UNICA
UNIVERSITÀ DEGLI STUDI
DI CAGLIARI

INFERENTIAL ANALYSIS OF THE CRITICAL TEMPERATURE IN SUPERCONDUCTORS WITH SPARK ON DOCKER

ESAME

LABORATORIO DI BIG DATA

SOMMARIO

INTRODUZIONE....2
CONTESTO DI ANALISI....2
DATASET....3
INFRASTRUTTURA HARDWARE-SOFTWARE....4
ESPLORAZIONE....5
ELABORAZIONE CON SPARK....6
ANALISI INFERENZIALE....7
RISULTATI....7
CONCLUSIONI....12
BIBLIOGRAFIA....13

INTRODUZIONE

Nel presente elaborato viene proposta un'analisi su Big Data per mezzo di un'infrastruttura Spark, utilizzando i modelli della libreria Python dedicata per effettuare elaborazioni di machine learning su un computer-cluster simulato in container Docker.

I dati analizzati riguardano i materiali superconduttori, con particolare riferimento a variabili circa le loro proprietà chimico-fisiche e la loro struttura molecolare. L'obiettivo è quello di individuare, tramite appositi modelli inferenziali, quali tra queste risultino maggiormente significative per spiegare la variabilità della temperatura critica dei materiali stessi, parametro di riferimento per quanto riguarda l'applicabilità del superconduttore in contesti pratici.
L'intera analisi viene effettuata tramite Apache Spark e la rispettiva libreria Python *pyspark*, distribuendo il carico di lavoro tra diversi container Docker creati al fine di simulare il comportamento di un computer-cluster reale.

CONTESTO DI ANALISI

Il contesto tecnico-scientifico di riferimento per l'analisi è quello dei materiali superconduttori.
La superconduttività è un fenomeno fisico che si esplica nella resistenza elettrica nulla e nell'espulsione del campo magnetico dall'interno di alcuni materiali. Questi, chiamati "superconduttori", manifestano il suddetto comportamento se posti sotto uno specifico valore di temperatura detto "temperatura critica". Tipicamente, si tratta di valori particolarmente bassi: uno dei migliori superconduttori (cuprato di mercurio, bario e calcio) presenta una temperatura critica di circa 133K (−140 °C)[1]. La sfida dei ricercatori è quindi quella di creare materiali che si comportino come superconduttori a temperature più elevate possibili, al fine di avvicinarsi alla temperatura ambiente per poterne facilitare l'applicazione a scopi ingegneristici e di ricerca.

Con l'analisi statistica si intende osservare, partendo dalle informazioni contenute nel dataset di riferimento, come varia la temperatura critica del materiale in base alla sua composizione chimica.
Vengono dunque proposti due modelli, uno di apprendimento supervisionato (Ridge/Lasso Regression) e uno di apprendimento non supervisionato (PCA, Principal Component Analysis), ritenuti adatti ad individuare le relazioni che legano le variabili a disposizione con quella di interesse.

L'aspetto tecnico intende invece simulare l'elaborazione in un computer-cluster, creando l'apposita infrastruttura software necessaria all'esecuzione in parallelo dei modelli sopra citati, avvalendosi di Apache Spark e Python per la parte di gestione software e di Docker per la parte di gestione hardware.

DATASET

Il dataset oggetto di analisi è "Superconductivity Data[2]", reperito sul portale web UC Irvine Machine Learning Repository. Contiene due file in formato *.csv*, ciascuno dei quali presenta dati raccolti su 21263 materiali superconduttori:

- **train.csv:** ai fini dell'analisi rinominato con il nome più significativo di "**superconductivity.csv**", contiene 21263 occorrenze e 82 colonne, ciascuna delle quali presenta informazioni sulla consistenza atomica dei materiali e sul loro comportamento dal punto di vista fisico-chimico
- **unique_m.csv:** ai fini dell'analisi rinominato con il nome più significativo di "**molec_structure.csv'**, contiene 21263 occorrenze e 88 colonne, ciascuna delle quali rappresenta uno specifico elemento e la quantità dello stesso presente in ogni materiale

Tutte le feature di interesse sono di tipo quantitativo.

Nella seguente tabella viene sintetizzato il significato delle variabili presenti in *superconductivity.csv*:

VARIABILE	UNITÀ DI MISURA	DESCRIZIONE
Number of elements	Integer	Numero di elementi
Atomic Mass	Atomic mass units (AMU)	Somma delle masse a riposo di protoni e neutroni
First Ionization Energy	Kilo-Joule per mole (kJ / mol)	Energia necessaria per rimuovere un elettrone di valenza
Atomic Radius	Picometro (pm)	Raggio atomico
Density	Kilogrammi per metro cubo (kg/m^3)	Densità a temperatura e pressione standard
Electron Affinity	Kilo-Joule per mole (kJ / mol)	Energia liberata quando viene aggiunto un elettrone a un atomo neutro, isolato e allo stato gassoso
Fusion Heat	Kilo-Joule per mole (kJ / mol)	Energia necessaria alla transizione da solido a liquido senza variazioni di temperatura
Thermal Conductivity	Watt per metro-Kelvin (W/(m*K))	Conduttività termica
Valence	Nessuna unità di misura	Numero tipico di legami chimici formati dagli elementi
Critical temperature	Gradi Kelvin (K)	Temperatura critica associata al superconduttore

Tabella 1: significato delle variabili contenute nel dataset superconductivity.csv (alias di train.csv)

Ognuna delle precedenti variabili è espressa con i seguenti indici di posizione e variabilità: media, media ponderata, media geometrica, media geometrica ponderata, entropia, entropia ponderata, range, range ponderato, deviazione standard, deviazione standard ponderata.
Una panoramica approfondita delle formule di calcolo si può trovare nella documentazione associata al dataset.

Il significato delle colonne presenti nel dataset *molec_structure.csv* è invece il seguente:

VARIABILE	DESCRIZONE
Elementi	Colonne (86) relative alla quantità di singoli elementi
Material	Sintesi della formula chimica del materiale
Critical temperature	Temperatura critica associata al superconduttore

Tabella 2: significato delle variabili contenute nel dataset molec_structure.csv (alias di unique_m.csv)

Il significato delle sigle dei singoli elementi chimici, ciascuno per ogni colonna, sono disponibili nel sito Lenntech.it[3].

Per evitare problemi di compatibilità con le librerie Java, riscontrabili in alcune configurazioni nell'utilizzo del metodo di caricamento di file *.csv* di *pyspark*, si è deciso di importare il dataset con la libreria *pandas* e poi fornire il dataframe come argomento del metodo *SparkSession.createDataframe*.

INFRASTRUTTURA HARDWARE-SOFTWARE

L'infrastruttura hardware-software proposta è intesa a simulare un computer-cluster di 4 unità, ciascuna delle quali è rappresentata da un container Docker appositamente inizializzato. Ogni container rappresenta un *workernode*, ciascuno connesso al *masternode* tramite Spark aperto in rete locale dalla piattaforma principale, sulla quale sono eseguiti gli stessi container.

L'architettura dell'infrastruttura è la seguente:

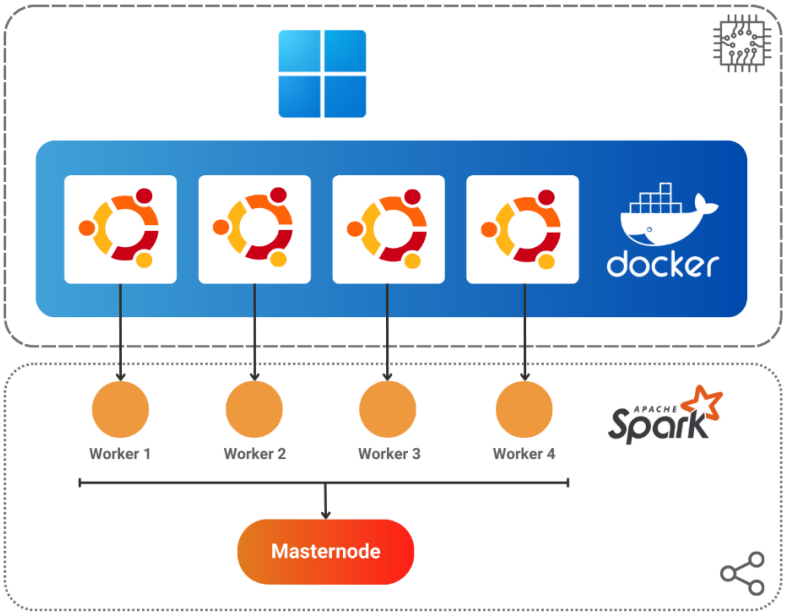

Figura 1: architettura del computer-cluster

Nella piattaforma principale Windows, vengono istanziati 4 container Docker, ciascuno dei quali esegue un'immagine di Ubuntu Linux accessibile tramite shell. Su ciascun container viene installato ed eseguito Spark, connettendo ogni *workernode* al *masternode* precedentemente istanziato dalla piattaforma principale. In questo modo, si simula un computer cluster con quattro unità Linux connesse come *workernode*.

ISTANZIAMENTO E CONNESSIONE DEI NODI

Per istanziare il *masternode* su Windows, si naviga dal terminale nella directory "*./SPARK_HOME/bin*" e si esegue il seguente comando:

```
spark-class org.apache.spark.deploy.master.Master
```

Nel terminale sarà fornito automaticamente il link all'interfaccia web di Spark.

Per poter creare le istanze dei *workernode*, è necessario installare Docker Desktop, dal quale si può gestire l'esecuzione dei container tramite interfaccia. Successivamente, si crea un container su immagine Ubuntu Linux, automaticamente scaricata dai server Docker, alla quale si accede tramite terminale di Windows.
Dalla shell del container si installano Spark e le sue dipendenze, nonché la stessa versione di Python 3 presente nella piattaforma Windows dalla quale si avvierà il nodo master (in questo caso, Python 3.11).
Una volta installato il necessario, si naviga nella directory "*./SPARK_HOME/sbin*" e si esegue il seguente comando:

```
sudo ./start-worker.sh spark://<IP>:<port>
```

dove IP e port sono l'indirizzo IPv4 e la porta del *masternode*.

Dall'interfaccia web di Spark è possibile monitorare sia i nodi e le sessioni connesse che le elaborazioni effettuate:

Figura 2: screenshot dell'interfaccia web di Spark con un workernode connesso da container Docker

ESPLORAZIONE

Si è deciso di esplorare graficamente le relazioni tra le variabili di risposta e la temperatura critica. Tale analisi esplorativa viene effettuata, vista la natura dei dati, solo sul dataset principale (*superconductivity.csv*).
Sono dunque stati prodotti 9 grafici, ciascuno per una delle macro-caratteristiche presenti nel dataframe. Avendo a disposizione diverse misure per ogni caratteristica, per il confronto è stata scelta la media geometrica ponderata.

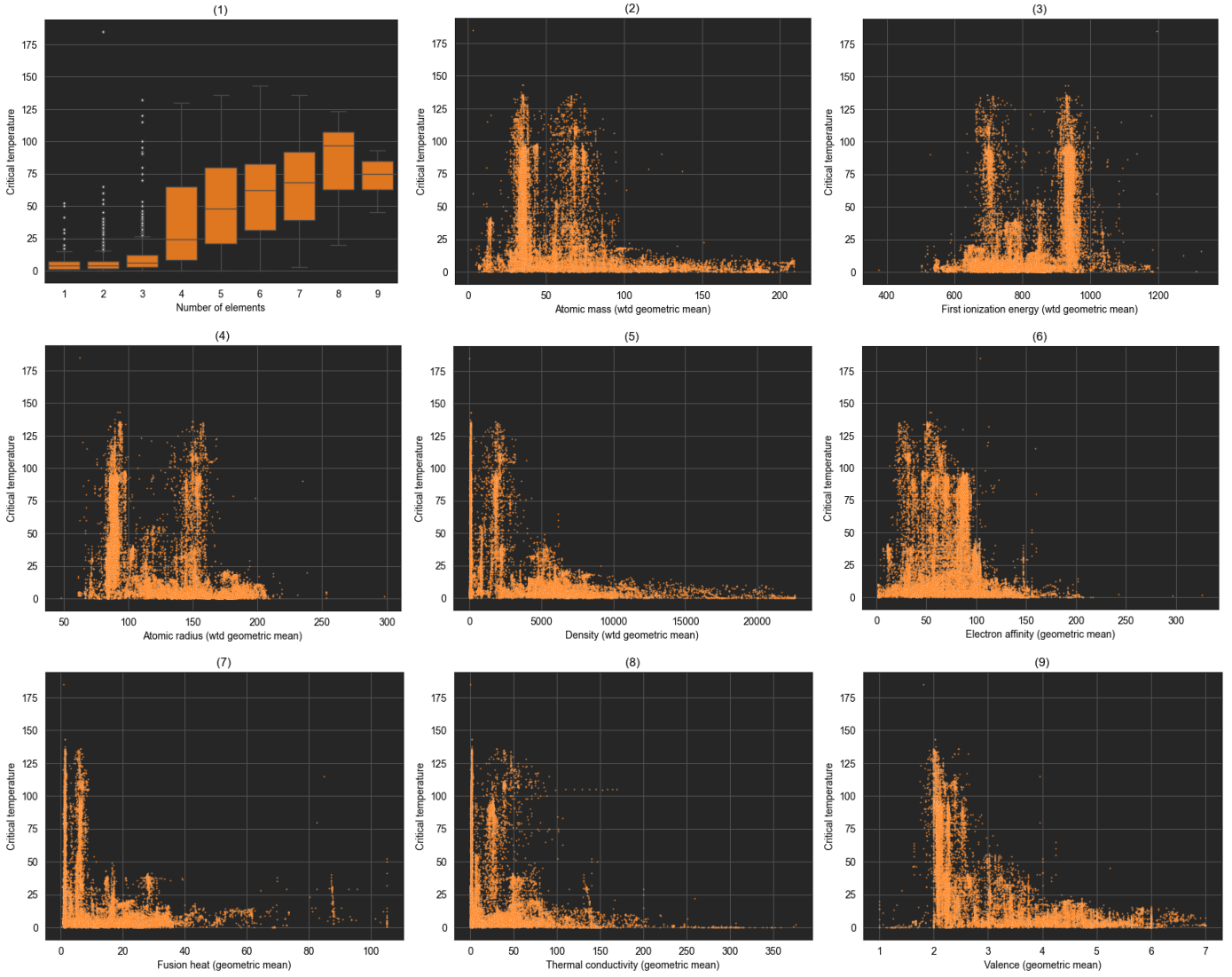

Figura 3: serie di scatterplot che mostrano la relazione tra la temperatura critica e le variabili: number of elements (1), atomic mass (2), fia (3), atomic radius (4), density (5), electron affinity (6), fusion heat (7), thermal conductivity (8), valence (9)

Si osserva che le feature presentano pattern simili nella forma della distribuzione, tendente in alcuni casi a normalità (es. fia) o spostate sulla coda di sinistra. Interessante risulta essere la relazione tra la temperatura critica e il numero di elementi, descritta nel boxplot (grafico 1).

Il dataset è stato creato calcolando diversi indici per ogni macro-variabile, poiché utile ai fini dell'analisi della superconduttività dei materiali dal punto di vista chimico-fisico. È dunque lecito attendersi una relazione di collinearità tra le variabili.

Ad esempio, si osservano i seguenti plot delle matrici di correlazione tra gli indici di *atomic mass* e *electron affinity*:

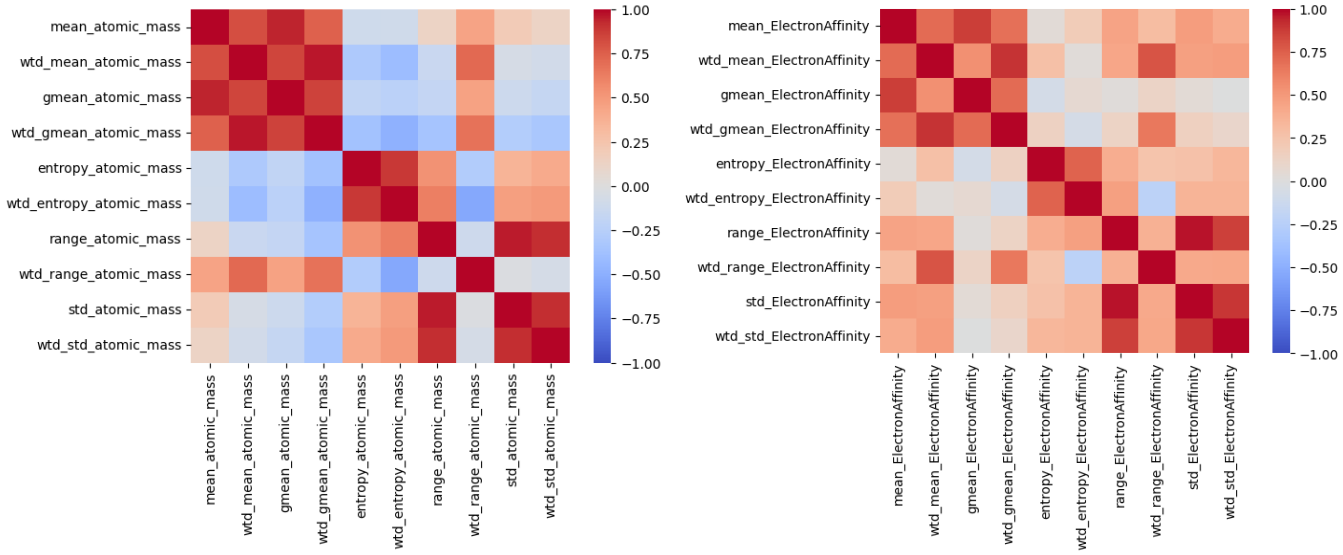

Figura 4: corrplot degli indici di atomic mass (sinistra) e electron affinity (destra)

Si può constatare una certa correlazione tra le diverse medie calcolate e tra gli indici di variabilità. Per evitare perdite di informazione dovute alla rimozione delle variabili e allo stesso tempo diminuire l'influenza della collinearità, si decide di utilizzare regressione con regolarizzazioni L1 e L2 per inferire le relazioni tra la temperatura critica e le variabili medesime.

ELABORAZIONE CON SPARK

Una volta istanziato il *masternode* e connessi i *workernode*, si collega una sessione pySpark e si eseguono le elaborazioni esposte nei notebook Jupyter associati al presente report. Per verificare la riuscita della parallelizzazione del lavoro nel cluster, ecco un esempio estratto dall'elaborazione della funzione *ClusterHandler.fit_lr()*, la quale si occupa di eseguire il tuning della Linear Regression in Cross Validation:

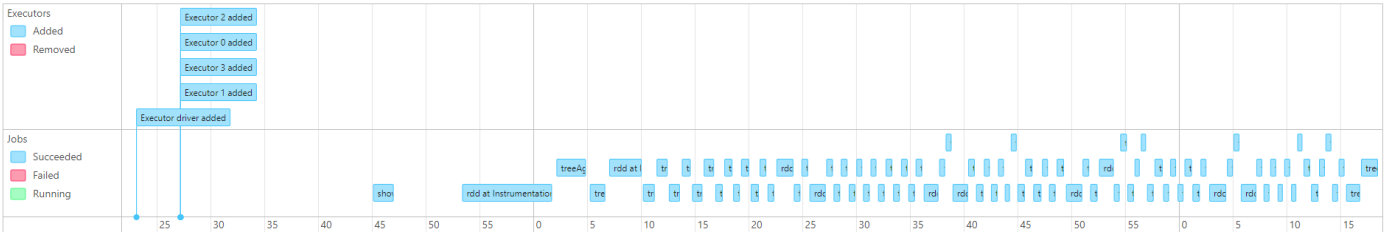

Figura 5: screenshot della timeline di elaborazione di Spark di una Ridge Regression con 5-fold cross-validation

Nello screenshot della timeline osserviamo prima l'assegnazione effettuata da Spark dei 4 worker connessi, poi le specifiche elaborazioni in cross validation, sia per quanto riguarda l'addestramento per diversi valori del parametro *reg_param*, che per le successive operazioni di calcolo della metrica di riferimento.

Di seguito, il DAG (Directed Acyclic Graph) che mostra il flusso delle operazioni di partizionamento, parallelizzazione e mappatura effettuate per uno dei task appena descritti:

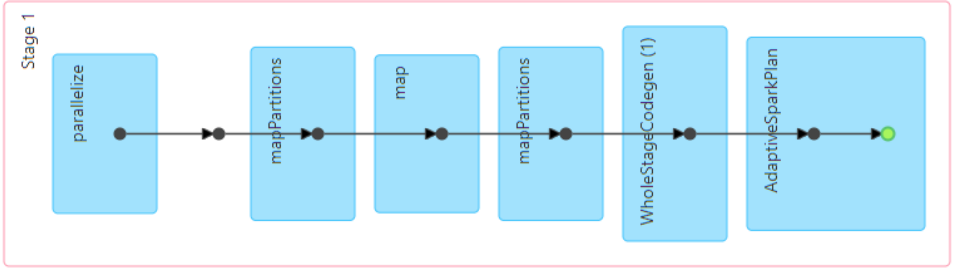

Figura 6: screenshot del flusso di lavoro per uno specifico task effettuato da Spark

ANALISI INFERENZIALE

Per formalizzare le relazioni tra le variabili di risposta, si decide di elaborare due modelli:

- **Ridge/Lasso Regression:** algoritmo di apprendimento supervisionato facilmente interpretabile, adatto a dataset interessati da collinearità delle variabili. Si avvale di un parametro, solitamente identificato con λ, per ridurre l'influenza dei coefficienti meno significativi, avvicinandone il valore allo zero. Tale parametro è selezionato sulla base della metrica R^2, scegliendo il modello che presenta il valore maggiore (più vicino a 1) e che quindi è meglio in grado di spiegare la variabilità del fenomeno. Il secondo parametro che viene selezionato è Elasticnet Param, il quale permette di controllare il comportamento della regolarizzazione che viene applicata. Se 0, il modello applica una regolarizzazione L2 (ridge), se 1 applica la regolarizzazione L1 (lasso) e se 0.5 una combinazione di entrambe (elasticnet). Per garantire robustezza all'analisi, viene eseguito l'addestramento con una 5-fold Cross Validation.
- **PCA - Principal Component Analysis:** algoritmo di apprendimento non supervisionato utile a suddividere la variabilità del fenomeno in *n* dimensioni, al fine di ridurre la dimensionalità del dataset iniziale e poter apprezzare l'influenza delle feature per ogni dimensione generata. È stato scelto un numero di dimensioni pari a 5. Si analizza dunque se le componenti principali estratte mantengono una relazione con la variabile di risposta. Essendo l'analisi preposta ad analizzare la variabilità del fenomeno con riferimento alla temperatura critica, si decide di eliminare quest'ultima dal dataset nel quale viene eseguita la PCA, richiamandola in sede di plot dei risultati per apprezzare la distribuzione della stessa nelle diverse dimensioni e, al contempo, evitando che la variabilità di quest'ultima vada a sovrapporsi alle altre feature.

Entrambi i modelli sono stati eseguiti sia su *superconductivity.csv* che su *molec_structure.csv* (in notebook separati). Prima della loro esecuzione, è stata eseguita un'operazione di regolarizzazione delle variabili.

RISULTATI

SUPERCONDUCTIVITY.CSV

Di seguito la sintesi dei risultati del tuning del parametro di regolarizzazione della regressione, eseguita sul dataset *superconductivity.csv*, approssimati alla terza cifra decimale:

REGULARIZATION PARAMETER	ELASTICNET PARAMETER	R^2
0.001	**0.0**	**0.735**
0.001	0.5	0.709
0.001	1.0	0.710
0.01	0.0	0.734
0.01	0.5	0.708
0.01	1.0	0.709
0.1	0.0	0.728
0.1	0.5	0.707
0.1	1.0	0.701
0.5	0.0	0.714
0.5	0.5	0.689
0.5	1.0	0.674
1.0	0.0	0.706
1.0	0.5	0.671
1.0	1.0	0.650

Tabella 3: risultati del tuning del parametro di regolarizzazione della Linear Regression

La combinazione con l'R2 maggiore è con un lambda di 0.001 e regolarizzazione L2.
Vengono dunque analizzati i coefficienti del miglior modello.

Di seguito, le 15 variabili che presentano i coefficienti più alti (per valore assoluto) estratti dal modello di regressione:

FEATURE	COEFFICIENT
entropy_fie	-84.793
wtd_entropy_Valence	-69.179
entropy_Valence	67.221
entropy_atomic_radius	51.914
wtd_entropy_fie	43.375
wtd_entropy_atomic_radius	42.576
entropy_atomic_mass	-35.762
wtd_entropy_FusionHeat	24.891
wtd_std_Valence	-23.222
wtd_gmean_Valence	-20.959
wtd_entropy_ElectronAffinity	-20.703
wtd_entropy_Density	-18.888
entropy_FusionHeat	-18.475
wtd_mean_Valence	16.844
entropy_Density	16.567

Tabella 4: coefficienti associati alle 15 feature più influenti (Regression)

Effettuata la PCA, si estraggono i seguenti valori di variabilità spiegata dalle 5 dimensioni del modello:

DIMENSION	EXPLAINED VARIANCE
0	0.389
1	0.105
2	0.095
3	0.079
4	0.059
Total	0.727

Tabella 5: variabilità spiegata dalle 5 dimensioni della PCA

Estraendo i coefficienti associati alle feature nella prima dimensione (migliori 7 per valore assoluto), quella che spiega maggiormente la variabilità, si osserva che sono associati alle stesse variabili dei migliori del modello precedente:

FEATURE	COEFFICIENT
range_fie	-0.1639
wtd_std_fie	-0.1635
range_atomic_radius	-0.1635
wtd_entropy_atomic_radius	-0.1634
wtd_std_atomic_radius	-0.1604
entropy_Valence	-0.1587
entropy_fie	-0.1566

Tabella 6: coefficienti associati alle 7 feature più influenti della prima dimensione (PCA)

Analizzando le macro-variabili, sebbene siano tendenzialmente distribuite tutte in ciascuna dimensione, è possibile individuare i seguenti “topic” basandosi sulle feature più frequenti e con i coefficienti più alti:

DIMENSION	CHARACTERISTICS	MOST IMPORTANT MACRO-VARIABLES
0	Atomic structure	first ionization temperature, atomic radius, valence
1	Atomic structure	atomic mass, atomic radius
2	Energetic and thermal properties	thermal conductivity, valence
3	Energetic properties	electron affinity
4	Thermal properties	thermal conductivity, fusion heat

Tabella 7: topic delle 5 dimensioni della PCA

Volendo associare le dimensioni alla temperatura critica, si esegue un plot delle singole osservazioni sulle dimensioni della PCA, colorando i valori in base al valore della variabile di risposta:

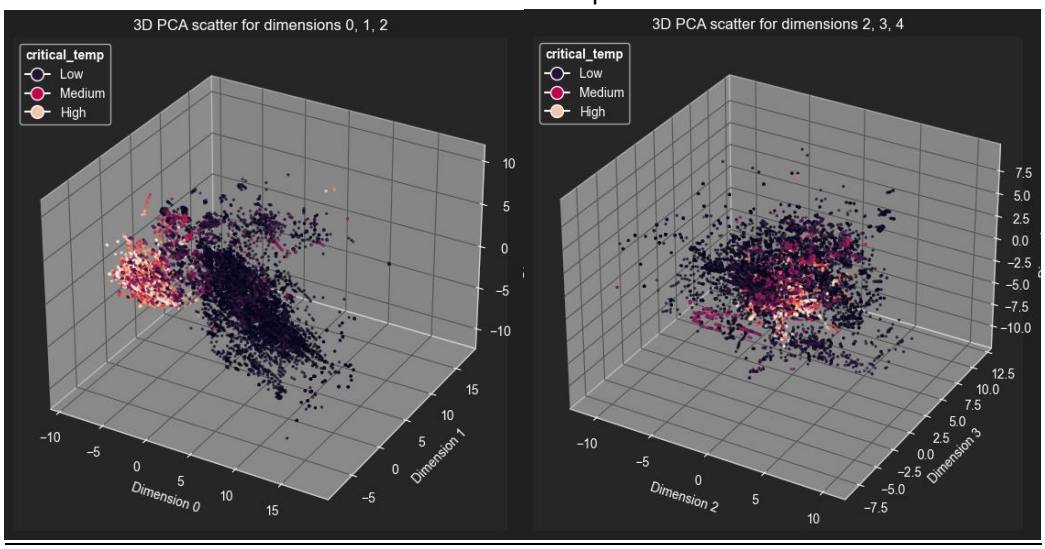

Figura 6: 3D plot delle dimensioni della PCA, rispettivamente [0,1,2] (sinistra) e [2,3,4] (destra)

Per apprezzare meglio l'effetto associato alla temperatura critica delle prime due dimensioni, specialmente riguardo alla prima, si eseguono dei plot 2D:

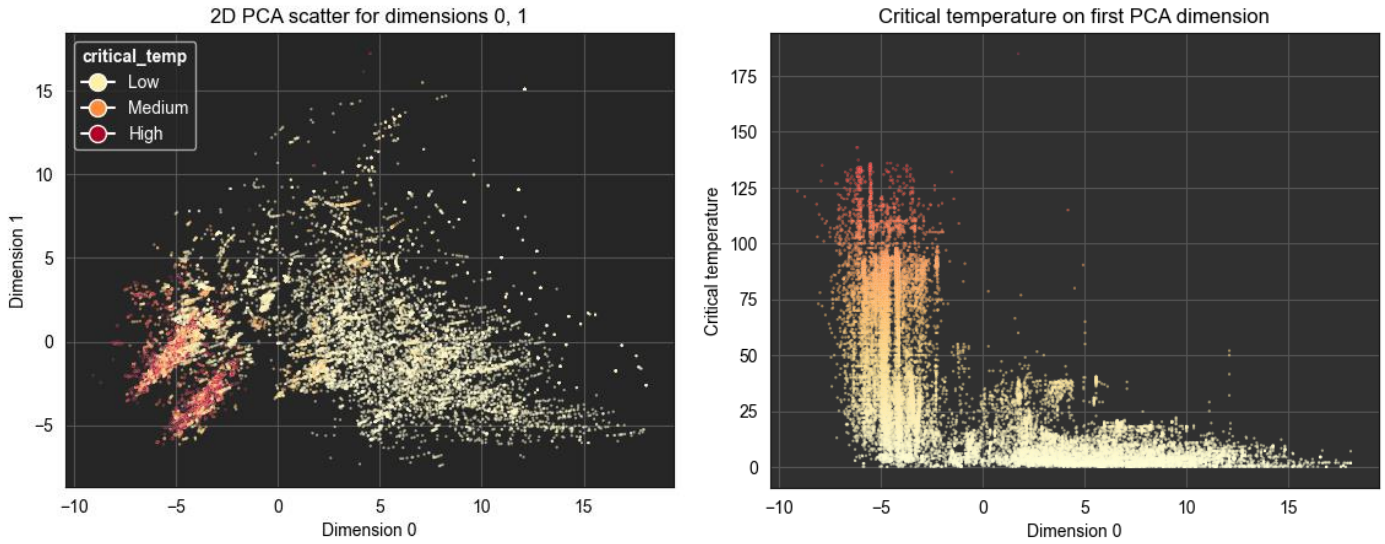

Figura 6: 2D plot delle prime due dimensioni della PCA (sinistra) e della temperatura critica sulla dimensione 0 (destra)

Mentre i superconduttori con una temperatura critica relativamente bassa risultano essere distribuiti eterogeneamente nella metà destra del grafico, quelli a conduttività media e alta sono concentrati nel primo quarto di sinistra, con valori della dimensione zero compresi tra -10 e 0. Meno efficace appare la medesima valutazione effettuata sulla dimensione 1, in quanto valori inferiori a 5 per questa dimensione sono associati a temperature critiche sia alte che basse.

MOLEC_STRUCTURE.CSV

Si analizzano ora i risultati del tuning del parametro di regolarizzazione lambda della regressione, eseguita sul dataset *molec_structure.csv*, approssimati alla quinta cifra decimale:

REGULARIZATION PARAMETER	ELASTICNET PARAMETER	R^2
0.001	0.0	0.52291
0.001	0.5	0.52297
0.001	1.0	0.52304
0.01	0.0	0.52285
0.01	0.5	0.52353
0.01	1.0	0.52417
0.1	0.0	0.52225
0.1	0.5	0.52753
0.1	1.0	0.53214
0.5	0.0	0.51967
0.5	0.5	0.54230
0.5	1.0	0.55370
1.0	0.0	0.51669
1.0	0.5	0.55183
1.0	**1.0**	**0.56894**

Tabella 8: risultati del tuning del parametro di regolarizzazione della Linear Regression

La combinazione migliore si ottiene con un lambda di 1.0 e regolarizzazione L1.
Di seguito, le 9 variabili che presentano i coefficienti più alti (per valore assoluto) estratti dal modello di regressione:

FEATURE	COEFFICIENT
Bario (Ba)	18.938
Mercurio (Hg)	11.821
Stronzio (Sr)	10.672
Carbonio (Ca)	5.811
Tallio (Tl)	4.334
Cerio (Ce)	-3.850
Silicio (Si)	-2.707
Bismuto (Bi)	2.082

Coefficienti associati alle 9 feature più influenti (Regression)

In particolare i primi quattro elementi, sono molto usati per la creazione di semiconduttori. Il coefficiente indica una relazione positiva tra la temperatura critica e la presenza degli elementi stessi, il che riflette la scelta dei ricercatori di utilizzarli per le loro proprietà chimiche.

Eseguendo la PCA, non si ottengono invece risultati entusiasmanti. Tuttavia, possono essere fatte alcune considerazioni basate sulla classificazione chimica dei superconduttori. Si estraggono dunque i seguenti valori di variabilità spiegata dalle 5 dimensioni del modello, approssimati alla terza cifra decimale:

DIMENSION	EXPLAINED VARIANCE
0	0.036
1	0.033
2	0.026
3	0.025
4	0.024
Total	0.144

Tabella 10: variabilità spiegata dalle 5 dimensioni della PCA

Osservando tuttavia le variabili ordinate per valore assoluto dei coefficienti, è possibile delineare alcuni pattern riconducibili alla classificazione chimica dei semiconduttori.

Per la dimensione 0, gli elementi più influenti risultano essere l'ossigeno (O), il rame (Cu), il bario (Ba) e l'ittirio (Y), i quali vengono comunemente assemblati nella creazione di superconduttori a base di rame detti "cuprati"[4].
Per la dimensione 1, troviamo come elementi principali arsenico (As), ferro (Fe) e platino (Pt), tipicamente combinati in superconduttori a base metallica, come i superconduttori a base di ferro[5].
Infine, un altro pattern interessante appare tra le variabili più influenti della dimensione 2, contenente elementi quali carbonio (C), rubidio (Rb), potassio (K) e idrogeno (H), ampiamente utilizzati per la creazione dei c.d. fullereni[6].

Di seguito, lo scatterplot 3D delle osservazioni sulle dimensioni della PCA, colorate per temperatura critica:

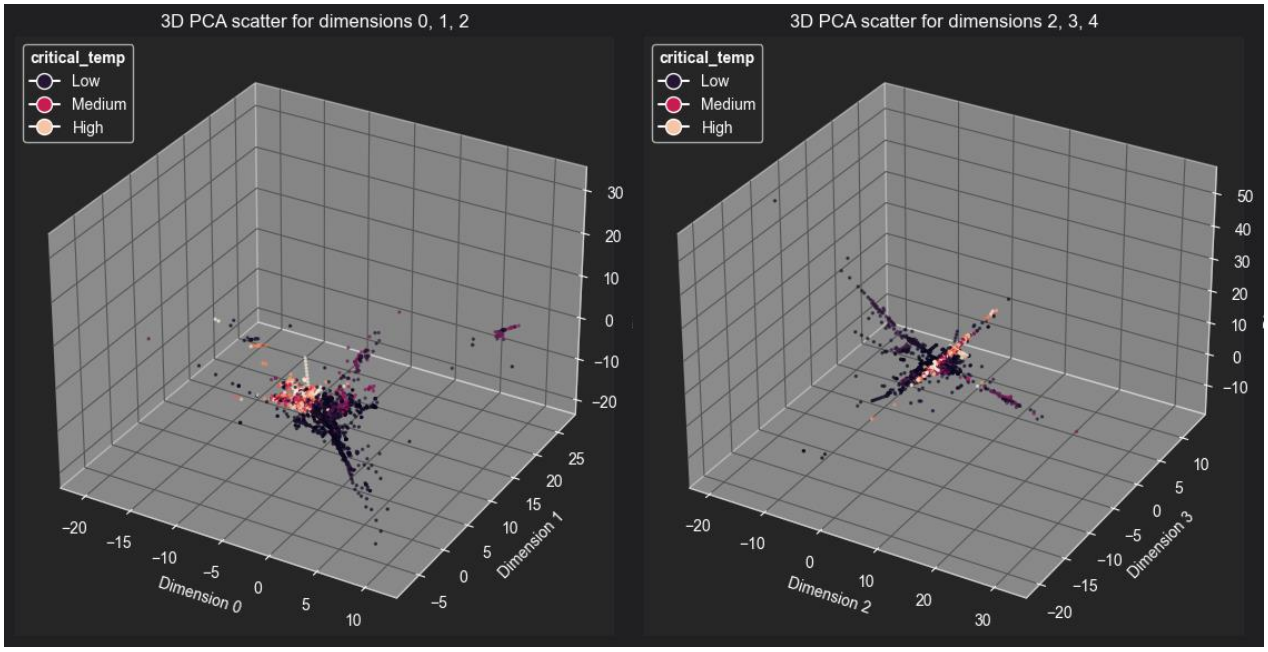

Figura 11: 3D plot delle dimensioni della PCA, rispettivamente [0,1,2] (sinistra) e [2,3,4] (destra)

Sebbene le dimensioni non riescano a spiegare sufficientemente la variabilità, è interessante osservare come, specialmente sulla dimensione 0, le osservazioni siano divise in due gruppi abbastanza distinti per temperatura critica bassa e medio-alta.
Per apprezzare maggiormente tale osservazione, si esegue uno scatterplot 2D sulle prime due dimensioni:

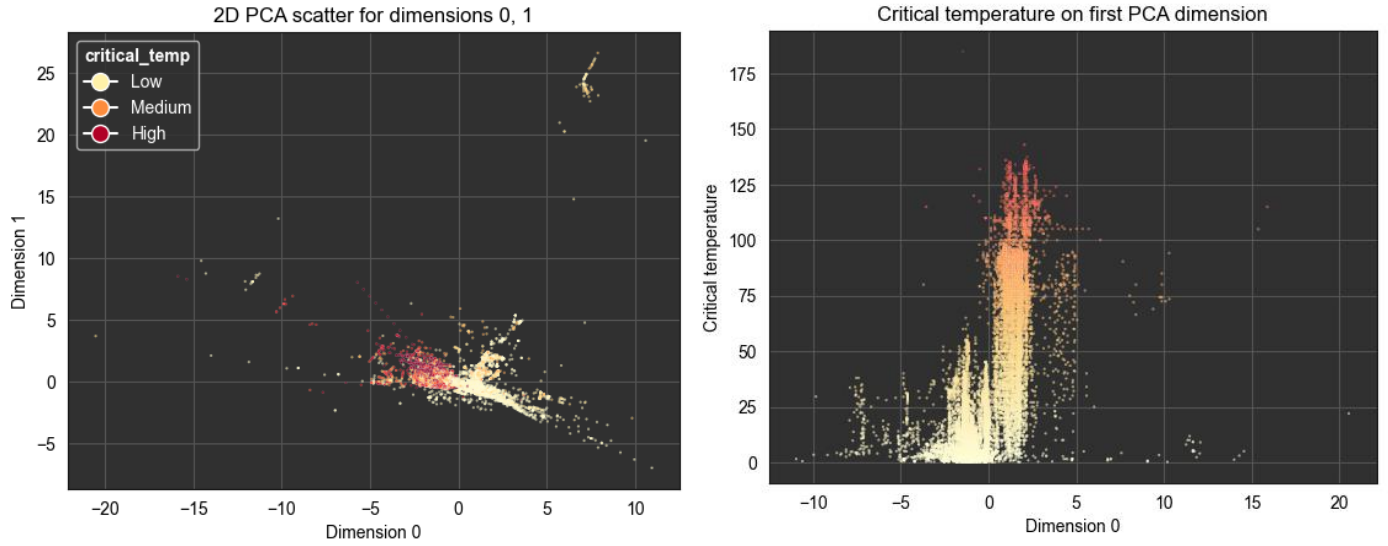

Figura 12: 2D plot delle prime due dimensioni della PCA (sinistra) e della temperatura critica sulla dimensione 0 (destra)

Si può concludere che la struttura molecolare dei materiali non rappresenta un'informazione sufficiente per spiegare la variabilità della temperatura critica ma, tuttavia, questa riflette le scelte di tendenza tra i ricercatori del settore nella costruzione di materiali superconduttori.

CONCLUSIONI

Utilizzando Apache Spark e Docker è stata effettuata un'elaborazione di machine learning su un computer-cluster simulato, distribuendo il carico tra i diversi container Docker atti a simulare diversi *workernode*, collegati alla stessa sessione attraverso il coordinamento di un *masternode*.

L'analisi, effettuata su un dataset di Big Data e consistente in una combinazione di due metodi, uno di apprendimento supervisionato e l'altro di apprendimento non supervisionato, ha permesso di osservare come le informazioni a disposizione nei due dataset, attinenti rispettivamente alle caratteristiche e proprietà atomiche e alla composizione molecolare di materiali superconduttori, permettano di tracciare una panoramica generale su quelli che sono i fattori che incidono sulla temperatura critica dei materiali medesimi.

BIBLIOGRAFIA

1. “Superconduttività ad alte temperature”, it.wikipedia.org
2. Hamidieh Kam (2018), “Superconductivty Data”, UCI Machine Learning Repository, https://doi.org/10.24432/C53P47
3. "Elementi chimici elencati in ordine alfabetico”, www.Lenntech.it
4. Encicplopedia online, “Cuprati”, www.treccani.it
5. Enciclopedia online, “Superconduttori a base di ferro”, www.treccani.it
6. Enciclopedia online, “Fullereni”, www.treccani.it