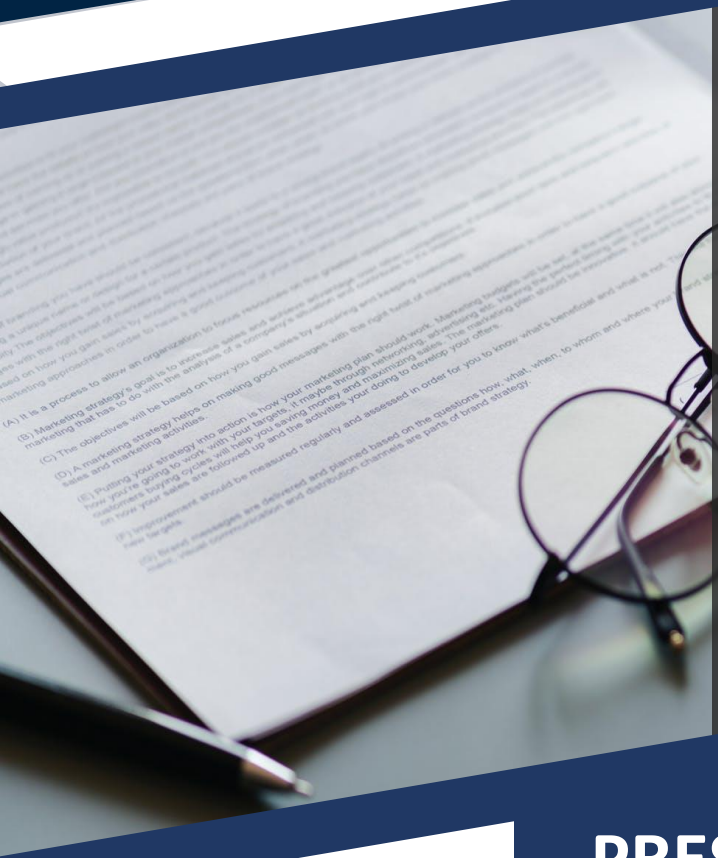
idata®
Inteligencia Analítica de Clientes



DOCUMENTACIÓN METODOLÓGICA

Prueba técnica
Problema #1

PRESENTA A

ecopetrol

Octubre de 2019

Derechos de Autor y Confidencialidad del documento

Medellín: Transversal 5ª #45-189- Bogotá: Calle 99 #10-57
Santiago de Chile: Camino El Alba 9500 Of 323
Línea de Servicio al Cliente (+57) (4) 444 97 60
contacto@idata.com.co_ www.idata.com.co

Versión	Responsable	Modificación	Fecha
V01	Idata	Creación de Documento	01/10/2019

Medellín: Transversal 5ª #45-189- Bogotá: Calle 99 #10-57
Santiago de Chile: Camino El Alba 9500 Of 323
Línea de Servicio al Cliente (+57) (4) 444 97 60
contacto@idata.com.co_ www.idata.com.co

Contenido

Medellín: Transversal 5ª #45-189- Bogotá: Calle 99 #10-57
Santiago de Chile: Camino El Alba 9500 Of 323
Línea de Servicio al Cliente (+57) (4) 444 97 60
contacto@idata.com.co_ www.idata.com.co

1. Contexto

Ecopetrol S.A. es la primera compañía de petróleo del país cuyo objeto social es el desarrollo, en Colombia o en el exterior, de actividades comerciales o industriales correspondientes o relacionadas con la exploración, explotación, refinación, transporte, almacenamiento, distribución y comercialización de hidrocarburos, sus derivados y productos.

Ecopetrol S.A. busca apoyar a uno de sus grupos de interés - Los pescadores del Valle Medio del Magdalena - por medio del uso de la analítica avanzada. La necesidad radica en que el precio del pescado que ellos venden depende 100% de la edad de este. Sin embargo, a los pescadores les lleva mucho tiempo calcular la edad del pescado, adicionalmente, cuentan con un conjunto de información donde se detallan diferentes características de los peces incluida su edad. De esta manera, la necesidad planteada por Ecopetrol S.A. consiste en planificar, diseñar, construir y "poner en productivo" un producto que solucione el problema de los pescadores, se espera que el producto sea consumible vía Web y tenga una excelente experiencia de usuario.

De esta manera, en el presente documento se muestran los objetivos del proyecto, la descripción y análisis de la información. Asimismo, se describe la metodología y técnicas utilizadas para la construcción del modelo.

2. Objetivo

2.1. Objetivo general

Desarrollar a través de herramientas analíticas un modelo de regresión, que apoye la necesidad puntual de Ecopetrol para predecir la edad de los pescados con el fin de apoyar la gestión del precio de estos para uno de sus grupos de interés.

Además, mediante un modelo de reconocimiento de imágenes se busca apoyar la identificación mediante imágenes y video de cascos industriales.

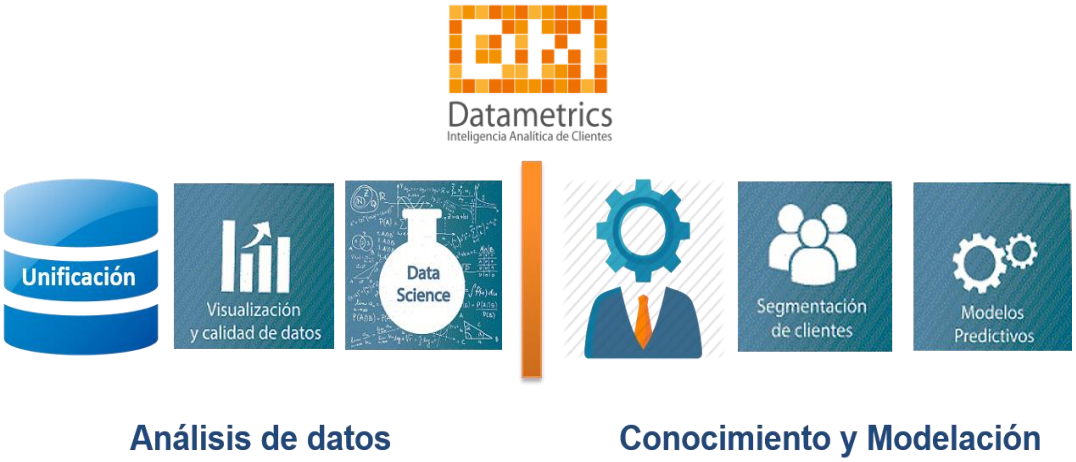
Medellín: Transversal 5ª #45-189- Bogotá: Calle 99 #10-57
Santiago de Chile: Camino El Alba 9500 Of 323
Línea de Servicio al Cliente (+57) (4) 444 97 60
contacto@idata.com.co_ www.idata.com.co

2.2 Objetivos específicos

- Realizar análisis descriptivo de la información.
- Construir un modelo analítico que permita la predicción de la edad de los pescados basado en las características disponibles de los mismos.
- Construir un modelo de reconocimiento en imágenes y video de cascos industriales.
- Industrializar ambos modelos mediante una aplicación web y el uso de microservicios.

3. Metodología

El desarrollo de soluciones analíticas se aborda desde nuestro proceso Datametrics de minería de datos, en el cual se cubren las etapas de extracción, análisis, calidad, conocimiento de la información, segmentación y construcción de modelos según la necesidad e interés estratégico del cliente.



4. Análisis de datos – Modelo No. 1

Parte del conocimiento de los campos que conforman la base de datos, identificando posibles bloques de información que componen a nivel individual cada unidad de análisis. Posteriormente se procede con la extracción y transformación de los datos en la herramienta de minería y se realiza un completo análisis descriptivo.

Medellín: Transversal 5ª #45-189- Bogotá: Calle 99 #10-57
Santiago de Chile: Camino El Alba 9500 Of 323
Línea de Servicio al Cliente (+57) (4) 444 97 60
contacto@idata.com.co_ www.idata.com.co

4.1. Descripción datos

Tabla 1. Ficha Técnica datos Ecopetrol

FACTOR DE ANÁLISIS	DESCRIPCIÓN
Fuentes de datos	CSV proporcionado por Ecopetrol (Separado por punto y coma).
Tamaño	173 kb
Número de registros	3.760 registros (aproximadamente) y 9 variables.
Ventana de tiempo transaccional	No se conoce la ventana de tiempo, es una única muestra.
Frecuencia temporal	La información no cuenta con una frecuencia temporal para su actualización.
Población Objetivo	3.760 pescados los cuales tienen la variable objetivo Rings (Edad) y sus correspondientes características.
Unidad de Análisis	Se cuenta únicamente con información de los pescados a analizar.
Bloques de Información analizados	Características propias de los pescados; sexo, longitud, diámetro, altura, peso total, peso pelado, peso de vísceras.
Hardware y Software	- Azure Databricks - Azure Datalake - DSVM (Data science virtual machine) - Azure Machine Learning Services. - Python

4.2. Exploración y análisis de datos

En esta sección se lleva a cabo un análisis exploratorio con el fin de identificar qué variables contiene información relevante para ser incorporada en el modelo. El análisis exploratorio consiste en examinar tablas de frecuencia que incluyen la media, la desviación estándar, el valor mínimo y máximo, los percentiles y los porcentajes, para lograr la identificación de valores atípicos o vacíos en cada uno de los campos que componen las bases de datos.

Medellín: Transversal 5ª #45-189- Bogotá: Calle 99 #10-57
Santiago de Chile: Camino El Alba 9500 Of 323
Línea de Servicio al Cliente (+57) (4) 444 97 60
contacto@idata.com.co_ www.idata.com.co

4.2.1. Sexo

En la tabla 2 se observa un total de 3.759 registros capturados de los cuales predominan los pescados de origen masculino con una participación del 39,6%, en segundo lugar, los pescados de sexo femenino representan el 34,5% del total, por su parte, el 25,9% de los registros de pescados presentan un sexo indeterminado.

Tabla 2. Tabla de frecuencia variable sexo

Sexo

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	F	1296	34,5	34,5	34,5
	I	974	25,9	25,9	60,4
	M	1489	39,6	39,6	100,0
	Total	3759	100,0	100,0	

4.2.2. Lenght, diameter, height, shuckedweight, viscera weight, wholeweight, shellweight

En la table 3, se pueden analizar un conjunto de variables que describen las características de los pescados tales como longitud, peso, diámetro entre otros aspectos. En general se observa que en la gran mayoría de variables todos los datos son válidos, a excepción de los campos Shuckedweight y Wholeweight.

La variable **Lenght** registra una validez del 100% de los datos, con una media de 0,54 y un valor mínimo y máximo de 0,205 y 0,815, respectivamente. Al analizar por percentiles, se tiene que, el 40% de los pescados tiene un longitud de 0,53 o menos, mientras que el 30% de los pescados presenta una longitud de 0,61 o más.

Con relación a la variable **diameter**, el 100% de los datos son válidos, se registra un promedio de 0,42 y un valor máximo de 0,650. El 80% de los registros de pescados presenta un diámetro de 0,5 o menos.

Medellín: Transversal 5ª #45-189- Bogotá: Calle 99 #10-57
Santiago de Chile: Camino El Alba 9500 Of 323
Línea de Servicio al Cliente (+57) (4) 444 97 60
contacto@idata.com.co_ www.idata.com.co

La variable **Height,** tiene todos los campos válidos, registra una media de 0,14; un valor mínimo y máximo de 0 y 1,130 respectivamente. Al analizar por percentiles, se observa que, el 30% de los pescados registra una altura de 0,165 o más.

Tabla 3. Tabla de estadísticos variables Lenght, diameter, height, shuckedweight, viscera weight, wholeweight, shellweight

Estadísticos

		Length	Diameter	Height	Shuckedweight	Viscera weight	Wholeweight	Shellweight
N	Válidos	3759	3759	3759	3718	3759	2381	3758
	Perdidos	0	0	0	41	0	1378	1
Media		,54790	,42762	,14659	,382788	,196634	,633782	,259652
Mediana		,56000	,43500	,15000	,364000	,185000	,626500	,250000
Moda		,550[a]	,450	,150	,1750[a]	,1715	,2225[a]	,2750
Desv. típ.		,097956	,081258	,037251	,1965936	,1034935	,2877672	,1297514
Varianza		,010	,007	,001	,039	,011	,083	,017
Mínimo		,205	,155	,000	,0170	,0055	,0425	,0155
Máximo		,815	,650	1,130	1,0300	,7600	2,5500	,8970
Percentiles	10	,41000	,31500	,10000	,139000	,069000	,261600	,100000
	20	,46500	,35500	,12000	,200000	,102500	,378500	,140500
	30	,50000	,38500	,13000	,252850	,131500	,466500	,177000
	40	,53000	,41000	,14000	,306800	,158500	,547000	,215000
	50	,56000	,43500	,15000	,364000	,185000	,626500	,250000
	60	,58500	,45500	,15500	,423500	,214500	,712000	,285000
	70	,61000	,47500	,16500	,484650	,244500	,798400	,319150
	80	,63500	,50000	,17500	,552600	,283000	,873000	,360000
	90	,66500	,52500	,19000	,650050	,336000	,953000	,430550

La variable **Shuckedweight** registra el 99% de sus campos como válidos, siendo el valor de la media de 0,38 y un valor máximo de 1,03. Por percentiles se encuentra que el 20% de los pescados tiene un peso estando pelado de 0,55 o más.

Con relación a la variable **Viscera weight**, el 100% de los datos son válidos, se registra un promedio de 0,196 y un valor mínimo y máximo de 0,0055 y 0,76. El 10% de los registros de pescados presenta un peso de la víscera de 0,336 o más.

La variable **Wholeweight** registra una validez del 43% de los datos, con una media de 0,63 y un valor mínimo y máximo de 0,425 y 2,55, respectivamente. Al analizar por percentiles, se tiene que, el 40% de los pescados tiene un peso entero de 0,54 o menos, mientras que el 30% de los pescados presenta un peso entero de 0,79 o más.

Medellín: Transversal 5ª #45-189- Bogotá: Calle 99 #10-57
Santiago de Chile: Camino El Alba 9500 Of 323
Línea de Servicio al Cliente (+57) (4) 444 97 60
contacto@idata.com.co_ www.idata.com.co

Finalmente, la variable **Shellweight** presenta una validez del 99% de los registros. Siendo el promedio de 0,25 y un valor máximo de 0,89. Por percentiles se observa que, el 20% de los pescados tiene un peso de la concha de 0,14 o menos, por su parte, el 30% tiene un peso de 0,31 o más.

4.2.3. Rings (edades)

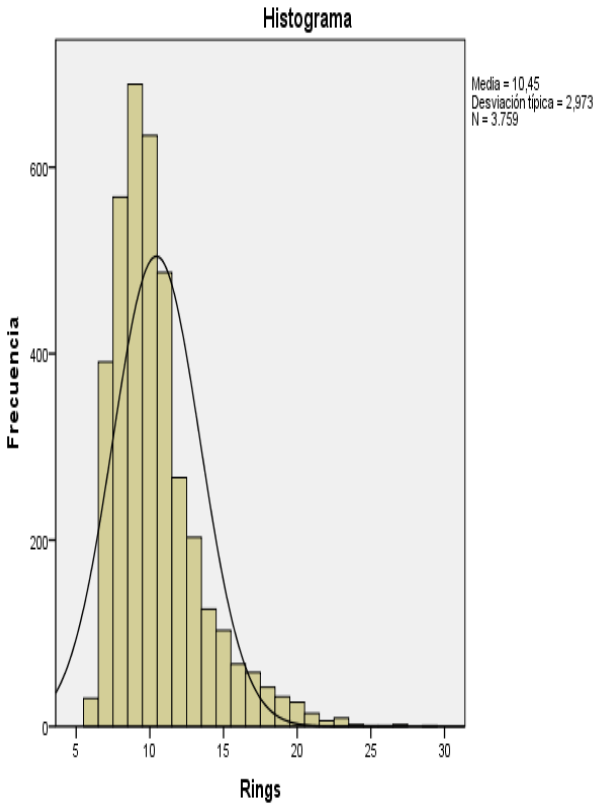
Se observa una media de 10,45 años para el registro total de pescados, con un valor mínimo y máximo de 6 y 29 años respectivamente. El valor de moda se ubica en 9 años, mientras que al analizar por percentiles se tiene que, el 40% de los pescados registra una edad de 9 años o menos, por su parte, el 30% de los pescados presentan edades de 11 años o más.

Tabla 3. Tabla de frecuencia e histograma Rigs

Estadísticos

Rings

N	Válidos	3759
	Perdidos	0
Media		10,45
Mediana		10,00
Moda		9
Desv. típ.		2,973
Varianza		8,837
Mínimo		6
Máximo		29
Percentiles	10	7,00
	20	8,00
	30	9,00
	40	9,00
	50	10,00
	60	10,00
	70	11,00
	80	12,00
	90	14,00



Medellín: Transversal 5ª #45-189- Bogotá: Calle 99 #10-57
Santiago de Chile: Camino El Alba 9500 Of 323
Línea de Servicio al Cliente (+57) (4) 444 97 60
contacto@idata.com.co_ www.idata.com.co

4.3. Conocimiento y modelación – Modelo No. 1

Las actividades que componen esta segunda fase son las más especializadas dentro del conjunto de procesos definidos en la metodología Datametrics. En esta etapa con la implementación de metodologías propias del análisis estadístico de datos, se buscó optimizar el tratamiento de la información finalizando con la construcción de un modelo predictivo.

4.3.1. Modelo ejecutado

Con el propósito de obtener el primer de los objetivos se entrenó un modelo de Gradient boosting el cual tuvo en cuenta las variables Length, Diameter, Height, Shuckedweight, Viscera weight, Wholeweight, Shellweight.

- El modelo arrojo un rms de 2.4.
- Brinda una estimación de cuales son las variables mas importantes para la clasificación.
- Las variables de mayor importancia para el presente ejercicio son: Shellweight, Shuckedweight, Wholeweight.

4.4. WebService

Con el fin de industrializar el modelo propuesto se ha puesto a disposición el modelo en Azure Machine Learning services basado en un cómputo de Kubernetes con lo cual se garantiza una estabilidad en el servicio y se admite así una alta concurrencia al modelo, como resultado se obtiene un punto de entrada que recibe peticiones POST con los parámetros del modelo y devuelve el resultado del modelo para ser interpretado en este caso dentro de un WebApp pero que podría ser integrado en otros ámbitos de Ecopetrol.

4.5. WebApp

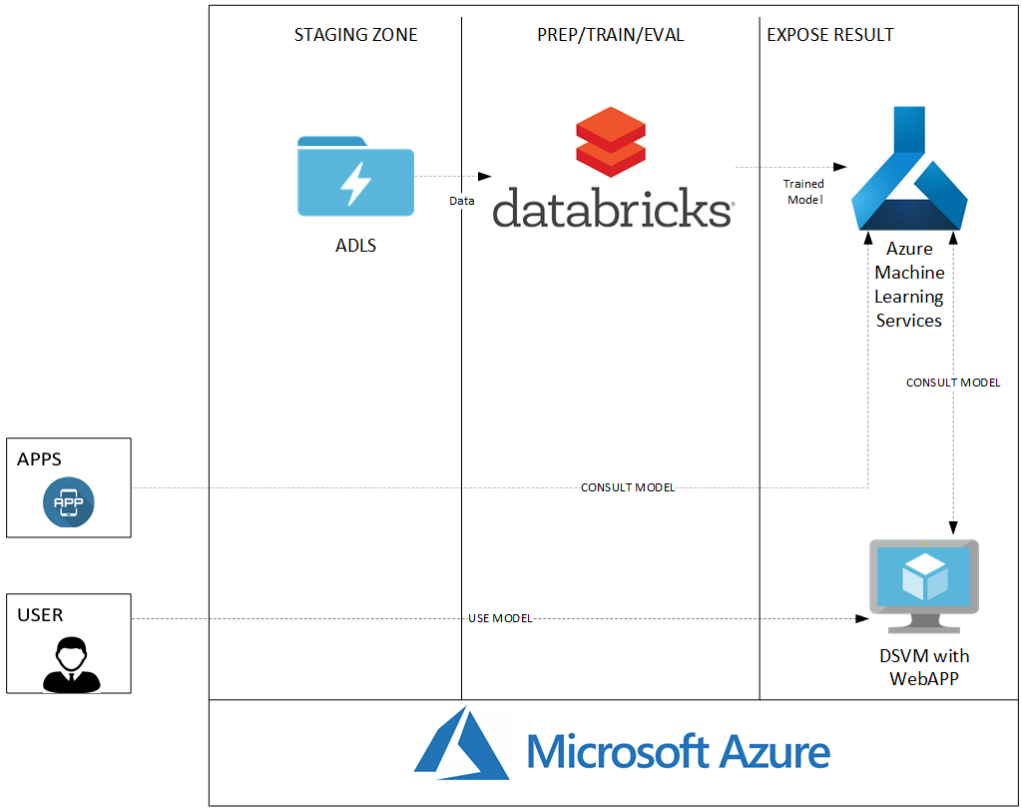
Con el fin de ofrecer una interfaz grafica para el uso del modelo predictivo y así ser de utilidad en el día a día de los pescadores del valle medio, se ha creado una aplicación web desarrollada en Python con framework Flask y desplegada en una Data Science Virtual Machine con el fin de garantizar una alta disponibilidad y escalabilidad. Esta WebApp es la encargada de recibir de manera amigable y enviar los parámetros al WebService de la solución para luego mostrar los resultados al usuario final

Medellín: Transversal 5ª #45-189- Bogotá: Calle 99 #10-57
Santiago de Chile: Camino El Alba 9500 Of 323
Línea de Servicio al Cliente (+57) (4) 444 97 60
contacto@idata.com.co_ www.idata.com.co

5. Arquitectura referencia

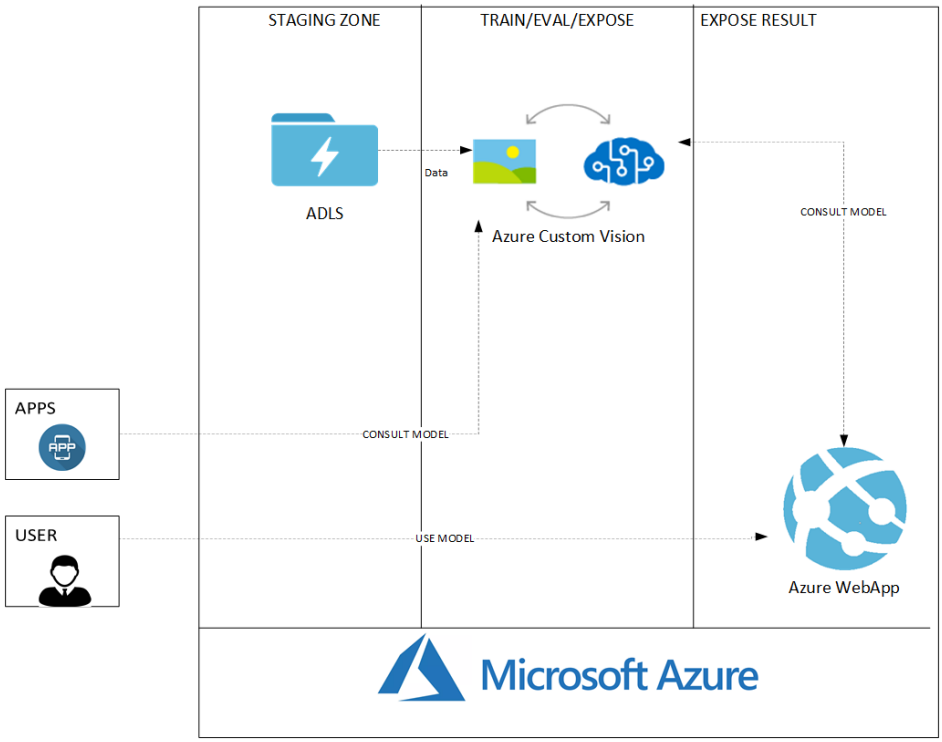
A continuación se presenta la arquitectura utilizada para el desarrollo del reto para la cual se desplegaron elementos dentro de la nube de Azure

Prueba 1



Medellín: Transversal 5ª #45-189- Bogotá: Calle 99 #10-57
Santiago de Chile: Camino El Alba 9500 Of 323
Línea de Servicio al Cliente (+57) (4) 444 97 60
contacto@idata.com.co_ www.idata.com.co

Prueba 2



Bibliografía

- Jiawei Han (2012). Data Mining Concepts and Techniques Third Edition

Medellín: Transversal 5ª #45-189- Bogotá: Calle 99 #10-57
Santiago de Chile: Camino El Alba 9500 Of 323
Línea de Servicio al Cliente (+57) (4) 444 97 60
contacto@idata.com.co_ www.idata.com.co