Centro de Investigación en Matemáticas (CIMAT). Unidad Monterrey

Proyecto 7: Un modelo estadístico basado en datos para predecir la temperatura crítica de un superconductor.

Victor Manuel Gómez Espinosa

Rafael Cruz Rodríguez

04 de junio de 2020

1. Objetivo e importancia del proyecto

Se busca reproducir los resultados obtenidos por (Hamidieh 2018), específicamente replicar los resultados del análisis multivariado de los datos, y sugerir una modificación en la implementación del modelo para predecir la temperatura critica de un superconductor.

Los superconductores, son materiales que conducen corriente con cero resistencia sólo a la temperatura critica o por debajo de esta, y son importantes ya que tienen (o podrían tener) muchas aplicaciones como en los sistemas de resonancia magnética, para mantener altos niveles magnéticos en el colisionador CERN e incluso podrían revolucionar la industria energética transportando y entregando energía con cero perdidas.

El modelo y teoría que prediga la temperatura critica de los superconductores es un problema que lleva desde 1911.

2. Introducción

Para este problema se cuenta con 2 conjuntos de datos (disponibles en http://archive.ics.uci.edu/ml/datasets/Superconductivity+Data) *train.csv* y *unique_m.csv* (Figuras 2.1 y 2.2 respectivamente), el primero consiste en 81 propiedades (Tablas 2.1 y 2.2) de 21263 super conductores y su correspondiente temperatura critica, mientras que el segundo conjunto contiene los elementos de los cuales esta compuesto cada super conductor, su fórmula química y su temperatura critica.

number_of_elements	mean_atomic_mass	wtd_mean_atomic_mass	gmean_atomic_mass	wtd_gmean_atomic_mass	entropy_atomic_mass	wtd_entropy_atomic_m
4	88.944468	57.862692	66.361592	36.116612	1.181795	1.062
5	92.729214	58.518416	73.132787	36.396602	1.449309	1.057
4	88.944468	57.885242	66.361592	36.122509	1.181795	0.975
4	88.944468	57.873967	66.361592	36.119560	1.181795	1.022
4	88.944468	57.840143	66.361592	36.110716	1.181795	1.129

rows × 82 columns

Figura 2.1: Conjunto de datos: *train.csv*

	H	He	Li	Be	B	C	N	O	F	Ne	...	Au	Hg	Tl	Pb	Bi	Po	At	Rn	critical_temp	material
0	0.0	0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	0	...	0.0	0.0	0.0	0.0	0.0	0	0	0	29.0	Ba0.2La1.8Cu1O4
1	0.0	0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	0	...	0.0	0.0	0.0	0.0	0.0	0	0	0	26.0	Ba0.1La1.9Ag0.1Cu0.9O4
2	0.0	0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	0	...	0.0	0.0	0.0	0.0	0.0	0	0	0	19.0	Ba0.1La1.9Cu1O4
3	0.0	0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	0	...	0.0	0.0	0.0	0.0	0.0	0	0	0	22.0	Ba0.15La1.85Cu1O4
4	0.0	0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	0	...	0.0	0.0	0.0	0.0	0.0	0	0	0	23.0	Ba0.3La1.7Cu1O4

5 rows × 88 columns

Figura 2.2: Conjunto de datos: *unique_m.csv.*

Variable	Units	Description
Atomic Mass	atomic mass units (AMU)	total proton and neutron rest masses
First Ionization Energy	kilo-Joules per mole (kJ/mol)	energy required to remove a valence electron
Atomic Radius	picometer (pm)	calculated atomic radius
Density	kilograms per meters cubed (kg/m^3)	density at standard temperature and pressure
Electron Affinity	kilo-Joules per mole (kJ/mol)	energy required to add an electron to a neutral atom
Fusion Heat	kilo-Joules per mole (kJ/mol)	energy to change from solid to liquid without temperature change
Thermal Conductivity	watts per meter-Kelvin (W/(m × K))	thermal conductivity coefficient κ
Valence	no units	typical number of chemical bonds formed by the element

Tabla 2.1: Propiedades del conjunto de datos: *train.csv*

Feature & Description	Formula
Mean	$= \mu = (t_1 + t_2)/2$
Weighted mean	$= \nu = (p_1 t_1) + (p_2 t_2)$
Geometric mean	$= (t_1 t_2)^{1/2}$
Weighted geometric mean	$= (t_1)^{p_1} (t_2)^{p_2}$
Entropy	$= -w_1 \ln(w_1) - w_2 \ln(w_2)$
Weighted entropy	$= -A \ln(A) - B \ln(B)$
Range	$= t_1 - t_2 \ (t_1 > t_2)$
Weighted range	$= p_1 t_1 - p_2 t_2$
Standard deviation	$= [(1/2)((t_1 - \mu)^2 + (t_2 - \mu)^2)]^{1/2}$
Weighted standard deviation	$= [p_1 (t_1 - \nu)^2 + p_2 (t_2 - \nu)^2)]^{1/2}$

Tabla 2.2: Medición para cada propiedad de la Tabla 2.1

Para la obtención de los datos *unique_m.csv*, (Hamidieh 2018) con base en las fórmulas químicas y temperaturas críticas, en R descomponen la formula en sus elementos (Figura 2.3) y obtienen sus propiedades (Tabla 2.2), previamente limpian la base de datos original y aplican directamente sobre estas propiedades un modelo de regresión lineal múltiple para utilizarlo como punto departida y comparación con otros métodos, se aplica también PCA sin lograr algún beneficio, y finalmente se aplica XGboost como el mejor método de los que implementan (Figura 2.4). Se utilizan las métricas de la raíz del error cuadrático medio (RMSE) y R^2 (R2).

Adicionalmente utilizan la opción de la similaridad del coseno para obtener formulas químicas parecidas y mejorar la predicción (Figura 2.5). Finalmente se menciona que no esperan buenas predicciones para super conductores recién descubiertos, así como también que en el futuro se podrían incluir otras características como la presión o la estructura del cristal para hacer mejores modelos.

```
RGui (64-bit) - [R Console]
File  Edit  View  Misc  Packages  Windows  Help

> makeup("NaCl")
Na Cl
 1  1
> makeup("CH4")
C H
1 4
> makeup("Yo975Yb0.025Ba2Cu3O")
     Yo      Yb      Ba      Cu       O
975.000   0.025   2.000   3.000   1.000
> makeup("Tm0.25Ba0.75Cu1OX")
  Tm   Ba   Cu    O    X
0.25 0.75 1.00 1.00 1.00
> makeup("Y1Ba2Cu3O7-Z")
 Y Ba Cu  O  Z
 1  2  3  1  1
Warning message:
In count.elements(formula) : NAs introduced by coercion
> makeup("SilV3")
Si  V
 1  3
> makeup("FCl")
 F Cl
 1  1
>
```

Figura 2.3: Descomposición de las fórmulas químicas en los elementos que la componen, utilizando R.

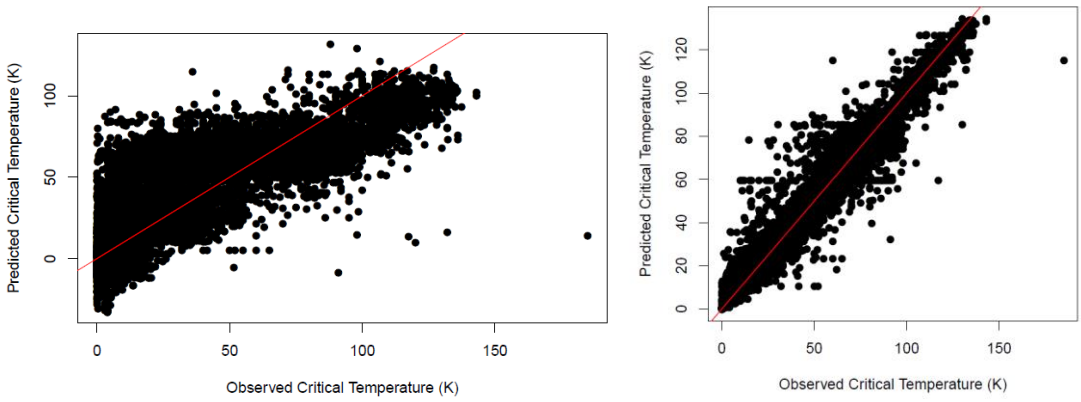

Figura 2.4: Modelo a) Regresión lineal múltiple. RMSE=17.58, R2=0.74, b) XGBoost RMSE=9.4, R2=0.92

```
RGui (64-bit) - [R Console]
File  Edit  View  Misc  Packages  Windows  Help

> predict_tc("Ba0.2La1.8Cu1O4", verbose = TRUE)
$prediction
[1] 24.44241

$info
     critical_temp           material
1            29.00     Ba0.2La1.8Cu1O4
934          28.00     La1.8Ba0.2Cu1O4
1053         31.00     La1.8Ba0.2Cu1O4
1277         25.60     La1.8Ba0.2Cu1O4
1798         21.00     La1.8Ba0.2Cu1O4
2272         23.50 La1.801Ba0.199Cu1O4
2342         20.90     La1.8Ba0.2Cu1O4
2907         32.50     La1.8Ba0.2Cu1O4
3533         16.50     La1.8Ba0.2Cu1O4
7041         17.90     La1.8Ba0.2Cu1O4
9684         38.00     La1.8Ba0.2Cu1O4
20338         9.38     La1.8Ba0.2Cu1O4
20653        23.40     La1.8Ba0.2Cu1O4

> predict_tc("MgB2")
[1] 35.50066
> predict_tc("Hg")
[1] 4.076086
```

Figura 2.5: Versión final del programa en R utilizando la similaridad del coseno.

Algunas observaciones importantes son que la mayoría de los superconductores contienen principalmente O (Figura 2.6) y para la mayoría, su temperatura critica es menor a los 50K (Figura 2.7).

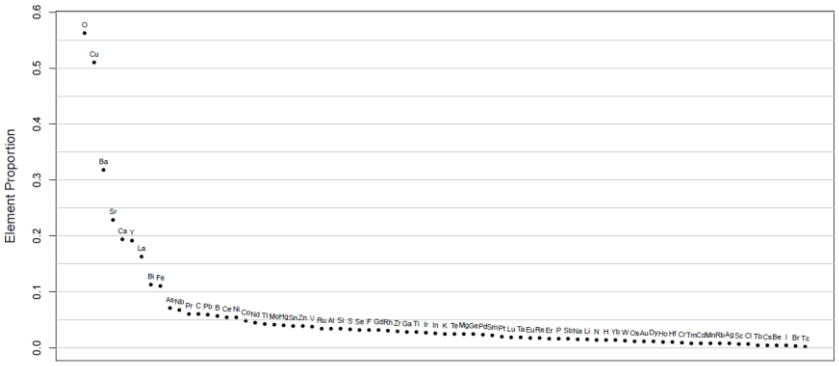

Figura 2.6: Proporción de los elementos en el conjunto de datos.

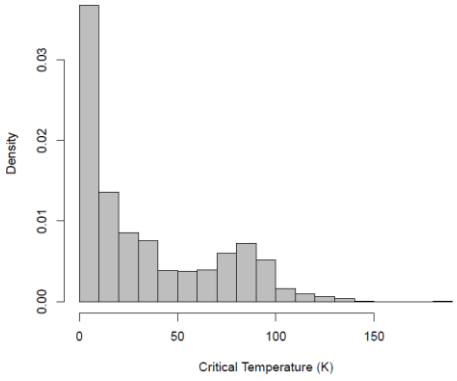

Figura 2.7: Distribución de las temperaturas críticas en los datos.

Las métricas a utilizar en este proyecto son la raíz de del promedio de los errores al cuadrado (RMSE) (1.1), R2 (1.2) y error absoluto medio (MAE) (1.3):

$$RMSE\left(y,\hat{y}\right)=\sqrt{\frac{1}{n}\sum_{i=0}^{n-1}\left(y_i-\hat{y}_i\right)^2} \tag{1.1}$$

$$R^2\left(y,\hat{y}\right)=1-\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2} \tag{1.2}$$

$$MAE\left(y,\hat{y}\right)=mediana\left(\left|y_1-\hat{y}_1\right|,...,\left|y_n-\hat{y}_n\right|\right) \tag{1.3}$$

3. Replica de los resultados de Hamidieh 2018

Para replicar los resultados del modelo de regresión lineal múltiple únicamente se utilizaron las 81 variables del conjunto de datos *train.csv* para ajustar el modelo y posteriormente utilizar este para predecir las temperaturas de toda la muestra (Figura 3.1).

Para las métricas RMSE y R2, se sigue un procedimiento que consiste en dividir aleatoriamente el conjunto de datos en conjunto de entrenamiento y de prueba en proporciones 2/3 y 1/3 respectivamente, ajustar el modelo con el conjunto de entrenamiento, utilizar el conjunto de prueba para predecir, obtener los indicadores (MSE y R2), repetir el procedimiento 25 veces, obtener el promedio de estos 25 indicadores, sacar la raíz a MSE (RMSE) y es lo que se reporta. Adicionalmente se incluyó también el error absoluto medio (MAE) debido a que es más robusto ante resultados atípicos.

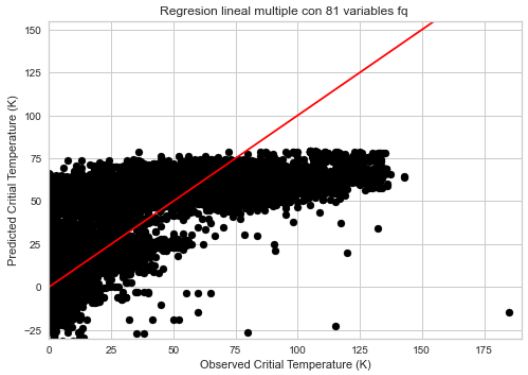

Figura 3.1: Modelo de regresión lineal múltiple con las 81 variables, RMSE=17.58, R2=0.74, MAE=10.21.

Posteriormente, aunque no se incluye en el trabajo de (Hamidieh 2018), se realiza un análisis del modelo y se observa que la gran mayoría de las variables no son significativas (Tabla 3.1) ya que varias de estas sus p valores son mayores a 0.2 y sus intervalos de confianza incluyen al 0, además de que el programa lanza la advertencia de colinealidad. También analizando los residuales, se nota que no siguen los supuestos, que la varianza no es constante y que los residuales no siguen una distribución normal (Figura 3.2).

	coef	std err	t	P>|t|	[0.025	0.975]
const	-20.8066	4.991	-4.169	0.000	-30.589	-11.025
number_of_elements	-3.4965	0.748	-4.674	0.000	-4.963	-2.030
mean_atomic_mass	0.8480	0.083	10.249	0.000	0.686	1.010
wtd_mean_atomic_mass	-0.9041	0.103	-8.773	0.000	-1.106	-0.702
gmean_atomic_mass	-0.5102	0.082	-6.226	0.000	-0.671	-0.350
wtd_gmean_atomic_mass	0.6468	0.098	6.625	0.000	0.455	0.838
entropy_atomic_mass	-35.9606	4.599	-7.819	0.000	-44.975	-26.946
wtd_entropy_atomic_mass	4.5545	3.638	1.252	0.211	-2.576	11.685
range_atomic_mass	0.2142	0.017	12.957	0.000	0.182	0.247
wtd_range_atomic_mass	0.0260	0.022	1.177	0.239	-0.017	0.069
std_atomic_mass	-0.5608	0.063	-8.936	0.000	-0.684	-0.438
wtd_std_atomic_mass	0.0905	0.055	1.656	0.098	-0.017	0.198
mean_fie	0.1601	0.063	2.526	0.012	0.036	0.284
wtd_mean_fie	-0.1787	0.077	-2.311	0.021	-0.330	-0.027
gmean_fie	-0.1524	0.063	-2.432	0.015	-0.275	-0.030
wtd_gmean_fie	0.1984	0.076	2.598	0.009	0.049	0.348
entropy_fie	-118.6988	20.191	-5.879	0.000	-158.274	-79.123

Tabla 3.1: P valores e intervalos de confianza para algunas de las 81 variables del modelo.

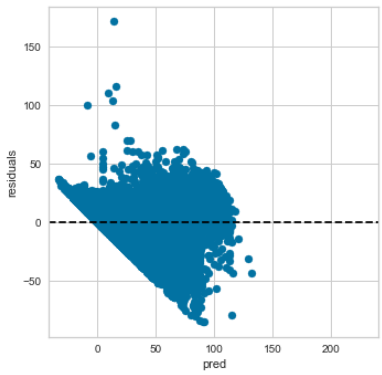

Figura 3.2: Residuales contra la predicción. Prueba de normalidad invariada Shapiro Test p-v=9.8e-45.

Finalmente, como se menciona que se intenta reducir la dimensión se utiliza PCA pero que se necesitan muchos componentes para obtener un resultado similar al anterior, por lo cual, aunque no se muestra se replica esto, encontrando que efectivamente con 6 componentes donde tenemos aproximadamente el 80% de la varianza acumulada el modelo no es mejor que el anterior (Figura 3.3).

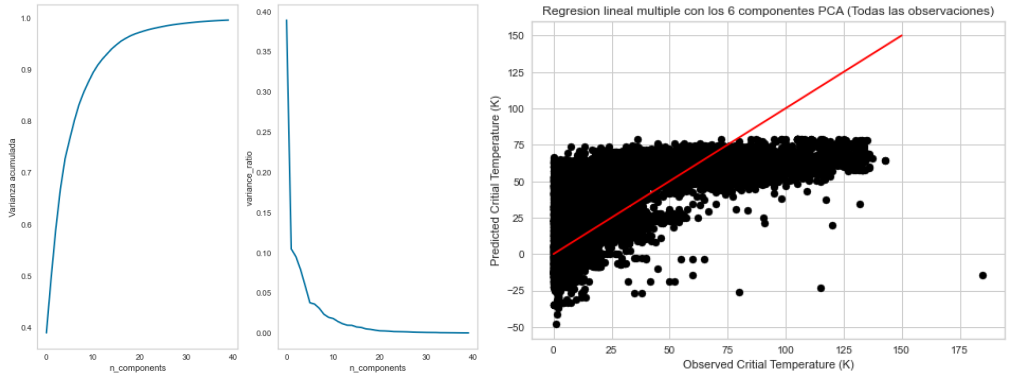

Figura 3.3: a) PCA Varianza acumulada y variance ratio. b) Modelo de regresión lineal múltiple con 6 componentes PCA. RMSE=22.63, R2=0.56, MAE=15.52

4. Modificación en la implementación

Primero, ante el problema de colinealidad se explora la matriz de correlación y se encuentra que efectivamente hay correlación fuerte entre algunas variables (Figura 4.1 a), también se busca probar si las variables siguen una distribución normal multivariada, lo cual es falso (Figura 4.1 b).

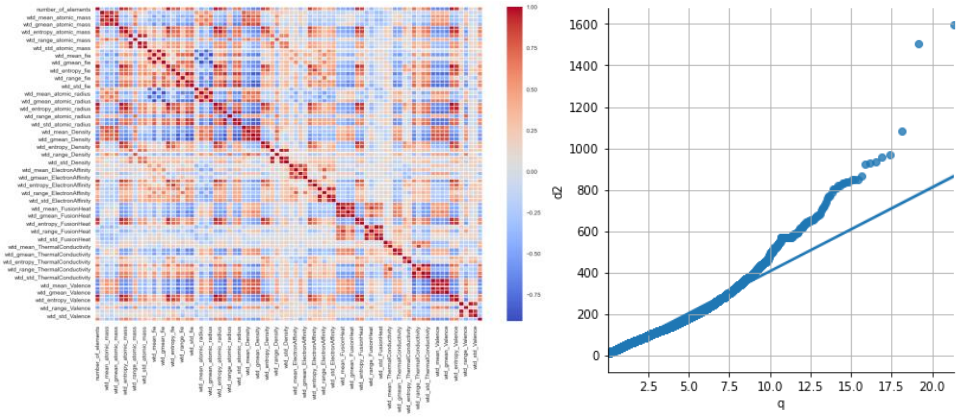

Figura 4.1 Para las 81 variables del conjunto de datos *train.csv*, a) Matriz de correlación. b) Gráfico chi-cuadrado donde se observan varios posibles datos atípicos, y prueba de normalidad multivariada Looney & Gulledge Test al 95% rechazada.

Ante los problemas anteriores, primero se aplicó la transformación de yeo-johnson (para evitar problemas con valores negativos) para aproximar los datos a normales ya que es uno de los supuestos para varias de las técnicas multivariadas, segundo como existe correlación entre varias variables y además son muchas variables, es deseable trabajar con menos variables para así quizá poder percibir patrones, y evitar el problema de colinealidad en el modelo, por lo cual se propone utilizar análisis de factores.

Para aplicar análisis de factores primero se aplicó la prueba de esfericidad de Barlett rechazando (p-val=0), se determinó el número de factores a utilizar mediante el criterio de la varianza acumulada y el grafico del codo (Figura 4.2) determinando 6 factores que es donde se encuentra el cambio en las pendientes y se cuenta con aproximadamente el 78% de la varianza acumulada.

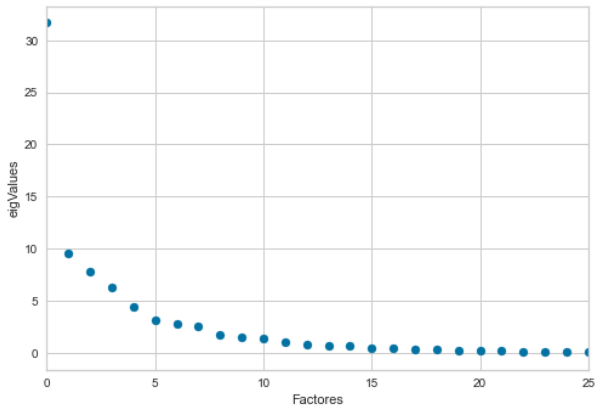

Figura 4.2: Valores propios contra el numero de factores. Varianza acumulada con 6 componentes aproximadamente 78%.

Posteriormente a los datos completos se les aplicaron los métodos por máxima verosimilitud (Figura 4.3) y por componentes principales (Figura 4.4) para 6 factores, después se dividió la muestra a la mitad aleatoriamente y se les aplico ambos métodos, se comparó si las soluciones coincidían y se eligió la más consistente, es decir la que no cambió mucho en lo general y principalmente en la interpretación de los factores por lo que finalmente se optó por el método de componentes principales. Por ejemplo, observe las cargas, los 2 cuadros superiores en la Figura 4.3, observe que los factores cambian para el método de máxima verosimilitud, mientras que para el método de componentes principales (Figura 4.4) la mayoría de los factores son consistentes. También otra razón para no elegir el método de máxima verosimilitud fue porque mostró problemas de convergencia y que la interpretación de los factores por el método de componentes principales parece más sencilla y consistente.

Para el método de componentes principales, el primer factor hace referencia a características atómicas principalmente la entropía, el segundo a la masa, tercero a la afinidad electrónica, cuarto al calor de fusión, quinto la energía de ionización y sexto la conductividad térmica.

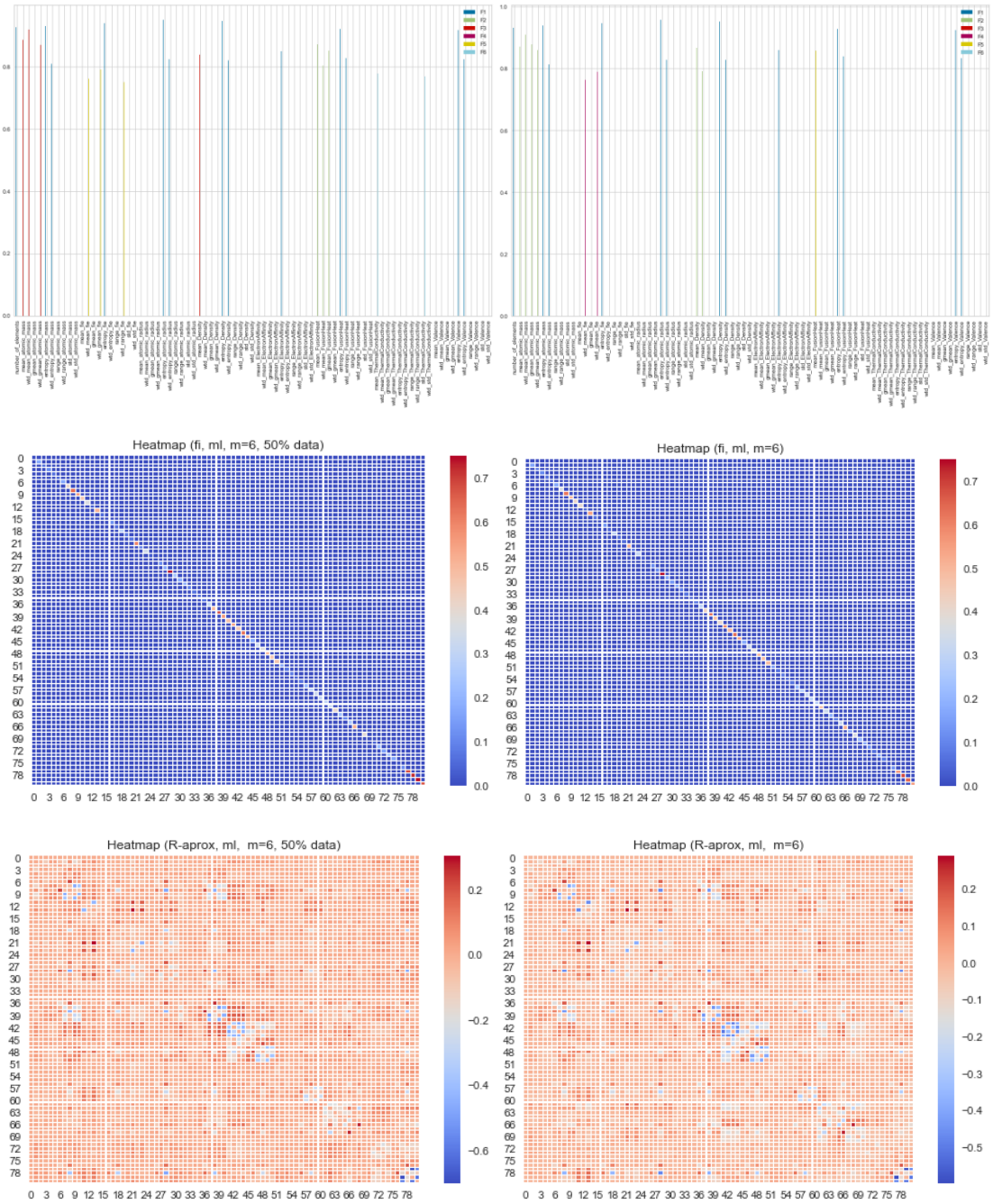

Figura 4.3: Soluciones $\left(L, \Psi, R - \hat{R}\right)$ por método de máxima verosimilitud, lado izquierdo para la mitad de los datos, lado derecho todos los datos.

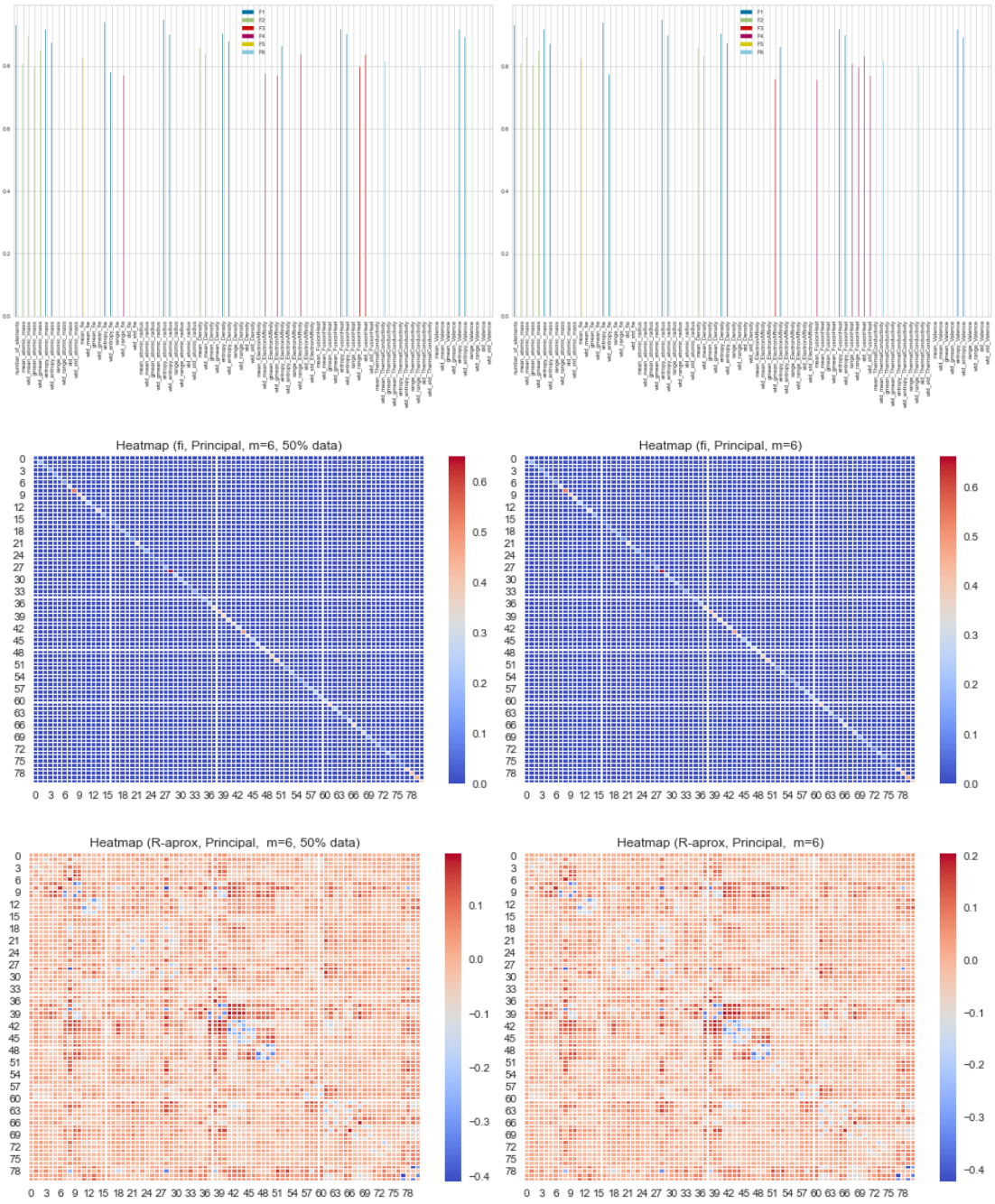

Figura 4.4: Soluciones $\left(L, \Psi, R - \hat{R}\right)$ por método de componentes principales, lado izquierdo para la mitad de los datos, lado derecho todos los datos.

En la Figura 4.5, se pueden observar en los primeros dos componentes las puntuaciones y la respectiva temperatura critica, donde puede observarse que parece que mientras mas al extremo derecho se encuentre del primer factor, parecen encontrarse los de mayor temperatura critica, pero con algunos datos atípicos que no siguen este comportamiento, por ejemplo aproximadamente por (-2,-4) se puede observar un punto con la mayor temperatura y se encuentra más cargado al extremo izquierdo del primer factor.

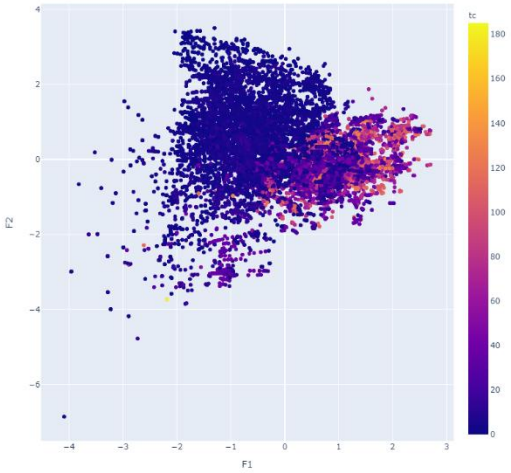

Figura 4.5: Proyección en los 2 primeros componentes de las puntuaciones, por análisis de factores.

Con esto, se volvió a ajustar el modelo y probarlo, sin embargo, los resultados nuevamente como en el caso de PCA no fueron satisfactorios, es decir se seguía pronosticando la temperatura muy por debajo de la real, por lo cual se trató de encontrar alguna razón para este comportamiento y se decidió hacer una exploración visual para tratar de encontrar algún patrón por lo que se decidió utilizar escalamiento multidimensional (MDS) para esta tarea.

Utilizando el conjunto de datos de las propiedades (*train.csv*) se obtuvo la matriz de distancias (euclidiana) y se utilizó el modelo clásico de MDS (porque debido a la cantidad de datos y el equipo de cómputo fue el único que logró realizarlo, previo a separar a la mitad los datos). Observe la Figura 4.6 que es el MDS en los primeros dos componentes con su correspondiente temperatura critica (y formula química), note que se aprecian aproximadamente 3 conjuntos similares con muchas posibles observaciones atípicas, observe que mientras mas a la derecha nuevamente parece que se encuentran las observaciones con mayor temperatura critica que las de lado izquierdo.

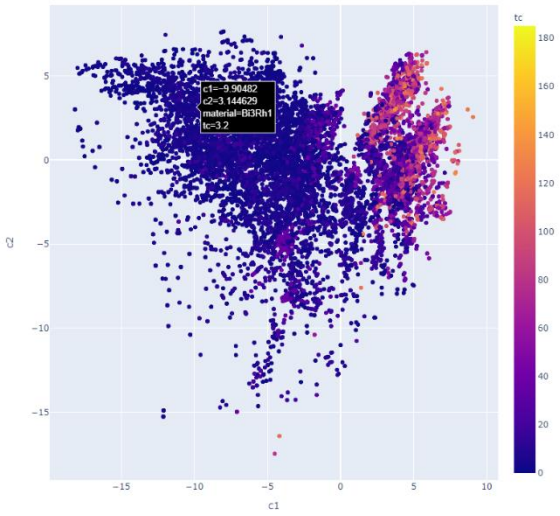

Figura 4.6: MDS clásico para los datos *train.csv* en los primeros dos componentes, con su temperatura crítica y formula.

Partiendo de esto, se trató de identificar a partir de las propiedades que tenían en común algunos de estos conjuntos, pero para tantas variables no fue tarea fácil (o muy obvia), una primera idea fue que se trataba del contenido de oxígeno (O), por lo cual se decidió intentar otro tipo de exploración visual pero ahora basado en los elementos, es decir utilizando el segundo conjunto de datos (*unique_m.csv*), para tratar de identificar algún patrón.

Para realizar lo anterior, se utilizó Kernel PCA con kernel coseno con 6 componentes, y al hacer una inspección en las proyecciones se encontraron patrones interesantes principalmente para los componentes 1,2 y 5. Observe la Figura 4.7 c) la cual representa la temperatura en el eje horizontal y en el eje vertical al primer componente, se pueden apreciar que se separa nuevamente en aproximadamente 3 conjuntos, donde los de mas arriba tienen menor temperatura que los que se encuentran abajo, también explorando por la formula química se observó que los de arriba no contenían oxigeno (O), mientras que los de en medio tenían a lo mucho 1 de oxígeno, y los de hasta abajo el contenido de oxigeno fue mayor a 1 (además se realizó prueba de diferencia de medias en estos conjunto y todas rechazan al 95%, Tabla 4.1), por lo cual el primer componente está relacionado al oxigeno (hacia los negativos hay más contenido de oxígeno) y al parecer esto esta relacionado con la temperatura critica. Por otro lado observe la Figura 4.7 b) que es la proyección de los componentes 5 y 1, nuevamente se observa la separación de los 3 conjuntos y en la parte superior de los que no contienen oxigeno se nota un dato atípico pues sobresale una observación en cuanto a temperatura, en cuanto al componente 5 la composición cambia, pero se puede notar revisando por formula que para algunos materiales con fórmulas químicas similares, la temperatura parece

incrementarse ligeramente si su composición o contenido de algún elemento cambia (principalmente oxigeno).

Lo anterior concuerda con lo observado en la Figura 4.7 a), que es la proyección en los primeros dos componentes, donde se observan diferentes grupos, algunos con menor o mayor contenido de cobre y oxígeno, otros con contenido de hierro con o sin oxigeno y materiales que no contienen oxígeno.

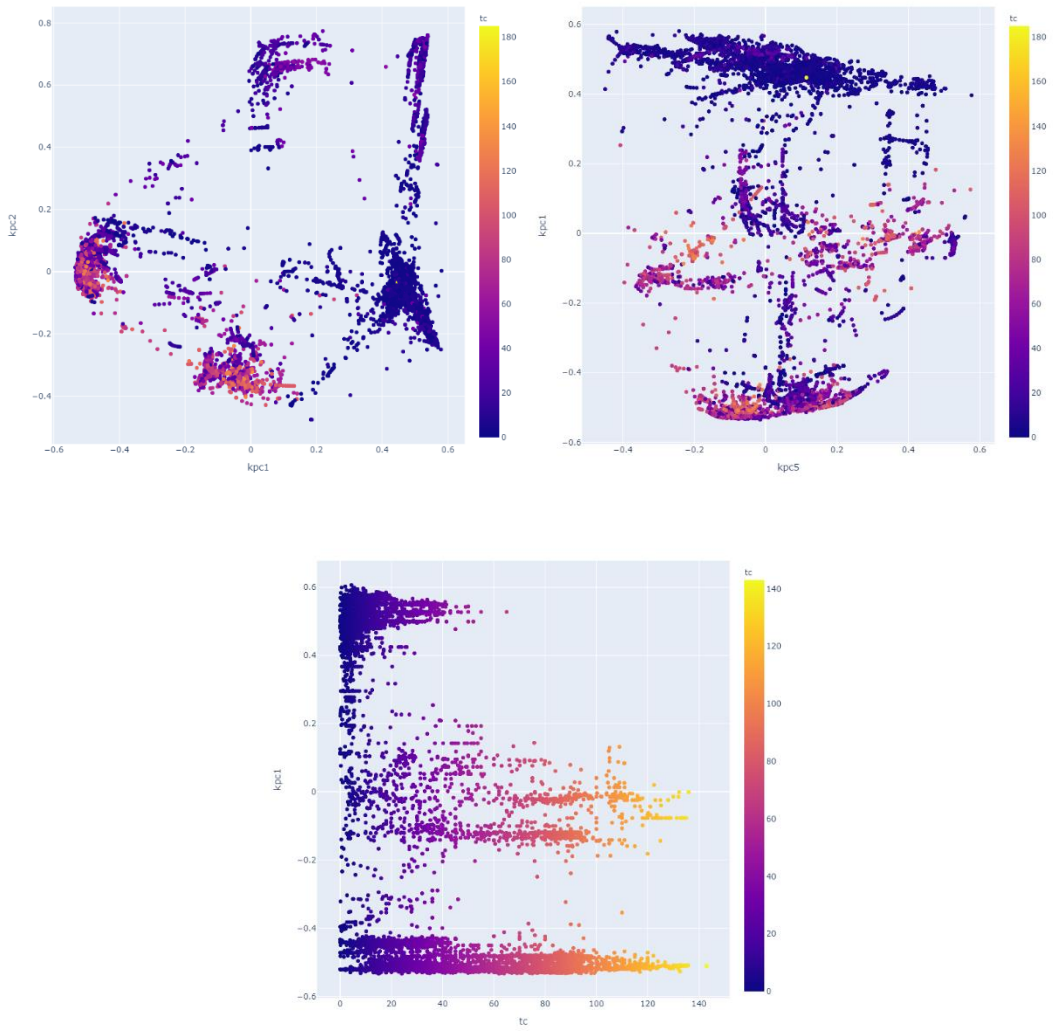

Figura 4.7: Kernel PCA con kernel coseno, proyecciones en los componentes a) 1,2, b) 5,1, c) temperatura critica, primer componente. Primer componente es representa el contenido de oxígeno.

Conjuntos	*Estadístico*	*Valor crítico*	*Resultado*
(O=0),(0<O<=1)	6435	113	Rechaza
(O>1),(0<O<=1)	6712	113	Rechaza
(O=0),(O>1)	338	113	Rechaza

Tabla 4.1: Prueba para probar diferencia de medias al 95% entre conjuntos con diferentes contenidos de oxígeno, bajo los supuestos de que las covarianzas son diferentes y el tamaño de muestra es mucho mas grande que el numero de variables. Todas rechazan.

Entonces según lo observado anteriormente, como hay muchos conjuntos según su composición (siendo quizá el mas importante el contenido de oxígeno), se decidió también como en el trabajo de (Hamidieh 2018) utilizar la similaridad del coseno para en base a un material de prueba, encontrar las 50 formulas químicas más parecidas dentro del conjunto de entrenamiento y sobre estas construir el modelo de regresión lineal múltiple utilizando los 6 factores principales.

Observe la Figura 4.8 que es el resultado de lo anteriormente descrito y note que ahora el modelo parece ser mucho mejor que el inicial, se observa que en general se ajusta bastante bien a la línea a 45 grados que representa que lo observado contra lo predicho coincide, claro se observan datos para los cuales o es menor o mayor y por lo tanto erróneo, pero si resultó mejor que la aproximación inicial. Sin embargo, un aspecto importante a destacar es que, haciendo el análisis del modelo para un material cualquiera, variando el número de fórmulas químicas similares para ajustar el modelo, resultó sensible a esto, ya que en algunos casos se tenían problemas de colinealidad, variables no significativas, datos atípicos, varianza no constante y los residuales no seguían una distribución normal univariada.

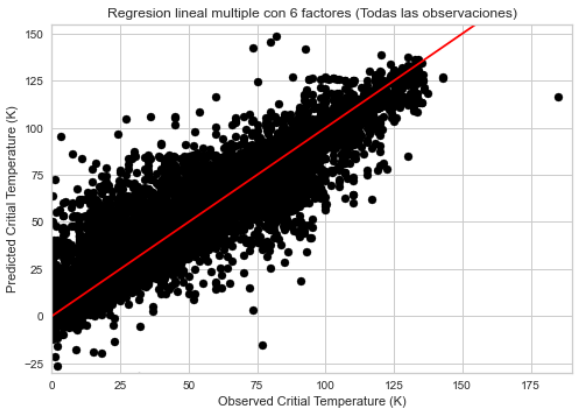

Figura 4.8: Regresión lineal múltiple para todo el conjunto de datos, utilizando la similaridad del coseno para encontrar las 50 formulas químicas más parecidas y ajustando el modelo con los 6 factores principales de esas formulas parecidas. MAE=2.61

5. Conclusiones

Después de analizar de diferentes formas los datos, tanto como por sus propiedades como por su composición química, es evidente que hay diferentes grupos de super conductores principalmente por su contenido de oxigeno (o sin este) y que los que contienen oxigeno tienden a tener temperaturas criticas mayores que los que no.

A su vez, dentro de estos grupos con diferentes contenidos de oxigeno parecen existir otros grupos con diferentes elementos, por ejemplo, los que contienen cobre o hierro, estos grupos tienden a sobresalir del resto y parecen tener temperaturas criticas mayores.

Al mismo tiempo esto no se cumple para todos, puesto que se observaron datos atípicos que por ejemplo perteneciendo a la categoría sin oxígeno y mostraron una temperatura critica muy alta, cercana a 180K.

Todo esto puede explicar porque la aproximación por la similaridad del coseno con los principales factores para ajustar el modelo de regresión lineal múltiple resultó mas preciso que los modelos con las 81 variables o con PCA para todo el conjunto de observaciones (Tabla 5.1).

Es posible que el modelo de regresión múltiple no sea el más adecuado y que existan mejores opciones (¿XGBoost?) o quizá como se menciona en el trabajo original no se estén considerando algunas propiedades (¿presión?) que puedan ser de mayor importancia para predecir la temperatura critica de un superconductor.

	MAE
Regresión lineal múltiple (RLM)	10.21
RLM con PCA	15.52
RLM con 6 factores y similaridad coseno	2.61

Tabla 5.1: Comparación de los modelos con su error MAE (error medio absoluto).

6. Bibliografía

Hamidieh, K. (2018), "A data-driven statistical model for predicting the critical temperature of a superconductor," *Computational Materials Science*, 154, 346–354. https://doi.org/10.1016/j.commatsci.2018.07.052.