01 EDA

Diabb Zegpi D.

El objetivo de este documento es registrar el proceso de análisis exploratorio de datos (EDA, por su sigla en inglés) sobre el conjunto de datos de los ensayos de HEA (**H**igh **E**ntropy **A**lloys).

Análisis de datos numéricos

Se inicia el análisis con la exploración descriptiva del dataset. Las dimensiones del conjunto de datos son 2477 registros por 4 variables. La Tabla 1 ilustra los estadísticos descriptivos de las variables numéricas del dataset.

Tabla 1: Estadísticos descriptivos de las variables numéricas: tasa de valores faltantes; media aritmética; desviación estándar; mínimo; percentil 25; mediana (p50); percentil 75; máximo.

variable	tasa de faltantes	media	desv. est.	mínimo	p25	p50	p75	máximo
dHmix	0%	-9.74	10.21	-76.44	-13.97	-7.01	-3.75	9.90
Elect.Diff	0%	0.16	0.18	0.00	0.11	0.13	0.18	3.92
VEC	0%	6.58	1.70	1.62	4.73	7.20	7.99	11.77

De la Tabla 1 destaca que ninguna de las 3 variables numéricas presenta datos faltantes. La variable `dHmix`, ostenta una alta desviación estándar respecto de su media ($CV = 1.05$), lo que sugiere una distribución con asimetría negativa. Lo contrario puede decirse de la variable Elect.Diff ($CV = 1.07$), que en conjunto con su alta dispersión, su distribución presenta asimetría positiva. Finalmente, la variable `VEC` es la que tiene menor coeficiente de variación ($CV = 0.26$), exhibiendo la distribución menos asimétrica de las 3. Las distribuciones de las variables numéricas se ilustran en la Figura 1. Los puntos de alta densidad de la `Elect.Diff` se explican por la alta dispersión de esta variable; una transformación matemática del tipo logaritmo o raíz cuadrada podrían reducir la asimetría de su distribución. En cambio, `VEC` posee una distribución bimodal, por tanto, un análisis bivariado y/o de clusters ayudaría a arrojar luces sobre el factor que afecta su distribución.

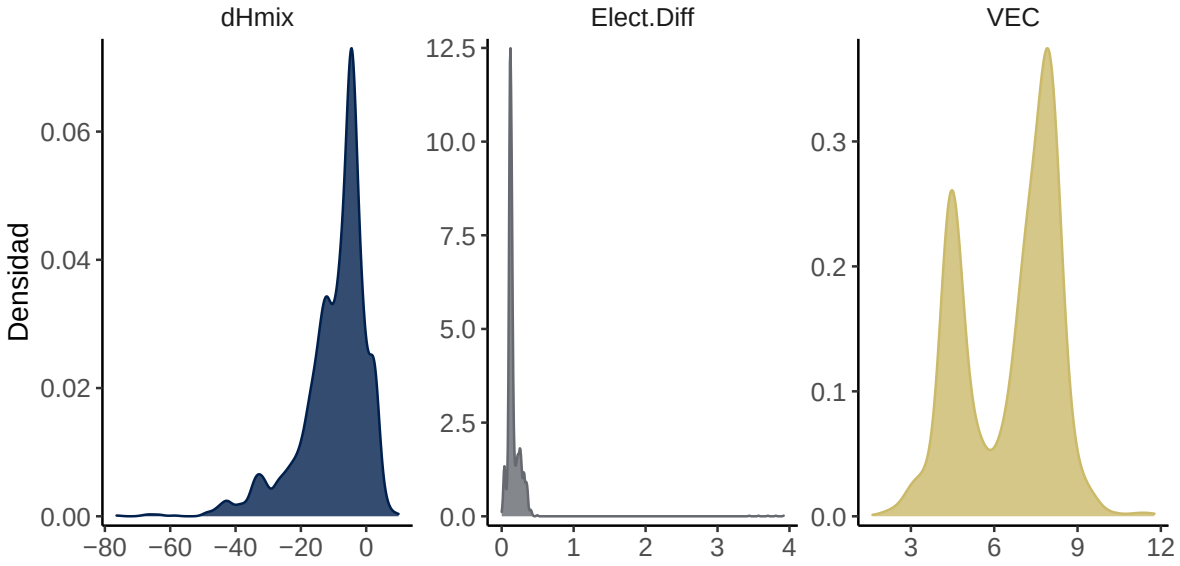

Figura 1: Estimación de densidad de las variables numéricas mediante kernel gaussiano.

Las relaciones entre las variables numéricas se ilustran en la Figura 2; en el gráfico se `dHmix` contra `Elect.Diff` se aprecia la presencia de outliers, valores extremos que sesgan la distribución de la variable ordenada. La gráfica de en medio demuestra la separación de conglomerados de la variable `VEC`, sin embargo, esta división de grupos no se perceptible solamente usando `dHmix`, debido a que los grupos son homogéneos en esta dimensión. Finalmente, la distribución de `VEC` a lo largo de `Elect.Diff` es aparentemente uniforme, con media y varianza estacionarias.

La investigación de los valores fuera de rango y su eventual corrección (si procediere), es uno de los objetivos relevantes para mejorar las propiedades de distribución de las variables, específicamente `Elect.Diff`.

Los diagramas de dispersión de la Figura 2 sugieren que no hay relaciones lineales entre las variables, pero sí existen conglomerados de observaciones. Dos métodos de clustering sugeridos para determinar los grupos presentes en los datos son: k-Means (clustering elíptico basado en centroides) y DBScan (clustering basado en densidades), con sus respectivas etapas de pre procesamiento.

Con fines predictivos, se recomienda la utilización de algoritmos que reconozcan separaciones en los datos, tales como los algoritmos basados en árboles de decisión (random forest, XGBoost, etc.) y algoritmos de proximidad entre vecinos (k-NN).

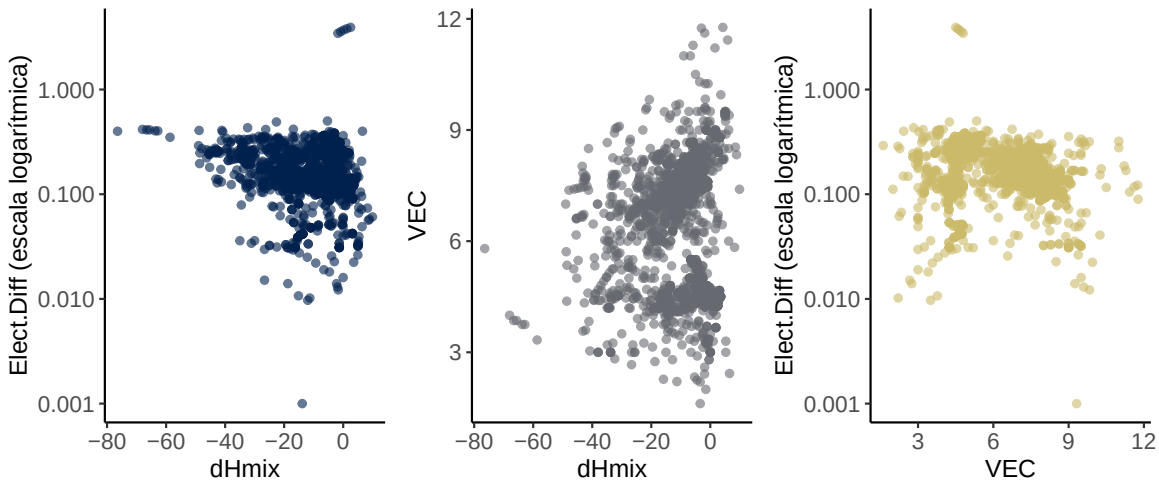

Figura 2: Nubes de dispersión de las variables numéricas. Elect.Diff es transformada por la función logaritmo con desplazamiento de 0.001.

Análisis de datos categóricos

La variable categórica en el dataset es `Phase`. A continuación los estadísticos descriptivos de esta variable (Tabla 2).

Tabla 2: Estadísticos descriptivos de la variable Phase.

variable	tasa de faltantes	n únicos
Phase	0.57%	80

Se aprecia que esta variable sí contiene valores faltantes, aunque marginales. Todos los valores faltantes tienen valores de `VEC` en el rango 7-8. Otra característica de esta variable es que tiene 80 niveles, equivalente a 30 aleaciones por `Phase`, en promedio.

Las 30 aleaciones más frecuentes se encuentran en la Tabla 3.

Analizar 80 niveles del factor `Phase` puede ser sobrecogedor. Por esta razón, se decide retener las categorías más frecuencias para simplificar el análisis. Tal como ejemplifica la Figura 3, las modas de la distribución de `VEC` son causadas por dos de los niveles más frecuentes: BCC y FCC.

Las distribuciones bien separadas convierten a `VEC` en candidato a predictor de la variable `Phase`.

Tabla 3: 30 Phases más frecuentes. Las Phases menos frecuentes fueron agrupadas en Otras; los valores faltantes se denotan NA.

Phase	n	Phase	n	Phase	n
BCC	501	BCC1 + BCC2	50	BCC + FCC	13
FCC	444	FCC+IM	47	FCC + FCC + BCC	13
IM	362	SS	24	BCC+FCC+B2	12
FCC + Im	216	BCC + FCC + Im	22	FCC+B2	12
AM	198	FCC + BCC + Im	21	BCC1 + BCC2 + Im	9
BCC + Im	126	BCC + Laves	17	SS+IM	9
Otras	106	FCC+BCC+IM	17	BCC + Im	8
FCC+BCC	57	BCC+B2	16	FCC + im	8
BCC+IM	51	BCC+FCC	16	FCC1+FCC2	8
FCC + BCC	51	NA	14	HCP	8

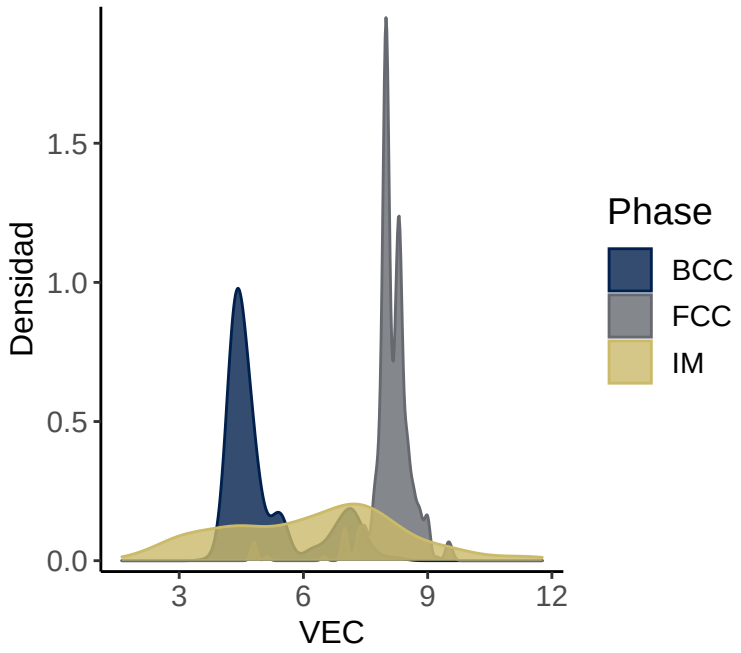

Figura 3: El factor que separa la distribución de VEC es la fase.