Ciencia de Datos Aplicada a Ciencias de la Tierra

Tema: Ejemplos de aplicaciones

Martín A. Díaz-Viera[1], Farhid M. Elisea Guerrero[2]

[1)] *mdiazv@imp.mx*, [2)] *felisea@imp.mx*

14 de marzo de 2024

Contenido I

Contenido II

Descripción de los datos

Los datos fueron tomados de la página Kansas Geological Survey, de la Universidad de Kansas (https://www.kgs.ku.edu/kgs.html).

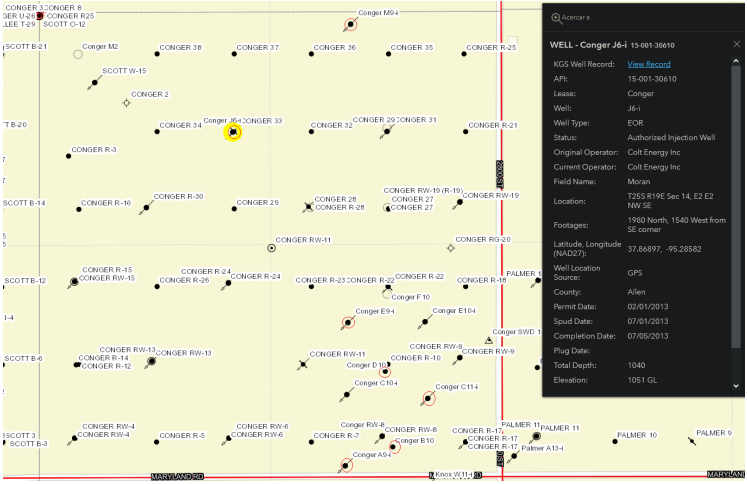
- Área de arrendamiento: Conger
- Operador: Colt Energy Inc
- Campo: Moran, County ALLEN, KANSAS, USA
- No. de pozos: 7
- Datos: Registros LAS
- Tipo de pozos: EOR

Descripción de los datos

Campo Moran, County Allen, Kansas, USA

- El desarrollo de este campo comenzó alrededor de 1903
- La producción se ha mantenido hasta la actualidad
- Las formaciones no están tan profundamente enterradas
- Producción de aceite por pozo pequeña (entre 25 y 5 bpd)
- El aceite es pesado (menor a 25 API)
- La geología consiste en capas alternas de calizas y lutitas y, localmente, de algunas areniscas.

Mapa de la ubicación de los pozos



Archivo de datos

Está en formato separado por comas (.csv) y contiene las variables:

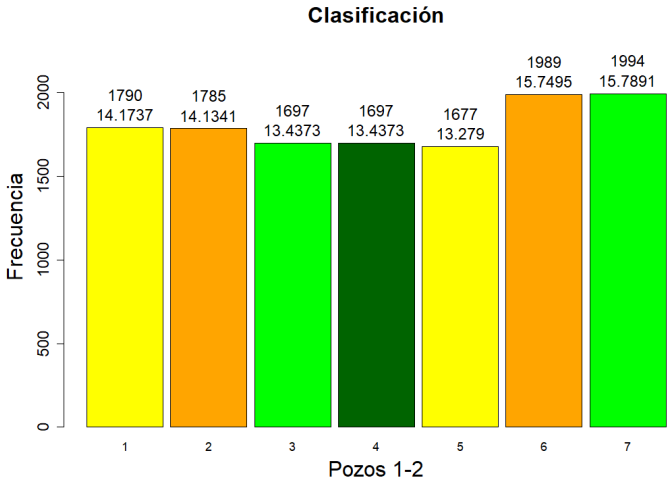
- count - número consecutivo
- Well - nombre del pozo
- Well.ID - identificador del pozo
- NAD83Long - longitud NAD
- NAD83Lat - latitud NAD
- Depth - profundidad
- GR - rayos gamma
- RILD - resistividad profunda
- RHOB - densidad
- NPOR - porosidad neutrón

Además varias clasificaciones hechas con métodos de agrupamiento K-medoides y K-means.

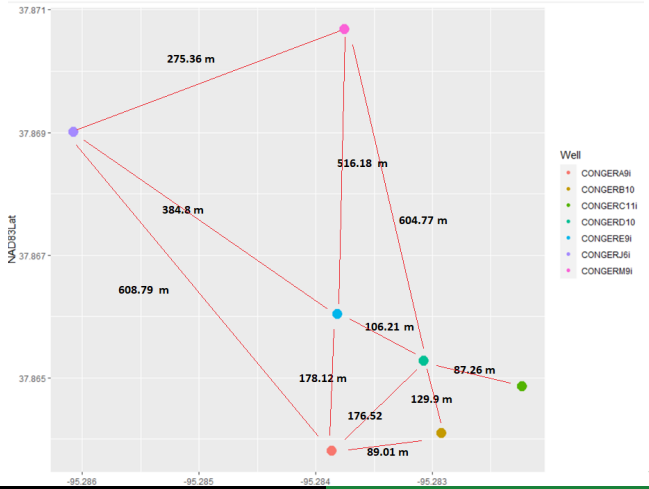
Pozos con sus intervalos y espesores

	Pozo	Prof.Min	Prof.Max	Espesor	ID.Pozo	Valores	Porcentaje
1	CONGERM9i	7.4676	311.2045	303.7369	7	1994	15.789
2	CONGERJ6i	8.6869	311.6617	302.9748	6	1989	15.749
3	CONGERE9i	29.8707	285.2962	255.4255	5	1677	13.279
4	CONGERD10	29.8707	288.3443	258.4736	4	1697	13.437
5	CONGERC11i	28.9563	287.4298	258.4735	3	1697	13.437
6	CONGERB10	16.7642	288.6491	271.8849	2	1785	14.134
7	CONGERA9i	16.7642	289.4111	272.6469	1	1790	14.174

Número de muestras por pozo



Sección de pozos

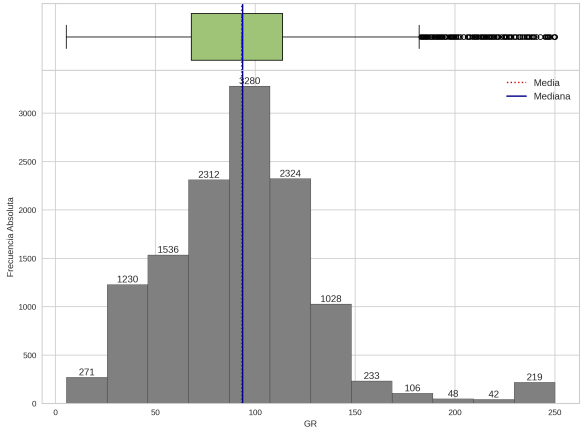


Análisis estadístico univariado

	GR	RILD	RHOB	NPOR
Muestras	12629.00	12629.00	12629.00	12629.00
Minimo	5.40	3.82	2.10	0.00
1er Cuartil	67.94	8.66	2.46	0.16
Mediana	93.62	11.22	2.52	0.20
Media	93.13	18.29	2.51	0.19
3er Cuartil	113.62	15.90	2.58	0.24
Maximo	250.00	242.72	2.95	0.30
Rango	244.60	238.91	0.85	0.30
Rango Intercuartil	45.68	7.24	0.12	0.08
Varianza	1563.44	500.53	0.01	0.00
Desviacion Estandar	39.54	22.37	0.11	0.05
Simetria	1.00	4.00	-1.15	-0.88
Curtosis	2.92	20.18	2.75	0.52

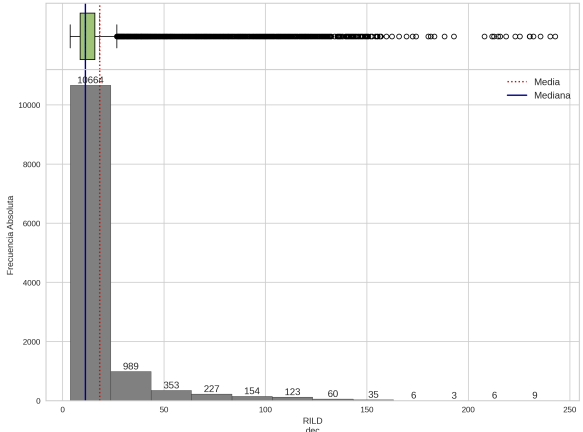
Análisis estadístico univariado

Histograma y Boxplot de GR



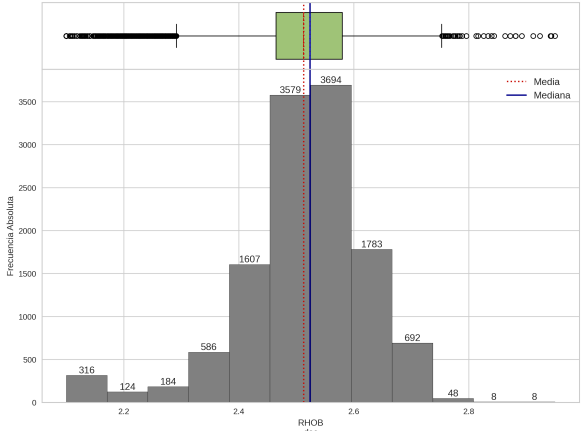
Análisis estadístico univariado

Histograma y Boxplot de RILD



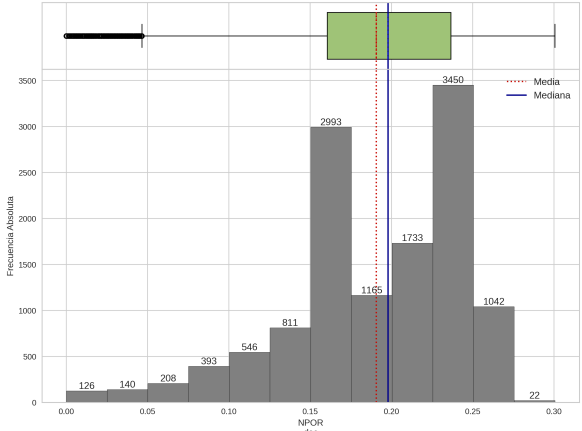
Análisis estadístico univariado

Histograma y Boxplot de RHOB



Análisis estadístico univariado

Histograma y Boxplot de NPOR



Análisis estadístico bivariado

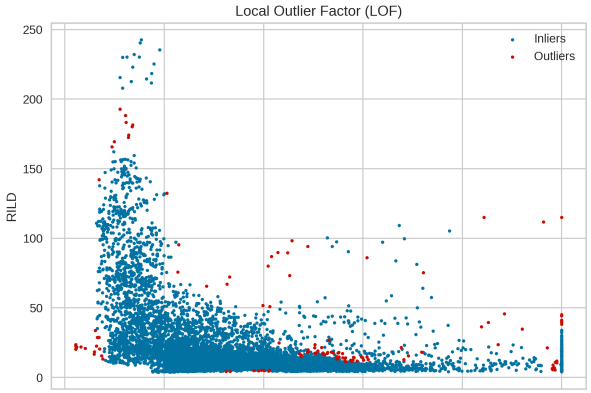
- Global,
- por intervalos,
- por secciones, etc

Análisis estadístico bivariado

- Matriz de gráficos de dispersión + histogramas
- Matriz de gráficas de calor (con medidas de dependencia)
- Medidas de dependencia (Pearson, Spearman y Kendall)
- Análisis más detallado donde existan las mayores dependencias
- Gráficos de dispersión + histogramas + boxplot + medidas de dependencias + valores atípicos

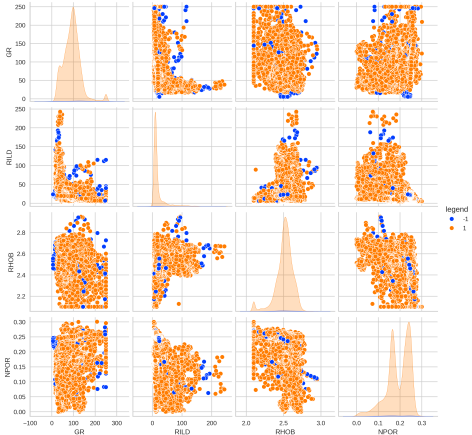
Análisis estadístico multivariado

Detección de valores atípicos usando el método Local Outlier Factor (LOF)



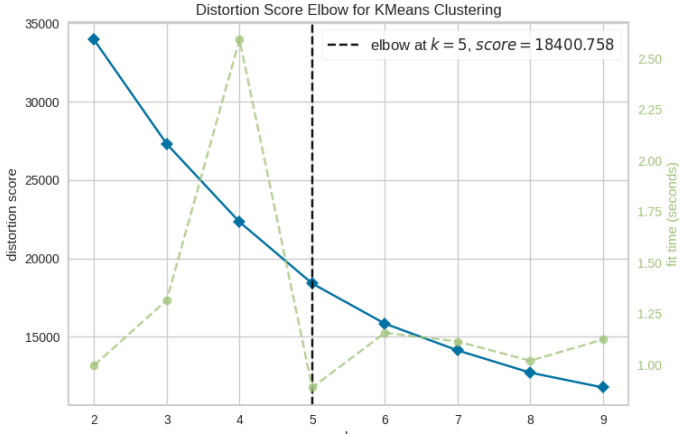
Análisis estadístico multivariado

Detección de valores atípicos usando el método Local Outlier Factor (LOF)



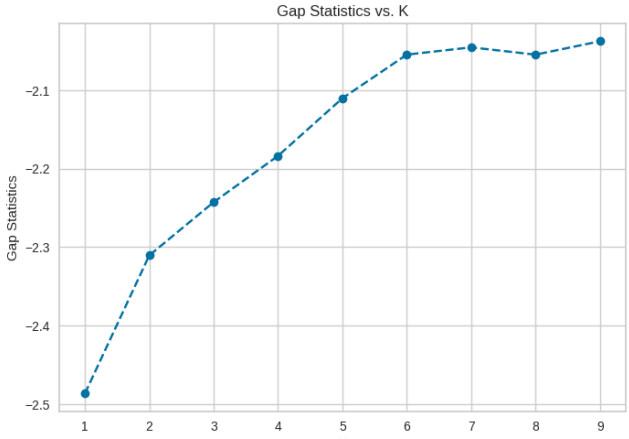
Determinación del número óptimo de grupos

Mediante el método del codo (elbow)



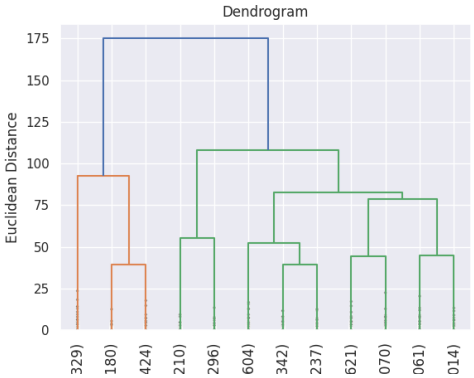
Determinación del número óptimo de grupos

Mediante el método del Gap



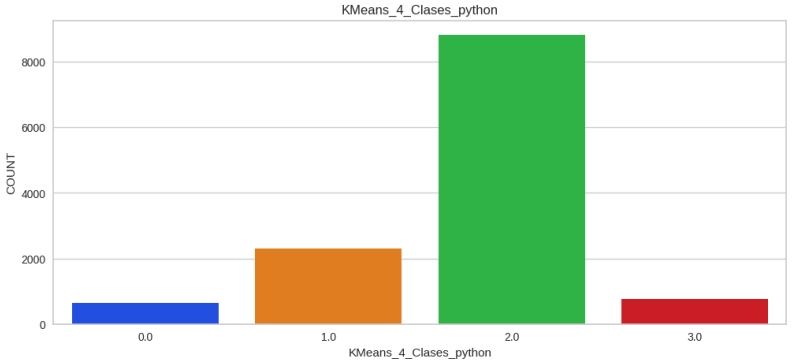
Determinación del número óptimo de grupos

Agrupamiento jerárquico usando distancia Euclidiana y el método de Ward



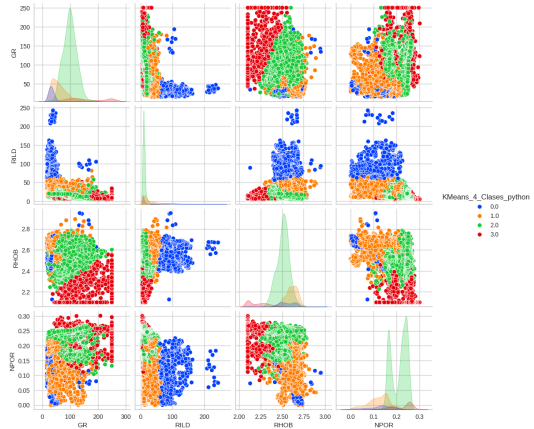
Aplicación de k-medias con 4 clases

Histograma del conteo por clases



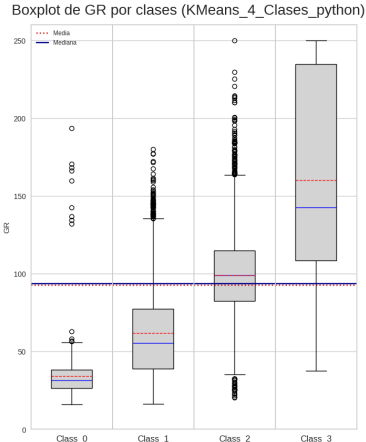
Matriz de dispersión por clases

Matriz de dispersión por clases



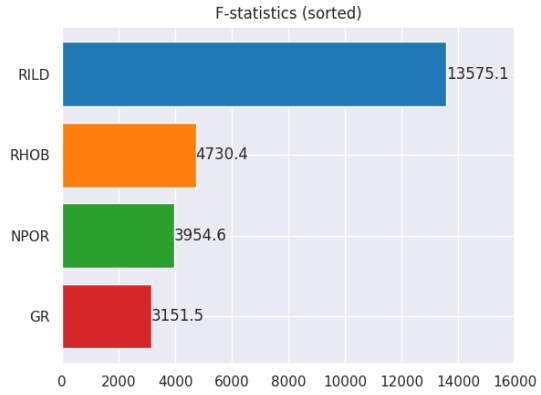
Boxplots de Rayos Gamma (GR) por clases

Boxplots de Rayos Gamma (GR) por clases



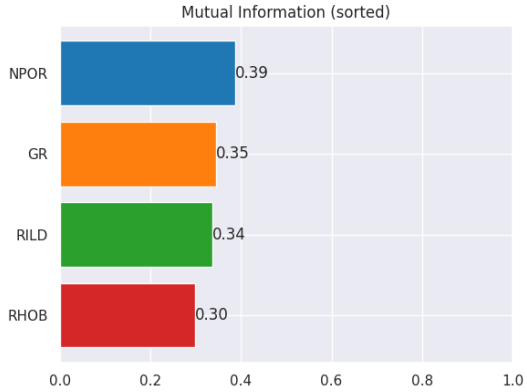
Selección de variables

ANOVA/F-Statistics



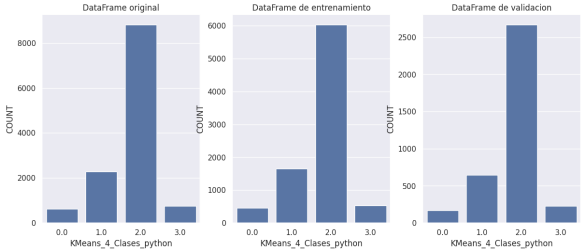
Selección de variables

Información mutua



Muestras de entrenamiento y de validación

Se seleccionan los subconjuntos de entrenamiento (70 %) y de validación(30 %)



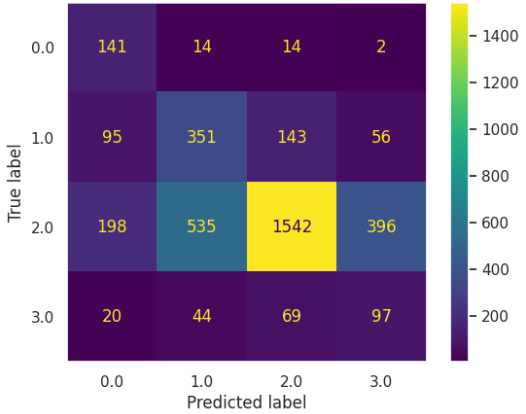
Muestras de entrenamiento y de validación

Resumen de las métricas de clasificación usando el método de Bosques Aleatorios

Model	Accuracy	Precision	recall	F1	AUC ROC
Random Forest Train	0.535809	0.458103	0.651816	0.472379	0.866113
Random Forest Test	0.492063	0.414121	0.584604	0.419849	0.818617
Random Forest (Extra) Train	0.476531	0.392891	0.523931	0.395085	0.778470
Random Forest (Extra) Test	0.474307	0.385232	0.517728	0.386049	0.769503
Random Forest (PCA) Train	0.402606	0.400660	0.537438	0.379741	0.802291
Random Forest (PCA) Test	0.393866	0.382261	0.476179	0.358758	0.757744
Random Forest (AdaBoost) Train	0.621382	0.495182	0.690393	0.534274	0.832392
Random Forest (AdaBoost) Test	0.573312	0.432653	0.591950	0.459034	0.756424

Muestras de entrenamiento y de validación

Matriz de confusión del método de Bosques Aleatorios con Adaboost



Referencias

[1] DellÁversana, P. , "Cross-disciplinary Machine Learning - Part 1: Applications to rock classification, well log analysis and integrated geophysics", 182 pags. (2020).

[2] Leijnen, S. and Veen, F. V., *The neural network zoo*. Proceedings, **47(1)** (2020).

[3] Rosenblatt, F., "The Perceptron: A Probabilistic Model For Information Storage And Organization In The Brain", *Psychological Review* **65 (6)** 386-408 (1958).

[4] Samuel, A. L. . "Some studies in machine learning using the game of checkers", *IBM Journal of Research and Development*, **3:3** 210–229 (1959).

Agradecimientos

- Este trabajo ha sido parcialmente financiado por los proyectos D.61037 y D.62002 del Instituto Mexicano del Petróleo.

Gracias!

Bienvenidas las preguntas!!!