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

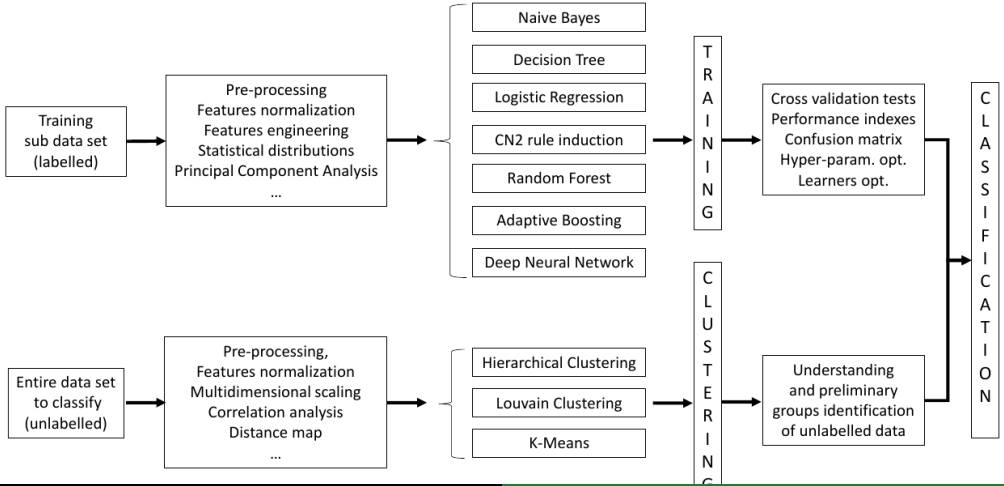
Caso de estudio

Aplicación a la clasificación de muestras de rocas

Aplicación a la clasificación de muestras de rocas

El presente caso de estudio está basado en el capítulo 1:
"Machine learning for rock samples analysis and classification" del libro
"Cross-disciplinary Machine Learning - Part 1: Applications to rock classification, well log analysis and integrated geophysics" (DellÁversana, 2020 [1])

Ejemplo de Flujo de Trabajo en Geociencias



Descripción de los datos

El conjunto de datos consiste en muestras de rocas con:

- diferentes tipos de rocas magmáticas (andesita, andesita basáltica, dacita y riolita),
- porcentajes de varios óxidos (SiO_2, TiO_2, K_2O y Al_2O_3) medidos en cada muestra.

Caso de Estudio (DellÁversana, 2020)

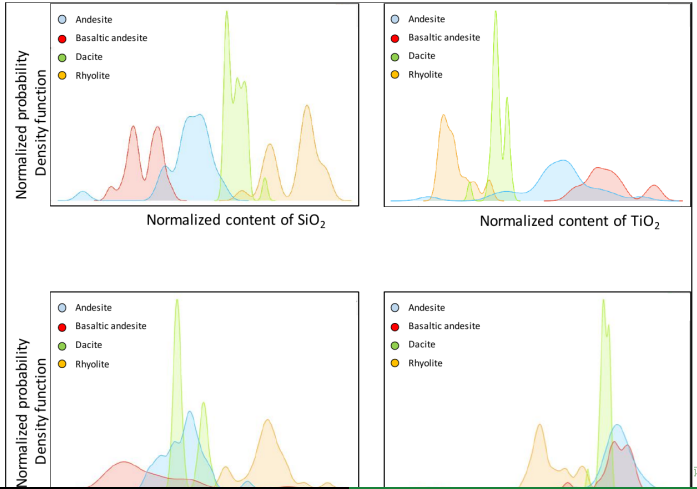
Se realizan dos pruebas diferentes:

- En la primera, se clasifica automáticamente (mediante aprendizaje automatizado) unas 1000 muestras pertenecientes a 4 clases: andesita, andesita basáltica, dacita y riolita.
- En la segunda prueba se clasifica unas 1500 muestras pertenecientes a 8 clases: andesita, andesita basáltica, dacita, riolita, granito, granodiorita, gneis y traquiandesita.

Métodos de aprendizaje supervisados

- La **Reglas de Inducción CN2** consiste en un algoritmo diseñado para la inducción eficiente de reglas simples de forma "si $<$ *condicion* $>$, entonces $<$ *predecir clase* $>$".
- **Bayes Ingenuo** es un clasificador probabilístico con un enfoque bayesiano que estima las probabilidades condicionales de la variable dependiente a partir de los datos de entrenamiento.
- **Árbol de Decisiones** es una técnica que funciona separando los datos en dos o más conjuntos homogéneos (o subpoblaciones).
- **Bosques Aleatorios** es un método de aprendizaje que utiliza un conjunto de árboles de decisión.
- **Reforzamiento Adaptativo** consiste en crear un clasificador fuerte como una combinación lineal de clasificadores "débiles".
- La **Regresión Logística** es un método comúnmente utilizado para la clasificación binaria.

Densidad de probabilidad normalizada



Selección de variables

Definición de algunos índices de la importancia relativa de las variables para una clasificación dada.

- **Info. gain** nos dice qué tan importante es una variable dada.
- **Gain ratio** es una relación entre la ganancia de información y la información intrínseca del atributo.
- **Gini** se puede considerar como un criterio para minimizar la probabilidad de clasificación errónea.
- **ANOVA** es la diferencia entre los valores promedio de la variable en diferentes clases.
- **Chi-Square** Dado un conjunto de datos sobre dos "eventos", podemos comparar el conteo observado O y el esperado E.
- **Relief** está relacionado con la capacidad de una variable para distinguir entre clases en instancias de datos similares.

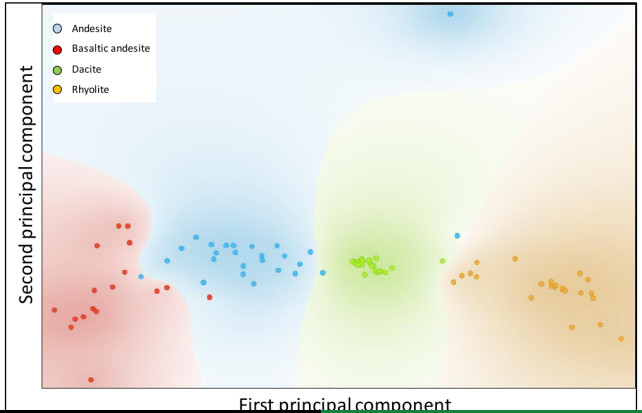
Selección de variables

La selección de variables es el proceso de extracción, clasificación y selección de las variables (o atributos) que mejor caracterizan al conjunto de datos.

	#	Info. gain	Gain ratio	Gini	ANOVA	χ^2	ReliefF
N SiO2		1.606	0.803	0.599	212.614	71.887	0.222
N CaO		1.509	0.755	0.551	252.737	68.742	0.259
N Fe2O3		1.400	0.700	0.498	191.196	64.762	0.255
N MgO		1.316	0.658	0.479	120.155	66.019	0.172
N TiO2		1.178	0.589	0.423	138.298	63.602	0.256
N MnO		1.105	0.553	0.383	8.212	57.969	0.081
N P2O5		1.063	0.532	0.371	45.265	55.229	0.115
N Na2O		0.919	0.460	0.313	14.106	39.444	0.043
N K2O		0.919	0.460	0.315	74.481	48.733	0.177
N Al2O3		0.832	0.416	0.277	36.346	46.890	0.118

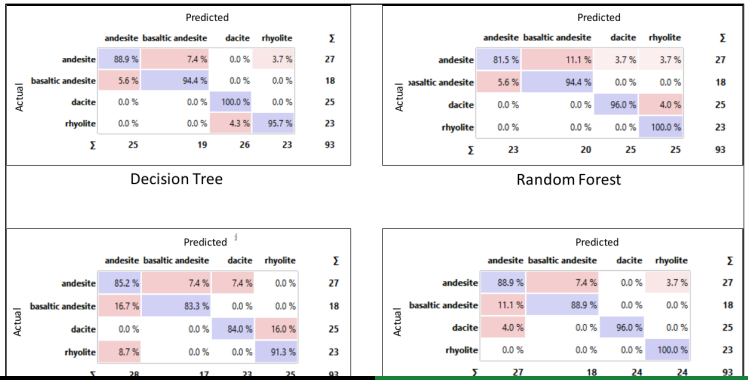
Análisis de componentes principales (PCA)

El PCA (Jolliffe, 2002) convierte un conjunto de variables correlacionadas usando una transformación ortogonal en un conjunto de variables no correlacionadas linealmente.



Validación cruzada

Los datos se dividen en dos subconjuntos: uno de entrenamiento y otro de validación. Existen varias técnicas de validación cruzada los métodos "K-fold", "Muestreo aleatorio" y "Omitir uno".



Indices cuantitativos del rendimiento

Existen varios índices cuantitativos para evaluar el rendimiento de clasificación de diferentes algoritmos

- **AUC** representa el grado o la medida de "separabilidad".
- **CA** es la proporción de ejemplos correctamente clasificados.
- La **precisión** es la proporción de verdaderos positivos entre las instancias clasificadas como positivas.
- La recuperación (**recall**) es la proporción de verdaderos positivos entre todas las instancias positivas en los datos.
- **F1** es una media armónica ponderada de precisión y recuperación.

Indices cuantitativos del rendimiento

Evaluation Results

Method	AUC	CA	F1	Precision	Recall
Tree	0.964	0.946	0.946	0.947	0.946
Random Forest	0.982	0.925	0.925	0.925	0.925
Naive Bayes	0.987	0.914	0.913	0.929	0.914
Logistic Regression	0.985	0.914	0.913	0.915	0.914
CN2 rule inducer	0.937	0.860	0.860	0.862	0.860
AdaBoost	0.957	0.935	0.935	0.936	0.935

Figura 6: Comparación de los resultados de seis algoritmos de ML.

Optimización de los hiperparámetros

- Una parte importante del flujo de trabajo de clasificación consiste en ajustar los hiperparámetros que son propiedades específicas de cada modelo.
- Por ejemplo, en Decision Tree, es el número mínimo de instancias en las hojas. En Random Forest, son la cantidad de atributos en cada división y la profundidad límite de cada árbol individual.
- Hay varios enfoques para ajustar los hiperparámetros:
  - Búsqueda en cuadrícula es un método de búsqueda exhaustivo de fuerza bruta.
  - Búsqueda aleatoria es un enfoque alternativo para muestrear diferentes combinaciones de parámetros.

Análisis de agrupamiento (K-means)

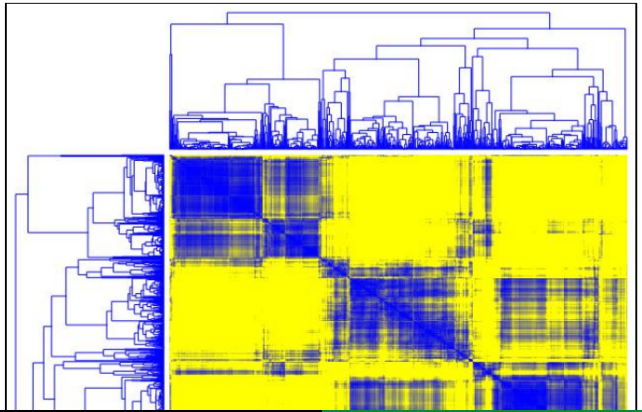
- Es uno de los métodos de agrupamiento más reconocidos por su antigüedad.
- Asigna las instancias del conjunto a clasificar a un total de k grupos especificados al inicio de la ejecución de forma aleatoria.
- Se minimizan las diferencias entre instancias de un mismo grupo y, al mismo tiempo se maximizan las diferencias entre grupos.
- Se actualizan las asignaciones a grupos hasta que ya no es posible mejorar las métricas o se alcanza un número máximo de iteraciones.
- En cada iteración, se calculan k puntos (o centroides) que se emplean para calcular las diferencias.

Análisis de agrupamiento (K-means)

- El análisis de agrupamiento es crucial en el aprendizaje automátizado no supervisado.
- Después de un agrupamiento robusto, es más fácil el proceso de clasificación.
- Por lo general, los datos se normalizan para un tratamiento equitativo de las variables.
- El primer paso importante es considerar algún tipo de distancia.
- Podemos elegir entre muchos tipos de métricas de distancia, como: Euclidiana, Manhattan, etc.

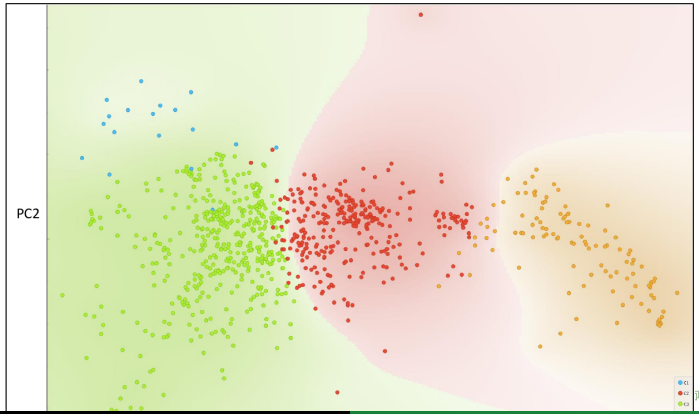
Mapa de distancias (K-means)

Como resultado del análisis de agrupamiento se obtiene una matriz de distancias que puede ser representada por un mapa de distancias



Gráfica de dispersión de CP (K-means)

Otra representación del análisis de agrupamiento es mediante una gráfica de dispersión de componentes principales.



Redes neuronales

- La primera red neuronal artificial "Mark I Perceptron", creada por Frank Rosenblatt (1958) [3], se basó en un modelo muy simple de conexiones neuronales.
- Supone que las conexiones entre las neuronas podrían cambiar a través de un proceso de aprendizaje supervisado que reduce el desajuste entre la salida real y la esperada.
- El resultado esperado proviene de un conjunto de datos de entrenamiento.
- Esa discrepancia se propaga hacia atrás por toda la red y permite actualizar los pesos de las conexiones.
- En otras palabras, el desajuste entre las respuestas reales y esperadas de la red representa la información necesaria para mejorar el desempeño del aprendizaje.

Redes neuronales

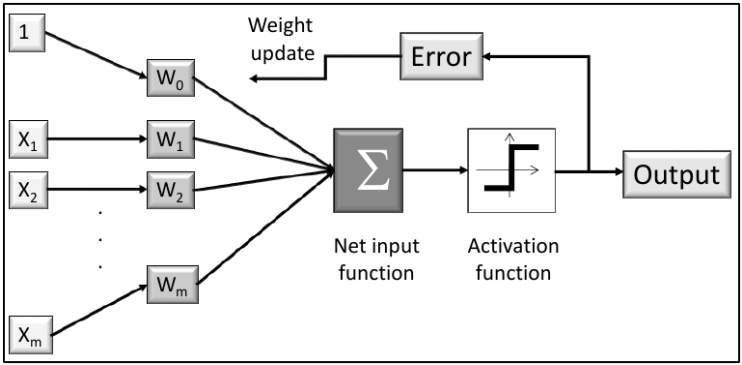


Figura 9: Función de una Red Neuronal.

Redes neuronales

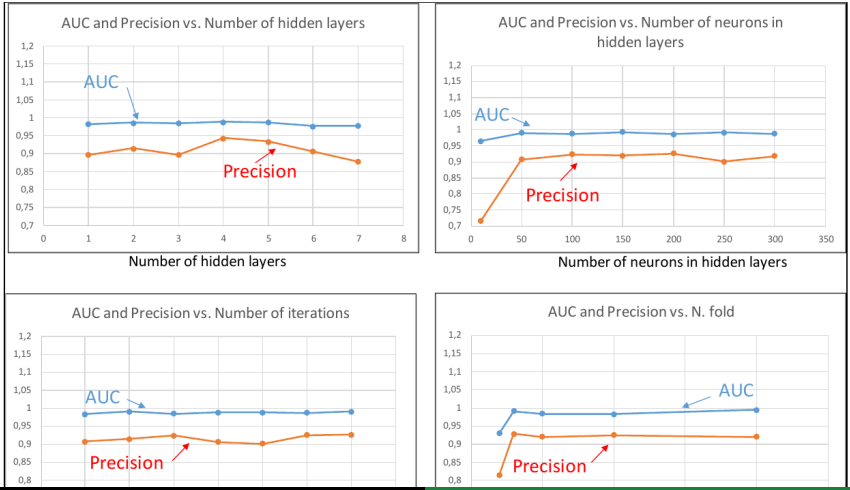
Flujo de aprendizaje de la red neuronal:

1. Propagación directa de los patrones de entrada (vector de datos de entrenamiento) desde la capa de entrada a través de la red para generar una salida.
2. Cálculo del error comparando la salida de la red con la salida deseada.
3. Retropropagación de errores, cálculo de derivadas con respecto a cada peso en toda la red y actualización del modelo.
4. Iteramos los pasos anteriores muchas veces (épocas o iteraciones) hasta obtener una convergencia adecuada (reducción de errores por debajo del umbral deseado).

Redes neuronales profundas

- **Red Neuronal Profunda** (DNN) es un tipo particular de red neuronal multicapa con más de una capa oculta.
- Inspirada en los modelos jerárquicos de nuestro sistema visual.
- Va desde las retinas, a la corteza visual, a la corteza occipital, y finalmente las áreas asociativas de alto nivel.
- Un nivel de la jerarquía se construye combinando entradas de unidades de un nivel inferior.
- Su organización jerárquica en capas le permite compartir y reutilizar información.
- Es posible seleccionar características específicas y rechazar detalles inútiles.

Optimización de Parámetros de DNN

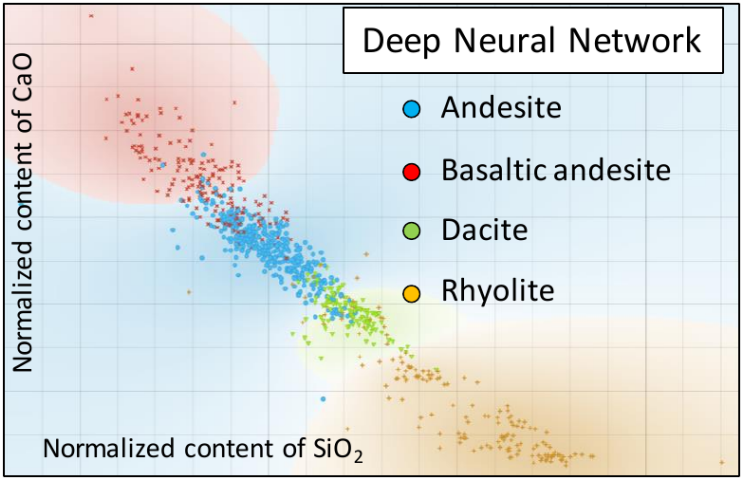


Indices de desempeño de siete MLM

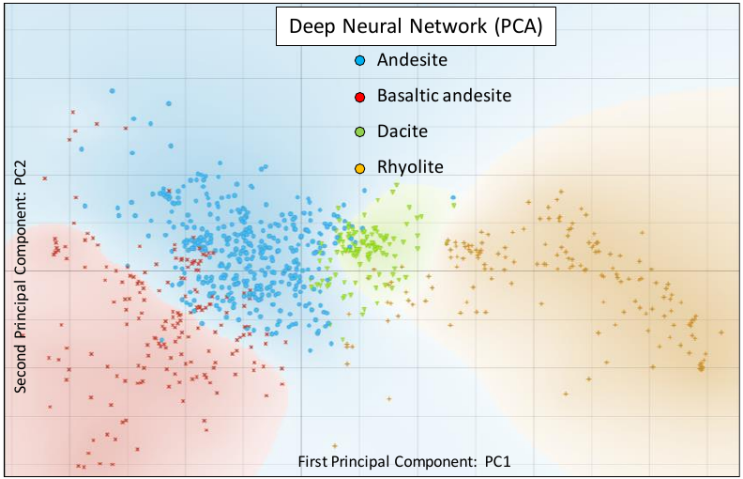
Method	AUC	CA	F1	Precision	Recall
Tree	0.964	0.946	0.946	0.947	0.946
Random Forest	0.990	0.914	0.914	0.915	0.914
Neural Network	0.989	0.914	0.913	0.932	0.914
Naive Bayes	0.987	0.914	0.913	0.929	0.914
Logistic Regression	0.985	0.914	0.913	0.915	0.914
CN2 rule inducer	0.937	0.860	0.860	0.862	0.860
AdaBoost	0.957	0.935	0.935	0.936	0.935

Figura 11: Indices de desempeño de siete MLM.

Clasificación con Redes Neuronales Profundas I (DNN)



Clasificación con Redes Neuronales Profundas II (DNN)



Etapas de la Ciencia de Datos

Etapas de la Ciencia de Datos

Flujo de Trabajo General

1. Generación de la base de datos (conformación, revisión y depuración)
2. Análisis exploratorio de los datos (univariado, bivariado, multivariado)
3. Métodos de aprendizaje no supervisados
4. Métodos de aprendizaje supervisados

Recomendaciones

- Los métodos de Aprendizaje Automatizado (ML) pueden integrarse en los flujos de trabajo de la modelación Geológico-Petrofísica de Yacimientos de manera exitosa.
- Pero deben hacerlo seguiendo las metodologías que incluyen los métodos del Análisis Estadístico de los Datos.
- Siempre probar más un método de Aprendizaje Automatizado (ML) para poder comparar.

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).