UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI BIOLOGIA

Corso di Laurea in Biotecnologie

ELABORATO DI LAUREA

Deep Eutectic Solvents e Intelligenza Artificiale: Approcci di Deep Learning per la Progettazione e la Caratterizzazione

Tutor
Prof. Sergio Rampino
Dipartimento di Scienze Chimiche

Laureando: *Pietro Dalla Pozza*

ANNO ACCADEMICO 2024/2025

Indice

1 Stato dell'arte **1**

2 Metodologia **4**

3 Risultati **11**

Bibliography **16**

Sommario

I solventi organici, fondamentali in numerosi settori industriali, sono purtroppo tossici ed inquinanti. Una recente alternativa è rappresentata dai Deep Eutectic Solvents, una miscela di due o più componenti che si presenta liquida a temperatura ambiente. Il problema è che esistono potenzialmente infinite combinazioni per creare DES, e la ricerca di quello "perfetto" per una determinata applicazione richiede molto tempo e denaro. L'obbiettivo di questo lavoro è quello di utilizzare delle tecniche di deep learning per predire le temperature di fusione di DES, date le loro composizioni e rapporti molari. Come base di partenza abbiamo sfruttato 8 modelli già pubblicati, utilizzati per essere ri-addestrati con un maggiore dataset di dati. Una ulteriore analisi è stata l'esclusione della temperatura di fusione dal dataset di training, per osservare la variazione di R2 e RMSE, a dispetto di un minore tempo di calcolo. Gli algoritmi qui addestrati (Linear Regression e Decision Tree) hanno mostrato indici di precisione leggermente migliori (5-8%) ai risultati precedentemente pubblicati. Questo lavoro da noi iniziato può essere continuato in futuro, espandendo ulteriormente il set di dati, o utilizzando l'intelligenza artificiale per predire anche altri valori chiave come densità e viscosità.

1 Stato dell'arte

I solventi organici sono sostanze di vitale importanza per gran parte delle aziende globali. Industrie farmaceutiche, mediche, petrolchimiche, ecc. utilizzano infatti questi composti per eseguire passaggi chiave della produzione, come ad esempio la sintesi e purificazione di fibre e materie plastiche, o l'estrazione e separazione dei principi organici dalle piante. Tuttavia, già da molto tempo, diverse ricerche hanno evidenziato come questi solventi organici non siano solamente tossici, ma anche molto costosi e dannosi se rilasciati nell'ambiente, provocando danni a lungo termine agli animali e agli esseri umani [1].

A partire dagli anni '90, con la crescente attenzione alla tutela dell'ambiente, all'interno della comunità scientifica si diffuse sempre di più il concetto di "green chemistry": un nuovo approccio alla chimica che implica l'utilizzo di solventi ecocompatibili e in quantità più moderate, ai fini di ridurre l'impatto ecologico e rendere la chimica più sostenibile, ma senza comprometterne l'efficienza o la qualità. Diversi gruppi di ricerca quindi si sono specializzati nel trovare nuovi solventi in grado di essere biodegradabili, riciclabili, non tossici e anche facilmente isolabili dai componenti estratti.

Nel 2003, un gruppo di ricerca inglese [2] scoprì per la prima volta una nuova tipologia di "green solvents": i deep eutectic solvents (DES). Questi DES, composti da una miscela di due o più sostanze in proporzione ben precisa, rappresentano un'alternativa più sostenibile, economica e meno pericolosa rispetto ai classici solventi organici. Solitamente, i DES binari (ovvero a due componenti) sono composti da un accettore (HBA) e da un donatore di legami idrogeno (HBD). I DES terziari e quaternari sono invece sviluppati utilizzando più tipi diversi di HBA e HBD. A seconda della natura chimica delle sostanze di partenza, essi sono categorizzati in 5 macro tipologie [Tab. 1], dove la III, formata dalla miscela di un sale quaternario e un HBD, è la classe più semplice, economica e meno tossica da preparare, e quindi più studiata e diffusa delle altre. Un vantaggio fondamentale dei DES è rappresentato dal fatto che sono liquidi a temperatura ambiente, anche se i singoli componenti presentano temperature di fusione maggiori. Variando poi le concentrazioni dei vari

Type	Formula	
I	Cat^+X^- zMClx	(M = Zn, Sn, Fe, Al, Ga, In)
II	Cat^+X^- zMClx yH2O	(M = Cr, Co, Cu, Ni, Fe)
III	Cat^+X^- zRZ	(Z = $CONH_2$, COOH, OH)
IV	MCl_x + RZ	(M = Zn, Al and Z = $CONH_2$, OH)
V	Non-ionic	

Tabella 1: *Tipologie di DES*

costituenti, varia anche la temperatura di fusione, fino a raggiungere il punto eutettico, dove il rapporto tra le componenti è tale da rendere la temperatura di fusione del DES il minimo possibile [Fig. 1].

I DES sono considerati più ecosostenibili rispetto ai classici solventi convenzionali, tuttavia, è importante sottolineare come l'impatto ambientale dei DES sia strettamente collegato alla loro composizione specifica. Ad esempio, alcuni DES potrebbero mostrare ugualmente un certo grado di tossicità o una limitata biodegradabilità [3]. Nonostante questi problemi, la caratteristica peculiare dei DES è che sono finemente regolabili (scelta dei componenti di base e le loro relative concentrazioni), e questa adattabilità permette di costruire alcuni "green solvents" applicabili in ambito scientifico ed industriale.

Al giorno d'oggi, questi Deep Eutectic Solvents vengono già sfruttati per la biocatalisi e sintesi organiche, sintesi di nanoparticelle, in metallurgia ed elettrodeposizione [4]. L'introduzione graduale dei DES nel mondo industriale però è molto lenta e faticosa. Il problema principale è rappresentato dal fatto che esistono un numero troppo elevato di combinazioni e concentrazioni possibili per formare DES, e scegliere la miscela esatta in base al caso specifico di applicazione richiede molto tempo e spesso non si arriva nemmeno a raggiungere il risultato più ottimale. Le proprietà chimico-fisiche dei solventi infatti determinano il loro campo di utilizzo. Nel nostro caso, uno dei parametri più cruciali dei DES è rappresentato proprio dalla temperatura di fusione, che ci mostra come, dato un certo rapporto specifico di componenti, se la miscela è liquida a temperatura ambiente, perciò utilizzabile. A questo parametro poi si aggiungono anche viscosità e densità, poiché sono direttamente coinvolti nella progettazione dei processi tecnici. Queste tre proprietà quindi riflettono in

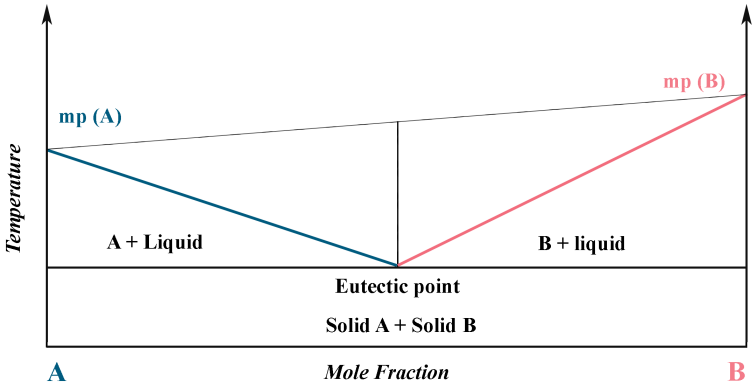

Figura 1: Rappresentazione schematica di un diagramma di fase di un DES a due componenti [5]

linea generale l'efficacia dell'estrazione del solvente.

La "generative chemistry", un campo di ricerca che adopera l'intelligenza artificiale generativa per creare molecole con determinate proprietà chimiche, è diventata sempre più sfruttata in tutti gli ambiti scientifici di interesse chimico-biologico. Sue applicazioni includono già la predizione e generazione di proteine, molecole organiche e inorganiche, la progettazione di farmaci e materiali bioattivi. Inoltre, studi recenti hanno evidenziato come l'intelligenza artificiale possa predire le proprietà chimico-fisiche dei DES. Tuttavia, i risultati concreti di questi sono veramente pochi, la predizione sulle temperature di fusione è molto spesso imprecisa, e tutti questi studi si limitano a testare dei gruppi ristretti di DES, senza quindi avere un algoritmo universale in grado di spaziare tra tutte le classi. Di conseguenza, usare questi modelli per una predizione generale di DES risulta poco efficiente.

La svolta è avvenuta nel 2023, in cui un gruppo di ricercatori [6], dopo aver appreso sperimentalmente le caratteristiche di alcuni DES, ha costruito un dataset composto da 402 elementi, riportando per ogni solvente la temperatura di fusione, così come la composizione percentuale delle sue componenti. Utilizzando questi dati, sono riusciti ad addestrare un modello self-attention autoregressivo in grado di predire e generare DES e le loro proprietà chimiche, ottenendo per la prima volta risultati promettenti e aprendo la strada a nuove frontiere. A questo, ha fatto seguito una nuova pubblicazione del 2024 [7],

in cui un gruppo di ricercatori ha ampliato questa analisi testando 8 diversi modelli ML (machine learning) per trovare il più preciso ed efficiente.

In quest'ultimo caso, anche il dataset usato è stato più grande, sfruttando un totale di 2303 righe di dati per studiare le temperature di fusione (e 4369 per densità/ 4216 per la viscosità, dato che questa ricerca non si limita solo alla temperatura), ottenuti mediante un'accurata ricerca bibliografica a partire da esperimenti del 2003. Il dataframe risultante è molto eterogeneo, infatti i tipi di HBA e HBD, così come la tipologia di DES risultanti, sono molto diversi tra di loro. Nonostante il risultato finale ottenuto non sia tra i più precisi (il modello più preciso, ovvero CatBoostRegressor, ha ottenuto un valore R2 di cross-validazione di 0,76 per la temperatura), forse dovuto proprio alla disomogeneità dei dati utilizzati, questo approccio sembra comunque molto incoraggiante per futuri miglioramenti.

In questo lavoro infatti prenderemo come riferimento i modelli della ricerca del 2024, e procederemo dapprima a testarli per provare a replicare i risultati; dopodiché proseguiremo applicando un dataset più ampio e omettendo alcuni campi di dati (come la temperatura di fusione delle componenti), per studiare il comportamento dei vari modelli e cercare di massimizzare il risultato finale.

2 Metodologia

Raccolta Dati

Il primo passaggio del lavoro è stata la creazione di un nuovo dataset "originale", più esteso dei precedenti dataframe già esistenti, in grado di studiare ed analizzare con precisione le temperature di fusione dei componenti e dei DES risultanti Nel nostro caso il dataset sarà quindi una fusione di dati dei due articoli precedentemente descritti, dato che la maggior parte degli elementi del primo dataframe non sono presenti anche nel secondo. Il primo articolo ci fornisce un file con 401 elementi (contro i 402 dichiarati nella pubblicazione), mentre il secondo 2259 elementi (contro i 2303 dichiarati nella pubblicazione). Logicamente, dato che i due dataset sono stati creati da due diversi gruppi di

ricerca, i dati saranno disomogenei tra di loro, perciò utilizzeremo del codice python dapprima per sistemarli e successivamente per unirli.

In primis, è stata eseguita la conversione delle temperature di fusione del primo dataset da gradi Celsius a Kelvin. Durante la successiva fase di unione, 8 "DES composition" sono stati trovati uguali in entrambi i dataset: presentano cioè gli stessi componenti agli stessi rapporti molari. Tuttavia è stato deciso di tenere entrambe le coppie di dati nel file finale, dato che riportano delle leggere variazioni nelle temperature di fusione delle componenti (± 10 Kelvin al massimo di differenza). Questa accortezza viene fatta per preservare la maggior quantità possibile di informazioni; eventualmente questi dati verranno poi riprocessati prima di essere elaborati da un modello ML. Una di queste 8 coppie comuni però presenta temperature di fusione completamente diverse tra i due dataset (2-hydroxyethyl(trimethyl)azanium;chloride con tetradecan-1-ol, concentrazione 0.6/0.4; nel primo dataset la temperature di fusione (T-melt da ora in avanti) del DES è 352.18 K mentre la seconda 390.25 K), e per evitare futuri errori si è deciso perciò di scartare entrambe le copie dello stesso dato.

In seguito ad una rapida riorganizzazione dell'ordine tabelle e dei nomi dei campi, si è proseguito con l'unione vera e propria. Abbiamo ottenuto così un dataset finale con 2658 dati, con incluse anche le 7 coppie di DES "uguali". Nel file finale così ottenuto, troviamo che ogni elemento (DES) è stato necessariamente descritto con la formula SMILES di entrambe le componenti (Smiles-HBA e Smiles-HBD), la temperatura di fusione di entrambe le componenti (Tmelt-HBA e Tmelt-HBD), i rapporti molari (X-HBA e X-HBD) e la temperatura di fusione del DES risultante (Mixed-Tmelt) [Tab. 2]. Lo SMILES è una sigla che sta per Simplified Molecular Input Line Entry System, ed è un tipo di annotazione testuale lineare per rappresentare una molecola, e che ci permette anche una rappresentazione facilmente leggibile e computabile delle molecole. In aggiunta al nostro dataset, come campi "facoltativi", la maggior parte degli elementi possiederà anche il nome per esteso dei due componenti (Name-HBA e Name-HBD), il tipo di DES rappresentato da un numero intero da 1 a 5 (DES-T), e il reference/DOI della ricerca da cui sono stati recuperati questi dati (REF-DOI).

SMILES-HBA	Tmelt-HBA	SMILES-HBD	Tmelt-HBD	X-HBA	X-HBD	MIXED-Tmelt
Cl[Fe](Cl)Cl	580.15	C[N+](C)(C)CCO.[Cl-]	575.15	0.67	0.33	294.15
Br[Zn]Br	567.15	C[N+](C)(C)CCO.[Cl-]	575.15	0.67	0.33	311.15
Cl[Zn]Cl	563.15	C[N+](C)(C)CCCl.Cl	513.15	0.67	0.33	296.15
Cl[Zn]Cl	563.15	C[N+](C)(C)CCO.[Cl-]	575.15	0.50	0.50	314.77

Tabella 2: *Esempio elementi del Dataset*

Un lavoro simile è stato fatto anche per i dataset contenenti i dati di viscosità e densità. In questo caso però non abbiamo compiuto nessun lavoro di espansione del dataset, ma ci siamo limitati a prendere i dati del medesimo articolo [7] e costruire un dataset unico dove vengono descritti, per un determinato DES e la temperatura di stato, la densità e la viscosità del composto. Questo dataframe è stato creato solo per ottenere un unico file più facile da manipolare per esperimenti futuri, ma che in questo lavoro non verrà utilizzato, dato che ci limiteremo semplicemente ad analizzare e predire le temperature di fusione. Dopo la creazione dei nuovi dataset possiamo quindi procedere con il processamento dei dati e l'addestramento dei modelli.

Data Processing

Innanzitutto abbiamo scaricato presso il link a disposizione tutti i modelli ed i programmi dell'articolo, e creato un nuovo ambiente sul computer personale dove verranno scaricate poi tutte le librerie Python necessarie per far eseguire i software. In regime di partenza, si è notato subito che alcune formule SMILES non erano corrette, dato che si generavano costantemente errori nell'esecuzione dei programmi; si è quindi creato un nuovo modulo, chiamato valid-smiles.ipynb, che analizza rapidamente la validità di tutti gli SMILES presenti nel dataframe. Tre entry di dati, contenenti tutti la stessa molecola C1=CC=C2C(=C1)C=CC(=O) (2-Iodonaphthalene-1-carboxylic acid) presentavano uno SMILES invalido e dunque sono stati completamente rimossi. In questo breve programma, inoltre, per le 7 coppie già discusse in precedenza di DES definiti "uguali", gli elementi comuni sono stati aggruppati compiendo una media matematica dei campi Tmelt-HBA/HBD e Mixed-Tmelt.

Successivamente abbiamo utilizzato il programma data-analysis.ipynb per estrarre e rappresentare le informazioni statistiche riguardanti il dataset da po-

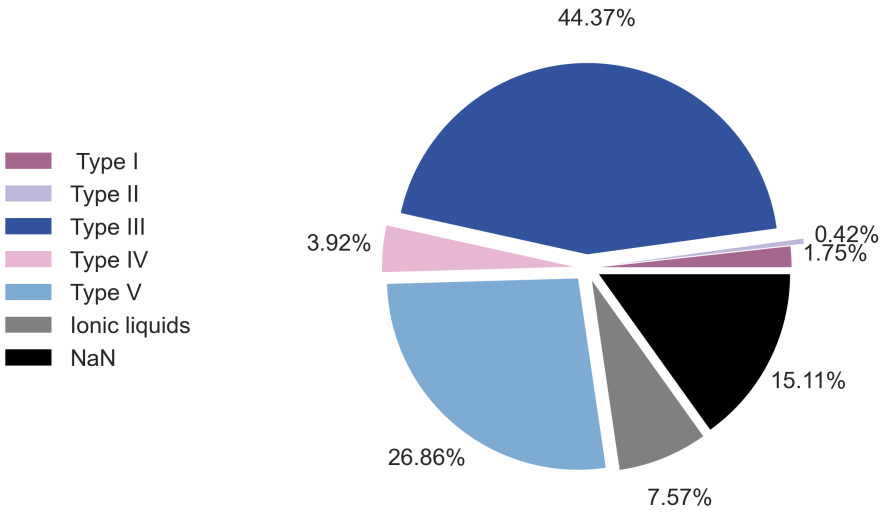

Figura 2: Ratio of different types of DES

co creato. In questo modo, analizzando le mode delle componenti dei DES, abbiamo trovato che il composto HBA più presente è il 2-hydroxiethyl(trimethyl) azanium;chloride, che appare in ben 104 coppie diverse nel nostro dataset (i successivi a scendere sono tutti sotto le 20 apparizioni), mentre per il componente HBD abbiamo il Decanoid Acid, Dodecanoid Acid ed Urea che compaiono rispettivamente 32, 31 e 27 volte (e anche qua abbiamo poi a scendere diversi componenti che appaiono intorno alle 15 volte). Per quanto riguarda il sistema più popolare, abbiamo il 5-methyl-2-propan-2-ylphenol con (1R,2S,5R)-5-methyl-2-propan-2-ylcyclohexan-1-ol che viene studiato con 34 diversi rapporti molari, seguito dal lithium;nitrate con N-methylacetamide a 32. Passiamo ora ad analizzare la composizione percentuale dei tipi di DES [Fig. 2], dove possiamo notare come il dataset sia abbastanza sbilanciato, con quasi la metà dei dati appartenenti al Tipo III e solo lo 0,47% appartenente al tipo II. In verità la distribuzione rappresenta proprio il loro spettro di applicazione a livello industriale, e comunque sono partecipi almeno 10 dati per ogni tipologia, ottenendo quindi un dataset universale che rispecchia proprio gli obiettivi iniziali del lavoro. Da notare inoltre una grossa fetta di dati senza la classificazione, che tuttavia potrà essere facilmente definita in futuro con una rapida operazione di classificazione.

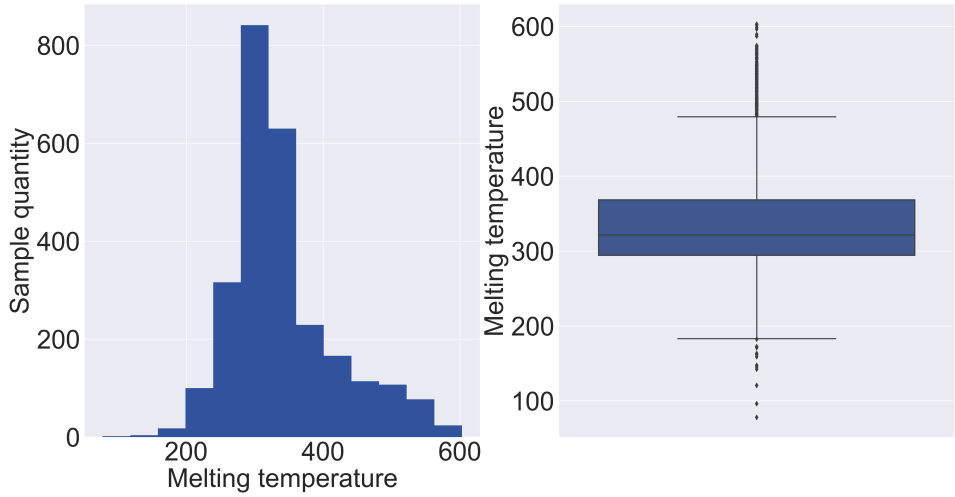

Figura 3: Melting Temperature Distribution

Il successivo passo è l'osservazione della distribuzione delle temperature dei DES. Possiamo in primo luogo notare [Fig. 3] come le temperature di fusione siano distribuite a formare una struttura simile a una curva gaussiana, dove il picco massimo viene raggiunto intorno ai 300 K con oltre 800 elementi. Osservando inoltre anche il boxplot, possiamo osservare che le minime scendono fino a 100 K, quindi sono liquide anche sotto 0 Celsius, mentre le massime superano i 600 K. Osserviamo ora il seguente grafico [Fig. 4] e notiamo le distribuzioni delle diverse temperature di fusione divise per Tipologie di DES: si nota logicamente che il Type II è quello con meno variazione perché possiede una quantità minore di dati, viceversa il Type III ne ha molti di più e la variabilità è inevitabilmente più elevata. Si osserva inoltre che, in generale, le temperature di Type IV sono le più basse, mentre gli Ionic liquids sono le più alte.

Finita l'analisi del nostro dataset, bisogna ora ultimare l'operazione di data processing. Dobbiamo finire di preparare i dati da dare "in pasto" ai modelli di machine learning. Per farlo utilizzeremo un programma detto mining-desc.ipynb, sviluppato da un gruppo di ricerca di nome acid-design-lab e utilizzato anche nel nostro articolo di riferimento. All'interno di questo programma si utilizza la libreria RDKit per estrarre dalle nostre formule SMILES i descrittori. I descrittori altro non sono che valori numerici (vettori) calcolati a partire da una formula chimica, e ne rappresentano le proprietà chimiche, fisiche e to-

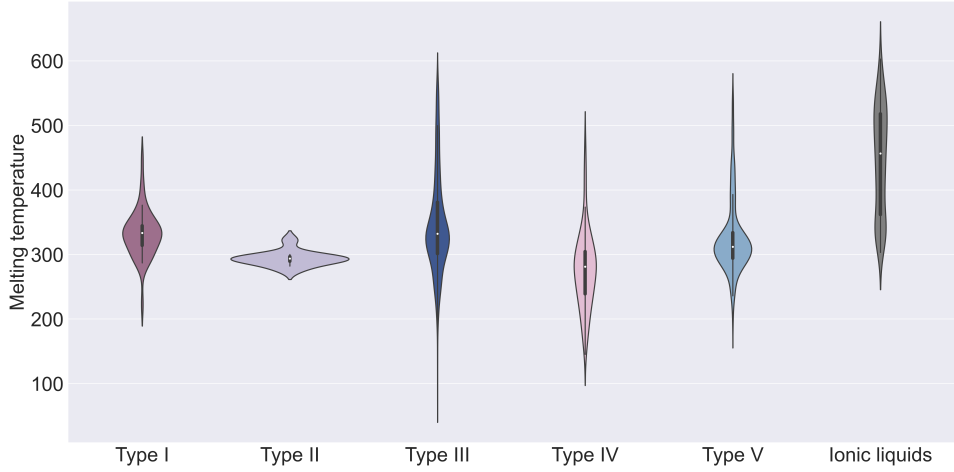

Figura 4: ViolinPlot for DES type

pologiche, che verranno poi utilizzate per addestrare modelli matematici o di machine learning. Il pacchetto RDKit ci fornisce oltre 200 descrittori per ciascun componente e per ogni composizione DES, tuttavia tanti di questi dati estratti sono ridondanti e multicollineari, oltre a causare problemi di consumo di memoria e complessità temporale. Utilizzeremo perciò solo i 21 descrittori che sono stati selezionati nell'articolo di riferimento, tra cui: descrittori geometrici come indice di sfericità, volume di Van Der Waals e raggio di girazione, descrittori strutturali come numero di atomi, numero di anelli aromatici e numero di gruppi, descrittori sperimentali come temperatura e frazioni molari (che sono già presenti nel dataset di partenza). Il dataset è ora pronto per iniziare l'addestramento dei machine learning models.

Nella successiva ed ultima fase, useremo il programma Melting-temperature-ML.ipynb dove sono contenuti tutti i modelli di Machine Learning. Gli algoritmi in questione utilizzati sono 8: Decision Tree Regression (DTR), un algoritmo di regressione supervisionata che utilizza un albero decisionale e divide lo spazio in regioni rettangolari, all'interno delle quali assegna un valore numerico medio come predizione; Random Forest Regression (RFR) che combina più alberi decisionali rispetto al precedente, e la previsione finale è la media delle previsioni dei singoli alberi; Gradient Boosting Regression (GBR), che utilizza molti alberi deboli in sequenza dove ogni nuovo albero impara a correggere gli errori commessi dai precedenti; Cat Boosting Regression (CBR) e XG Boosting

Regression (XGB), simili al precedente ma più avanzati; K-nearest neighbors Regression (KNR), che agisce direttamente sui dati e si basa sulla prossimità tra di loro; Support Vector Regression (SVR), che cerca una funzione che si discosta il meno possibile dai dati reali, entro una certa tolleranza; Multi Layer Perceptron (MLP), la famosa rete neurale artificiale feedforward composta da più strati di neuroni artificiali, di utilizzo molto popolare. A questi 8 modelli si aggiunge poi anche la Linear Regression (LR), ovvero la semplice regressione lineare che trova la retta migliore tra x (descrittori) ed y (Mixed-Tmelt).

In questo programma prima di tutto il dataset è stato suddiviso in 5 parti uguali. Queste cinque parti verranno poi utilizzate per addestrare i modelli cinque volte, utilizzando ogni volta 4 parti per il "training set" e 1 parte per il "test set". Le cinque parti sono state divise con il cosiddetto metodo "mixture out", assicurando quindi che tutti i DES con stessa miscela (ma rapporti molari diversi) fossero presenti solo in un determinato gruppo. Il set di dati di training è stato utilizzato per adattare i modelli di machine learning, mentre il set di test è stato utilizzato solo per valutarne le prestazioni e la capacità di generalizzazione.

Nella seconda parte del programma invece avviene il vero e proprio addestramento, e anche lo studio della performance di ciascun modello. Parlando proprio di performance, le prestazioni dei modelli sono state valutate utilizzando due metriche standard: il coefficiente di determinazione R2 e l'errore quadratico medio (RMSE). Il primo serve ad indicare la quantità totale dei dati variabili che viene spiegata dal modello: in generale ha valori compresi tra 0 e 1, dove 1 è il migliore dei modelli raggiungibili mentre 0 risulta un modello inutile (che non spiega nulla), ma può essere anche negativo se il modello è peggiore della media. RMSE invece misura quanto sbaglia in media il modello in generale: in questo caso, più basso è l'RMSE e migliore è la qualità della previsione, mentre con valori elevati vuol dire che le previsioni del modello sono molto discostanti dai valori reali.

3 Risultati

Analizziamo ora i risultati così ottenuti. Partiamo da una premessa: il tempo di esecuzione dei programmi. Purtroppo, la maggior parte dei modelli in questione sono troppo complessi per essere addestrati su un computer personale (ovvero il materiale a nostra disposizione), e spesso l'esecuzione di essi è andata in "overtime". Un tentativo è stato fatto provando a ridurre la mole di lavoro totale, dimezzando ad ogni tentativo la quantità di dati utilizzati per addestrare i modelli. I modelli tuttavia risultavano ancora troppo complessi per essere eseguiti, e solamente partendo con un dataset ristretto di 50 dati (pari a circa il 2% del totale) i modelli venivano svolti senza errori. Si tratta dunque di una base di partenza troppo ristretta per poter trarre conclusioni significative. Un altro tentativo è stato fatto utilizzando dei tools online, come Google Colab, ma anche questo approccio ha riscontrato i medesimi problemi. Gli unici output quindi che siamo riusciti ad ottenere con dataset completo sono quelli dei due modelli più semplici, in questo caso la Linear Regression e Decision Tree Regression. Pertanto, le nostre osservazioni si baseranno solamente su questi risultati ottenuti e nulla di più, tuttavia non escluderemo il fatto di provare a dedurre conclusioni più universali.

Il primo appunto è stato fatto eseguendo nuovamente l'intero programma con il set di dati originale. I risultati ottenuti per R2 e RMSE sono perfettamente uguali a quelli riportati nell'articolo. In secondo luogo, è avvenuto l'addestramento dei due modelli più elementari sempre con il dataset originale, ma omettendo le temperature di fusione delle due componenti HBA e HBD; e infine anche il nuovo dataset da poco ideato, sia con che senza i melting point. Per ciascuno dei due modelli avremo quindi 4 differenti output, che utilizzeremo per comparare sia R2, RMSE di cross validation, ma anche R2 e RMSE di training, per un analisi più accurata. Possiamo definire, come si può anche osservare in fig. 6 e 7 (distribuzione dei risultati per il primo test di cinque, ottenuto in questi casi con il dataframe nuovo, con le temperature di fusione), che questi due modelli sono molto poco precisi, e che anzi il modello DT sembra possa assumere solo alcuni determinati valori, probabilmente

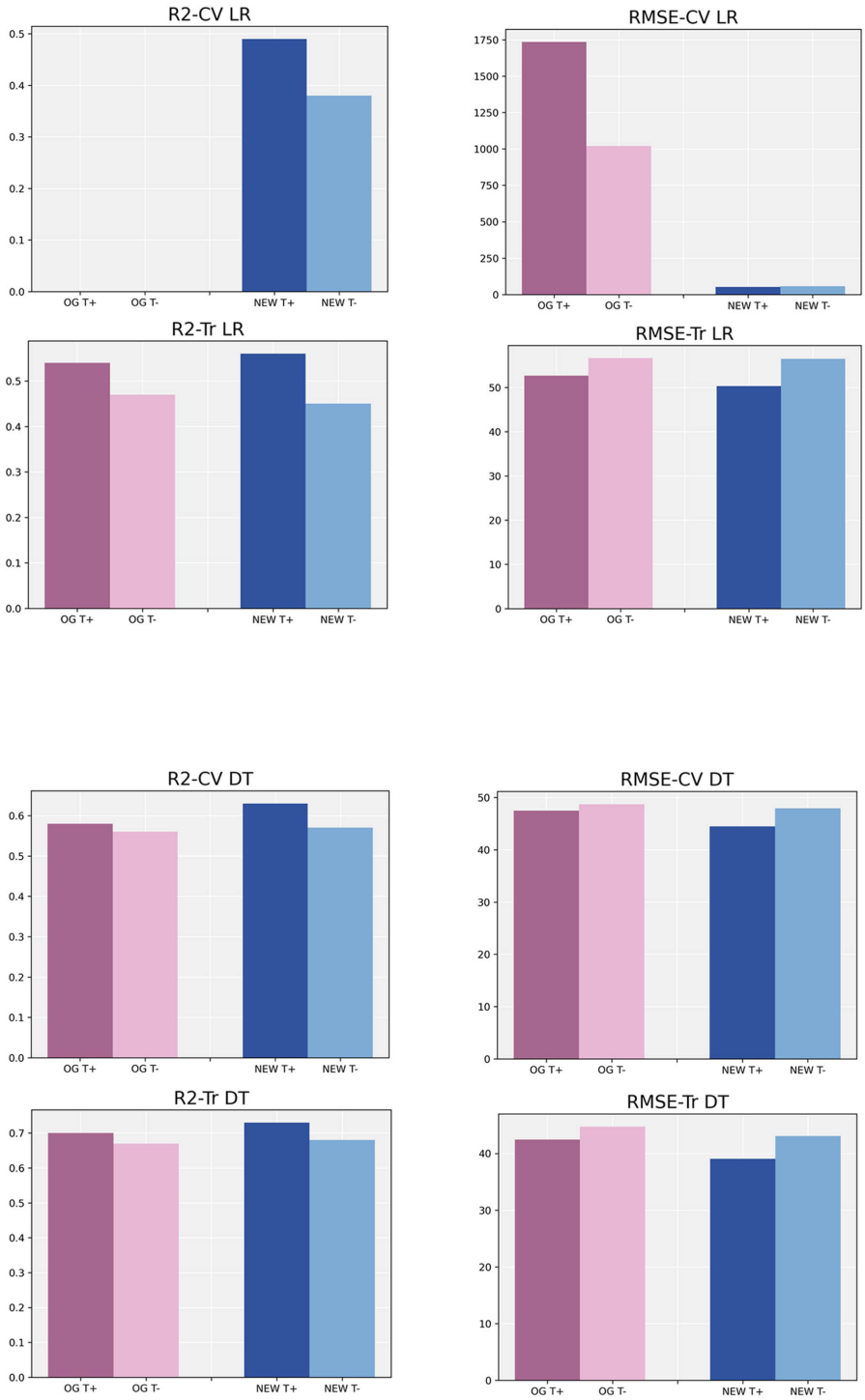

Figura 5: Risultati dei Modelli

causato proprio dalla sua bassa complessità di esecuzione.

Partiamo analizzando la Linear Regression. I risultati originali dell'articolo prevedono un R2-CV molto anomalo, pari a -1257.42, che stranamente "migliora" a -463.80 senza temperature (questi due risultati altamente negativi

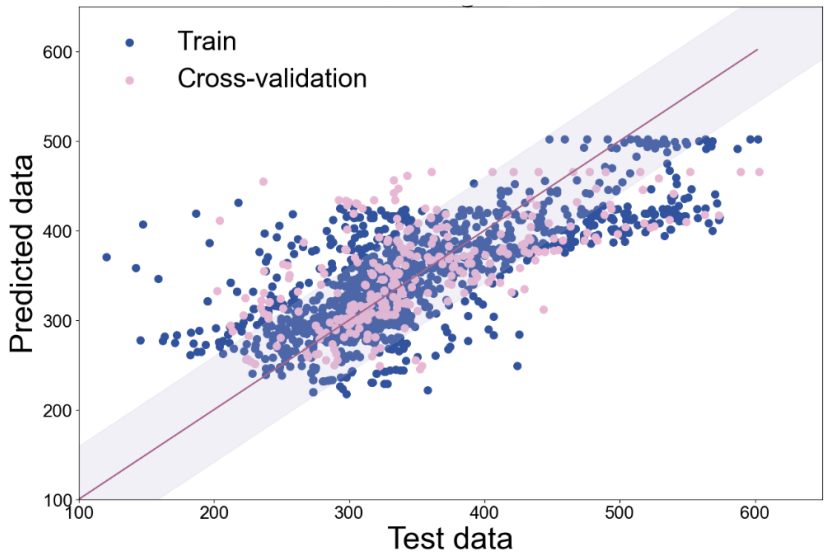

Figura 6: Linear Regression

Figura 7: Decision Tree

sono dall'addestramento del modello utilizzando la parte 5 di 5 del dataframe come y-test, dove R2 arriva fino a -4091.47, abbassando di conseguenza sotto lo zero la media generale). In entrambi i casi comunque possiamo evincere che il modello non spiega assolutamente nulla quindi completamente inaffidabile. Complementarmente, una seconda conferma di ciò e data dalla rappresentazione di RMSE-CV, inaspettatamente elevato, pari a 1735.46, che migliora a 1020.30 senza Tmelt. In questo caso, l'incremento della perfezione degli indici senza temperatura non deve essere un fattore preso in seria considerazione,

data la bassa precisione che riscontra il modello in entrambe le casistiche. Le osservazioni variano in positivo invece durante le analisi sul dataset di training, dove R2-Tr è di 0.54, che scende a 0.47 nel dataframe senza temperatura. Rmse-Tr invece è di rispettivamente 52.67 e 56.76. Queste tipologie di dati rimangono ancora troppo imprecise rispetto agli standard accettabili, ma si trattano comunque di elementi quantomeno "decifrabili" rispetto ai test sulla cross-validation. Ragionevolmente, in questo caso la precisione dell'algoritmo senza le Tmelt diminuisce, ma non di molto. Confrontiamo ora questi risultati con quelli ottenuti a partire dal nostro dataset. R2-CV rispettivamente con e senza temperatura è di 0.48 e 0.38, mentre RMSE-CV è di 52.74 e 57.86. Non è possibile attuare un confronto diretto con i test ottenuti a partire dal dataset originale, dato che quest'ultimo prevedeva dati del tutto atipici, tuttavia è facilmente riscontrabile il netto miglioramento avvenuto. Si può anzi aggiungere che i valori della cross-validation del dataset nuovo sono equiparabili ai valori di training del dataset vecchio. Proseguendo, R2-Tr è di 0.56 e 0.45, mentre RMSE-Tr è di 50.32 e 56.49. La differenza dei test di training tra i due dataset è molto più sottile delle precedenti. In questo caso inoltre l'eliminazione della temperatura pesa in maniera più negativa sul dataframe da noi creato (Delta negativo di 4.06 nel vecchio contro un Delta negativo di 6.17 nel nuovo).

Concentriamoci ora sull'analisi del modello Decision Tree, di migliore validità data la maggiore precisione relativa dell'algoritmo. In questo caso il valore di R2-CV con il dataset originale è di 0.58, che diventa 0.56 omettendo le temperature di fusione. Già da queste prime informazioni si può facilmente comprendere come il modello è decisamente migliore del precedente. Il valore di R2-CV del nuovo dataset è invece di 0.63, che diventa 0.57 senza temperature. L'incremento complessivo di R2-CV dal dataset di base a quello nuovo è di 0.05, o in percentuale dell'8.6%. Attuando un confronto tra i medesimi dataframe con e senza temperatura, si evince che il nuovo dataset ha una variazione maggiore dell'originale (delta di 0.06 contro 0.02), e dunque le Tmelt hanno un peso finale maggiore sull'accuratezza del modello. Proseguiamo poi con RMSE-CV, dove per il modello originale vale rispettivamente 47.46 (con Tmelt) e 48.72 (senza Tmelt), mentre per il dataset nuovo è di

44.49 e 47.91. Anche in questo caso il trend rispecchia le precedenti considerazioni, ove vi è una riduzione di RMSE pari al 7,3%; ed è sempre presente una variazione maggiore di Tmelt+/Tmelt- per il dataset nuovo (delta 3.42) contro il dataset originale (delta di 1.26). Passiamo in secondo luogo all'analisi dei dati di training. R2-Tr per il dataset originale vale 0.70, che diventa 0.67 senza temperatura, mentre per il dataset nuovo è di 0.74, che diventa 0.68 senza temperatura. Per queste tipologie di dati l'incremento di precisione tra dataset nuovo e vecchio è minore, pari "solamente" ad un 5.7%l. ciò è probabilmente dovuto dal fatto che i valori assoluti di R2-Tr sono comunque più alti rispetto a R2-CV, quindi è difficile mantenere un miglioramento percentuale di simil dimensione. Comunque, anche in questo caso il delta dovuto dall'assenza di temperatura è maggiore nel nuovo dataframe (0,05) che non nell'originale (0,03), con una differenza però meno accentuata delle precedenti. Analizziamo infine RMSE-Tr, che vale per il primo dataset rispettivamente 42.43 e 44.73, mentre per il secondo 39.04 e 43.09. La riduzione riscontrata è pari all'8% e la variazione data dall'assenza di temperatura è sempre maggiore nel dataset nuovo.

Osservando ora i dati nella loro interezza, possiamo quindi definire che l'addestramento dei modelli con il nuovo dataset ottiene dei risultati più precisi di un 5-8%, variabili a seconda delle condizioni. Con un certo grado di astrattezza, queste osservazioni potrebbero anche essere estese ai vari modelli che, per motivi di calcolo, non si è potuto completare l'addestramento. Ad esempio, per CatBoostRegression, nell'articolo definito come il modello più preciso, che infatti presenta già con il dataset originale un R2-CV molto elevato (0.78), è probabile aspettarci un miglioramento dell'1-2%. Inoltre, un'altra particolare osservazione è la somiglianza di precisione ottenuta con il nuovo dataset senza temperatura e quello originale (dotato di temperature di fusione) per entrambi i modelli, segno evidente di un miglioramento generale complessivo del sistema. Il nuovo dataframe esteso sembra quindi produrre risultati promettenti, e successive espansioni e sistemazioni di esso potrebbero quindi incrementare ulteriormente la precisione delle previsioni. Questo lavoro infine può anche non limitarsi solo alla temperatura di fusione, ma essere esteso anche a viscosità e

densità dei DES, per ottenere così un'analisi completa e di utile applicabilità in campo industriale.

Bibliography

[1] D.R. Joshi e N.Adhikari. *An Overview on Common Organic Solvents and Their Toxicity.* J. Pharm. Res. Int., 2019, pp. 1–18. DOI: 10.9734/jpri/2019/v28i330203.

[2] D. L. Davies A. P. Abbott G. Capper, R. K. Rasheed e V. Tambyrajah. «Novel solvent properties of choline chloride/urea mixtures». In: *Chem. Commun.* (2003). DOI: 10.1039/B210714G.

[3] M. Hayyan I. Juneidi e M. A. Hashim. «Evaluation of toxicity and biodegradability for cholinium-based deep eutectic solvents». In: *RSC Adv.* (2015). DOI: 10.1039/C5RA12425E.

[4] M. de la Fuente Revenga D. Lindberg e M. Widersten. «Deep eutectic solvents (DESs) are viable cosolvents for enzyme-catalyzed epoxide hydrolysis». In: *J. Biotechnol.* (2010). DOI: 10.1016/j.jbiotec.2010.04.011.

[5] K. A. Omar e R. Sadeghi. «Database of deep eutectic solvents and their physical properties: A review». In: *J. of Molecular Liquids* (2023). DOI: 10.1016/j.molliq.2023.121899.

[6] M. Wysokowski R. K. Luu e M. J. buehler. «Generative discovery of de novo chemical designs using diffusion modeling and transformer deep neural networks with application to deep eutectic solvents». In: *Applied Physics Letters* (2023). DOI: 10.1063/5.0155890.

[7] T. Rakhmanov V. Odegova A. Lavrinenko et al. «DESignSolvents: an open platform for the search and prediction of the physicochemical properties of deep eutectic solvents». In: *Green Chem.* (2024). DOI: 10.1039/D3GC04533A.