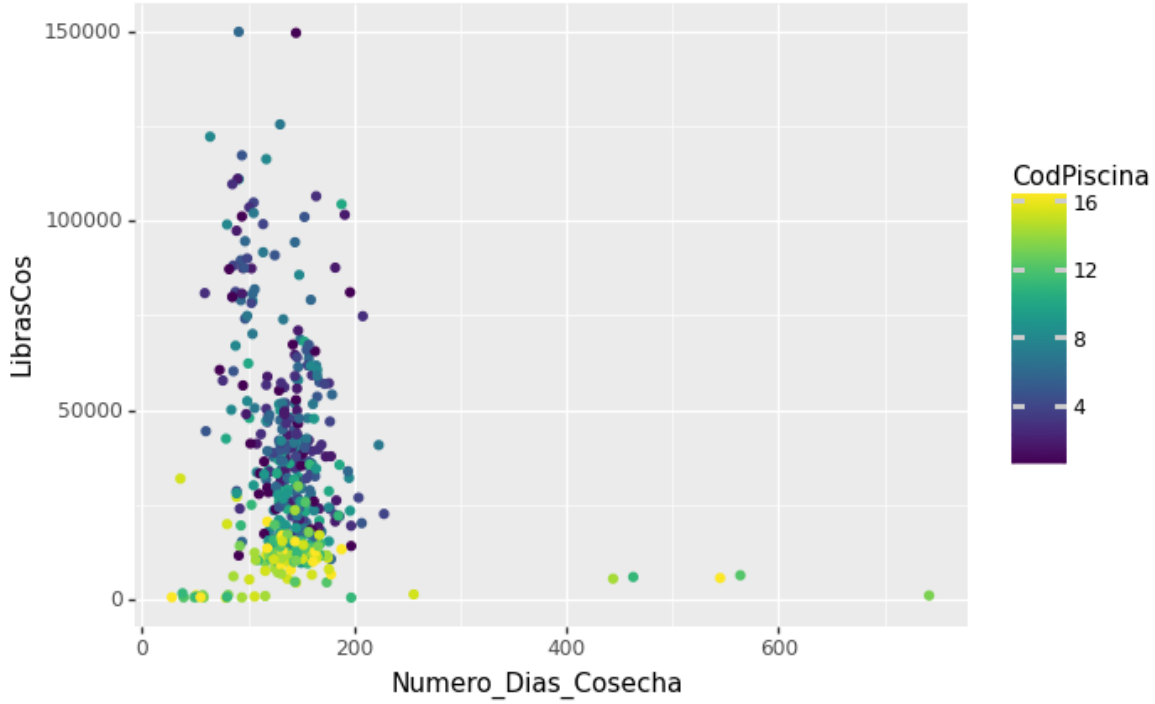
Fuente: Elaboración propia

Gráfico 5.5 Total de libras cosechadas según precipitación



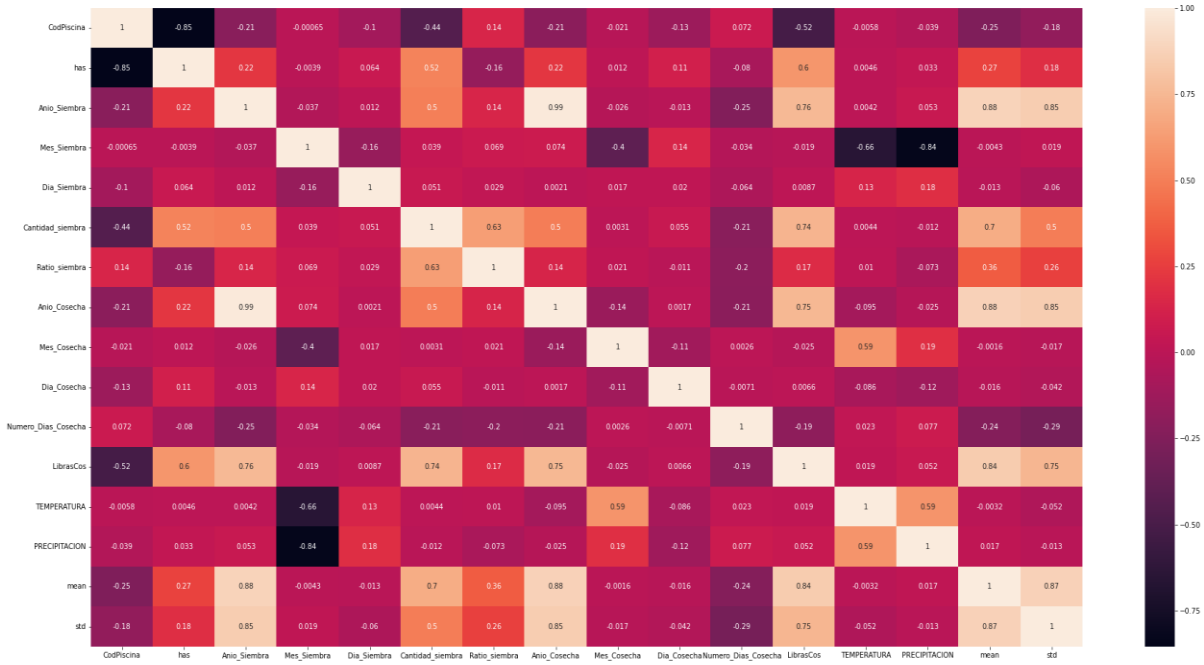
Fuente: Elaboración propia

Gráfico 5.6 Total de libras cosechadas por número de días de cosecha



Fuente: Elaboración propia

Gráfico 5.7 Matriz de correlación



Fuente: Elaboración propia