ensia The National School of Artificial Intelligence
المدرسة الوطنية العليا للذكاء الاصطناعي

# Predicting Chemical Solubility with Data Mining

**Group members:**
Ouari Meriem
Kara Laouar manel
Zerrouki Fella Manel
Belhadj Sara
Litim Chiraz

**Supervisors:**
Mohamed Brahimi
Sami Belkacem
Seif Eddine Bouziane

January 16, 2023

## 1.Introduction

Organic solvents play a central role in chemistry, influencing synthesis, catalysis and composition. However, predicting the solubility of various organic solvents is a significant challenge. Traditional computational methods such as QSPR and mechanistic models are limited, especially when predictions are extended beyond water. The current reliance on labor-intensive experiments hinders rapid solvent screening in the chemical industry. Despite advances, traditional models face practical limitations that require extensive experimental data.

Our project uses data mining and machine learning techniques where the main goal is to revolutionize solvent screening by providing a fast alternative to traditional methods.

## 2.Collecting data and the source + data description

The dataset is constructed using an open-source dataset called "BigSolDB: Solubility Dataset of Compounds in Organic Solvents and Water in a Wide Range of Temperatures." This dataset is quite extensive, containing 54,273 experimental solubility values. We measured these solubility values across a wide temperature range, spanning from 243.15 to 403.15 K, and in various organic solvents as well as water.

To enhance the dataset, we utilized RDKit, a potent tool in cheminformatics and machine learning. RDKit helps us compute and manipulate chemical information effectively. During the final stages of constructing the dataset, RDKit played a crucial role in filling in additional details. We added attributes such as molecular weight, number of hydrogen donors, number of rings, number of rotatable bonds, and polar surface area. The incorporation of these attributes was made possible through machine learning techniques, ensuring a more comprehensive and informative dataset for our solubility prediction project.

The following table provides a concise overview of the key attributes incorporated into the dataset and their respective descriptions.

| Attribute | Description |
|---|---|
| SMILES | Simplified Molecular Input Line Entry System (SMILES) for the chemical compound. |
| T,K | Temperature at which solubility was measured (in Kelvin). |
| Solubility | Experimental solubility values measured across a temperature range and solvents. |
| Solvent | The specific organic solvent used for measuring solubility. |
| SMILES_Solvent | Simplified Molecular Input Line Entry System (SMILES) for the solvent used. |
| Molecular Weight | The molecular weight of the compound, a measure of its mass. |
| Number of H-Bond Donors | The number of hydrogen atoms that are donors in the molecule. |
| Number of Rings | The number of rings present in the chemical structure. |
| Number of Rotatable Bonds | The number of bonds that allow free rotation around themselves. |
| Polar Surface Area | The surface area of the molecule occupied by polar atoms and groups. |

## 3.Cleaning the data

During the initial review of the dataset, we have observed that some columns had missing values. In particular, columns such as 'T,K', 'Solubility', 'Solvent', 'SMILES_Solvent', 'Molecular Weight', 'Number of H-Bond Donors', 'Number of Rings', 'Number of Rotatable Bonds', and 'Polar Surface Area' had 1556 missing entries. These missing values accounted for approximately 2.8% of the entire dataset. Therefore, we decided to drop these rows and get a cleaned dataset containing 52,717 rows. Then it turns out that there are rows with the same values in all columns. These duplicates, identified through the 'duplicated rows' function, amounted to 670 rows. In order to ensure the integrity of the dataset we removed duplicate rows, reducing the number of rows from 54,273 to 52,047.

The uniqueness of the dataset was further emphasized by examining the number of unique SMILES and solvents. The cleaned dataset consists of 830 unique SMILES and 128 unique solvents, highlighting the diversity and completeness of the dataset.

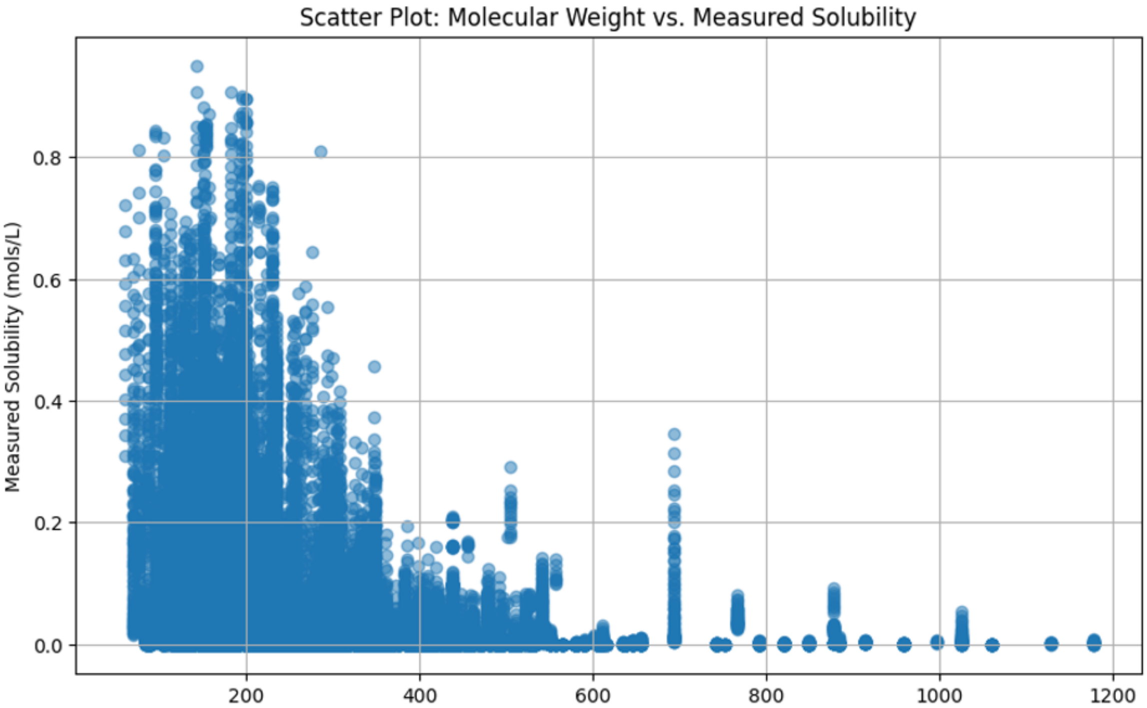
There were also some issues with invalid code, specifically with SMILES notation for chemical solvents. To address this, manual searches were conducted to identify and add the correct SMILES notations. One example is the solvent represented by the SMILES code "C(COCCOCCOCCOCCO)O," which corresponds to PEG-400 (polyethylene glycol with a molecular weight of 400). This information was added to ensure accurate representation in the chemical dataset.

# 4. Exploratory Data Analysis (EDA):

In this section, we delve into the exploration of our raw data, investigating its distribution across various levels of granularity and scrutinizing the relationships between Solubility and additional factors, such as Temperature (Kelvin). The primary objective of this analysis is to refine the dataset further, aiming to streamline the number of pertinent features initially outlined in the preceding section.

### 4.1 Molecular Weight vs. Solubility

We initiate our exploration with a scatter plot illustrating the relationship between molecular weight and measured solubility **(Fig. 4.1)**. A notable negative correlation is observed, where smaller molecular weights tend to exhibit higher solubility. As molecular weight increases, solubility tends to decrease, aligning with the established notion that smaller molecules are generally more soluble.



*Analysis:*

The clustering of points in the lower range of molecular weights signifies enhanced solubility, while increased dispersion at higher molecular weights indicates reduced solubility. This inverse relationship underscores the substantial impact of molecular weight on solubility characteristics.

### 4.2 Chemical Space Visualization Tool

Continuing our exploration, we employ a web-based chemical space visualization tool to showcase solubility trends across temperatures for a selected compound (20 selected compounds) **(Fig. 4.2)**. The tool highlights the influence of different solvents (10 most used solvents in our dataset)

*Analysis:*

Acetone and methanol emerge as highly effective solvents, displaying elevated solubility values. Other solvents exhibit average solubility, while water stands out with the lowest solubility compared to alcoholic solvents. The tool underscores the significant impact of solvent selection on solubility, with acetone and methanol proving particularly favorable. Users can leverage these insights for informed decisions regarding solvent selection based on desired solubility characteristics for specific compounds.

*Example:*

The example below illustrates the impact of solvents on the solubility of a selected compound, providing a tangible demonstration of the tool's capabilities.

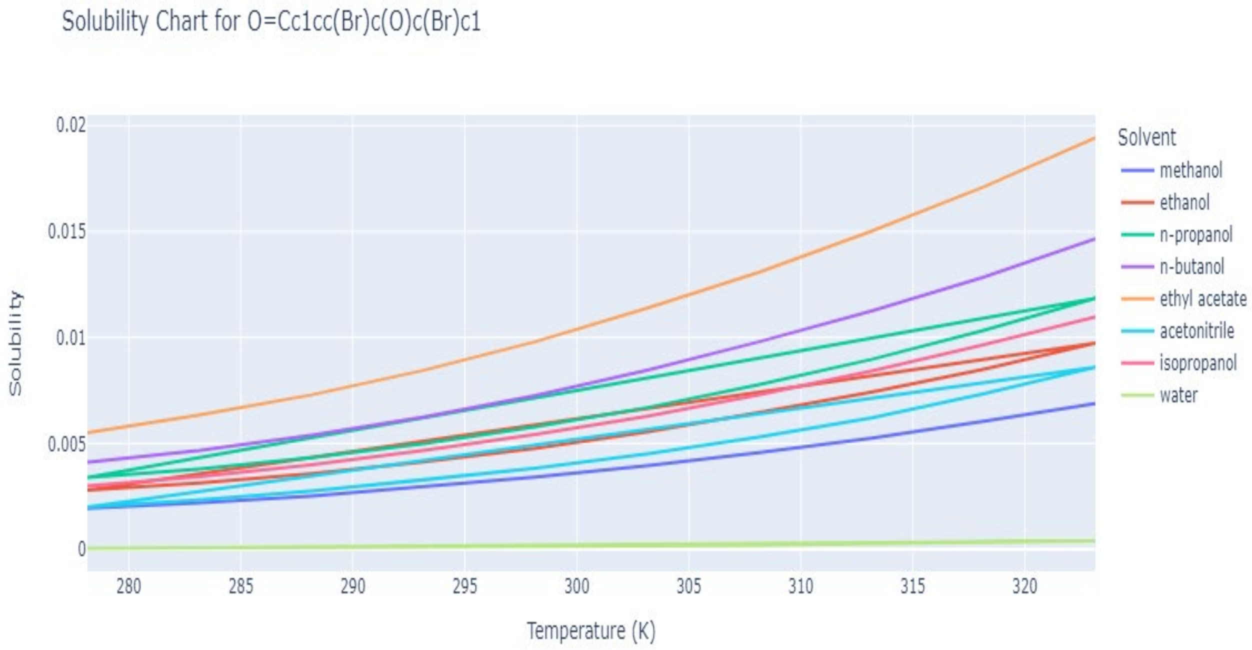


**Fig 4.2** Chemical Space Visualization Tool

### 4.3 Solvent Occurrence Barplot: Analyzing the Top 50 Solvents

This barplot visualization explores the distribution of the dataset's solvents, focusing on the 50 most prevalent. Each bar in the count plot corresponds to a solvent, with its frequency of occurrence within the dataset.

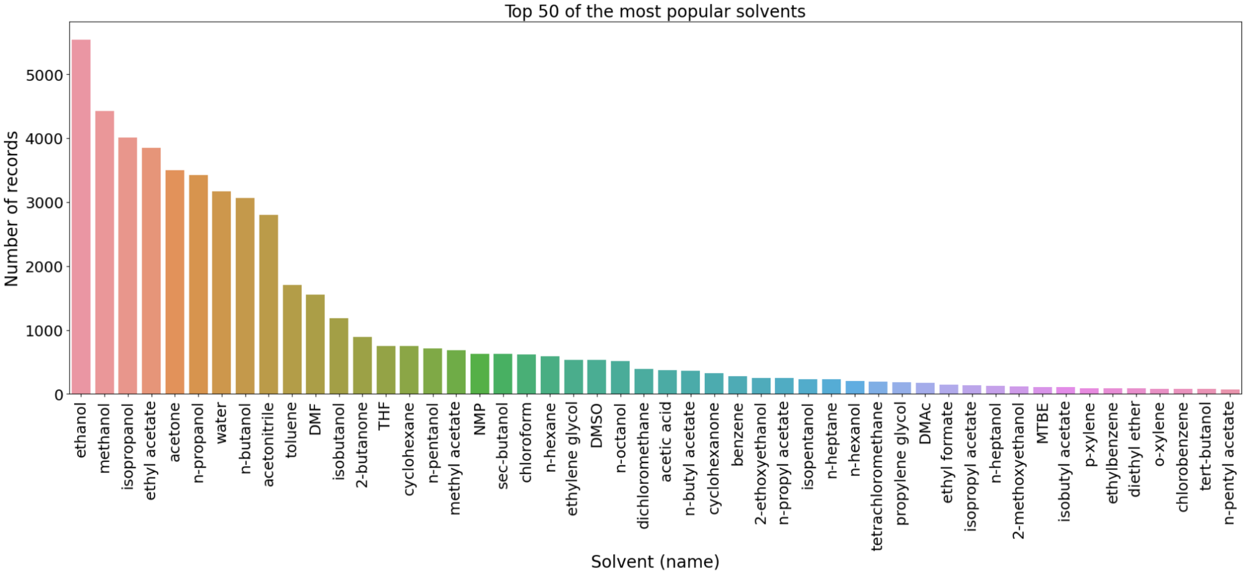
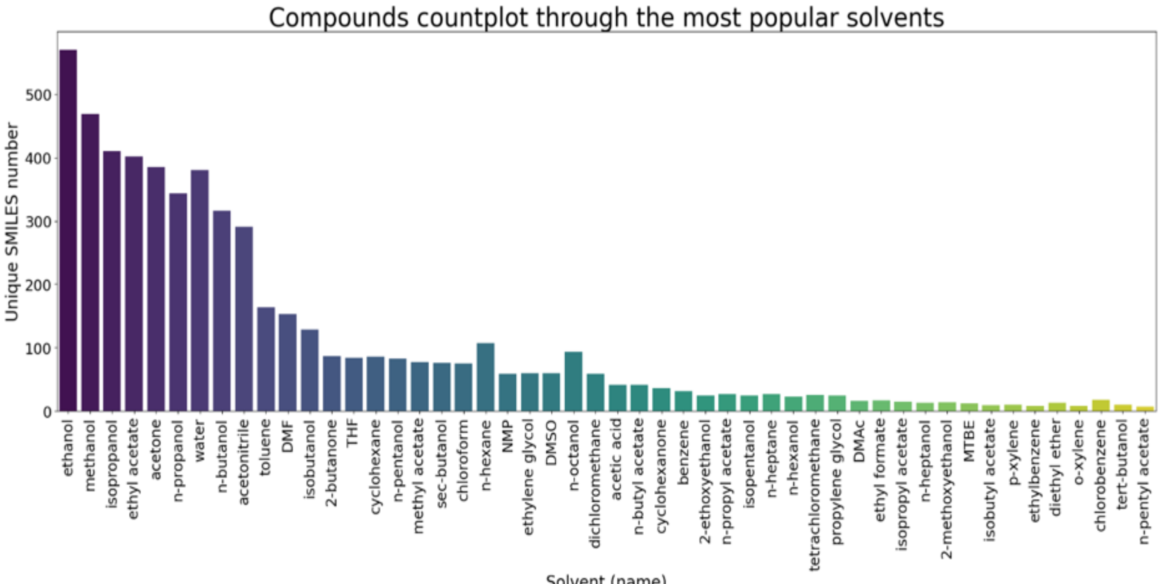


**Fig 4.3** Solvent Occurrence Barplot

Here in this section, we are visualizing the distribution of unique compounds across the top 50 solvents, this countplot offers a snapshot of chemical diversity within the dataset. Each bar represents a solvent, showcasing the number of distinct compounds studied. Explore the prevalence and variety of compounds in different solvent environments.



*Analysis:*
The dataset exhibits a notable bias towards a concentrated set of the top 50 solvents, contributing to 99.99% of recorded instances. Despite the presence of 128 unique solvents, this concentrated group significantly overshadows others, emphasizing their substantial influence. Among the top solvents, ethanol, methanol, and isopropanol stand out with substantial record counts, indicating their prevalent usage or relevance in experiments. Moreover, the dataset shows a preference for experiments involving various alcohols over other solvents, such as water, aligning with common practices in organic chemistry. This dominance of specific solvents underscores their potential importance in experimental setups, raising considerations about the diversity and representation of other solvents beyond the top 50. The visual representation in the barplot (**Fig 4.3)** provides insights into the distribution of records across these solvents, highlighting their varying occurrences within the dataset.

### 4.4 Solubility Distribution Across Top 20 Solvents

The histogram (**Fig 4.4)** visually represents the solubility distribution for the top 20 solvents in our dataset. Each subplot corresponds to a solvent, providing a snapshot of solubility patterns among frequently encountered chemicals.

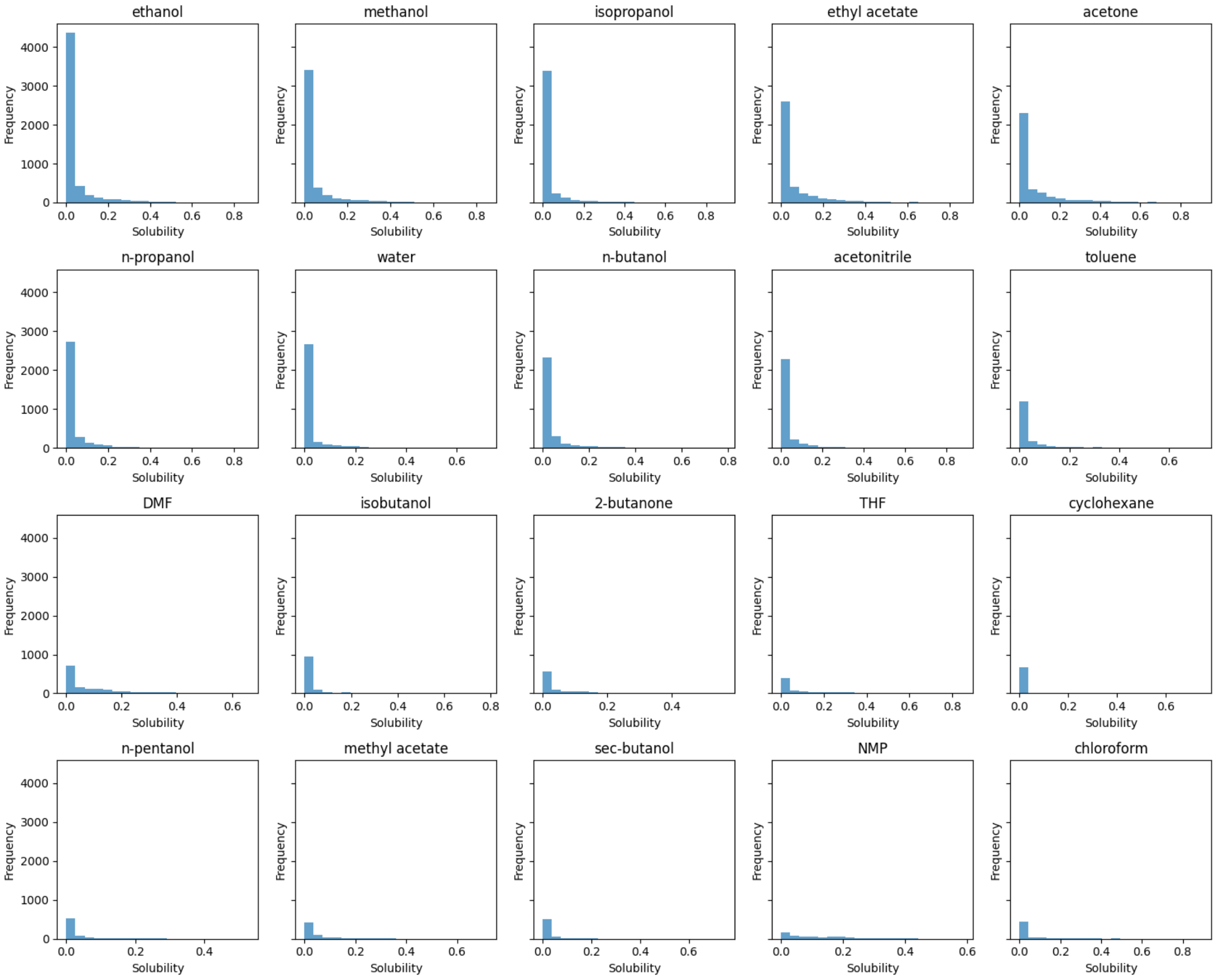


**Fig 4.4** Histograms:Solubility Distribution Across Top 20 Solvents

*Analysis:*

Ethanol, methanol, and isopropanol share similar distribution shapes, indicating potential similarities in solubility traits. Concentrated peaks around 0.0 and 0.1 on the solubility axis suggest a predominant range for these alcohols. Acetone and ethyl acetate exhibit broader distributions, implying diverse solubility behaviors compared to alcohols. Water, characterized by concentrated measurements at lower solubilities. Solvents like methyl acetate, sec-butanol, NMP, and chloroform present more uniform distributions, hinting at varied solvation tendencies across a wider solubility range. This concise analysis provides a quick understanding of solubility variations among the top 20 solvents.

## 4.5 Heatmap of the number of solubilities measured in the 30 most popular solvents at the most popular temperatures

The generated heatmap (**Fig 4.5**) provides valuable insights into the distribution of solubility measurements within the dataset, shedding light on the prominent solvents and temperatures involved.

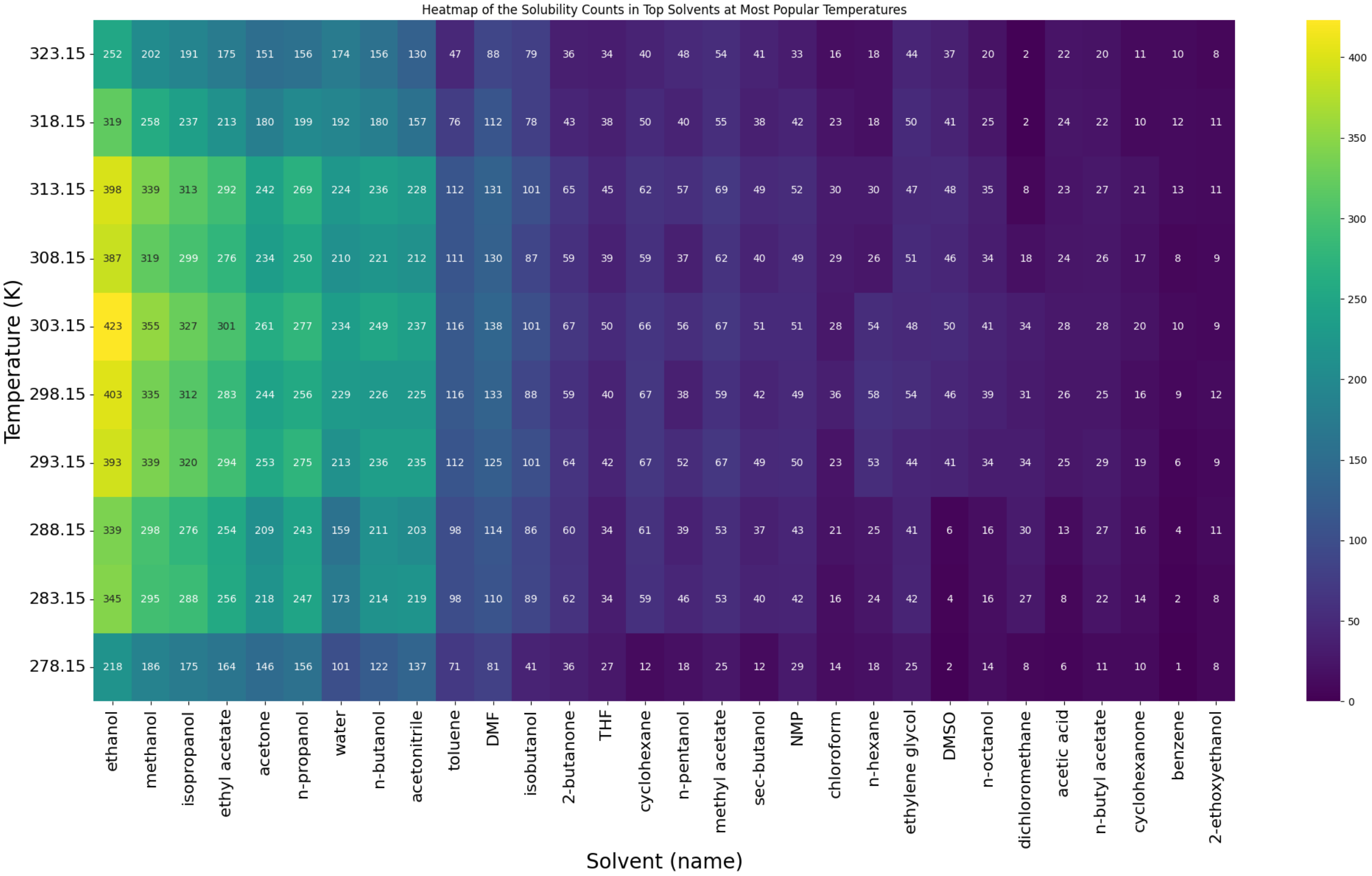


**Fig 4.5** Heatmap of the number of solubilities measured in the 30 most popular solvents at the most popular temperatures

*Analysis:*

The top-ten solvents (ethanol, methanol, isopropanol, ethyl acetate, acetone, n-propanol, water, n-butanol, acetonitrile,toluene) collectively represent 67.33% of the dataset.

Notably, water, despite being the seventh most frequent solvent, accounts for only 6.01% of the total records. The dataset reflects a bias towards experiments conducted in alcohols rather than water, aligning with common practices in organic chemistry.

totaling 423 measurements, is associated with the solvent ethanol at a temperature of 303.15 K. This particular combination stands out as the most frequently employed condition in solubility studies. The significance of this finding lies in the widespread use of ethanol at 303.15 K, indicating its potential as a standard condition or a particularly effective medium for investigating solubility in various compounds.

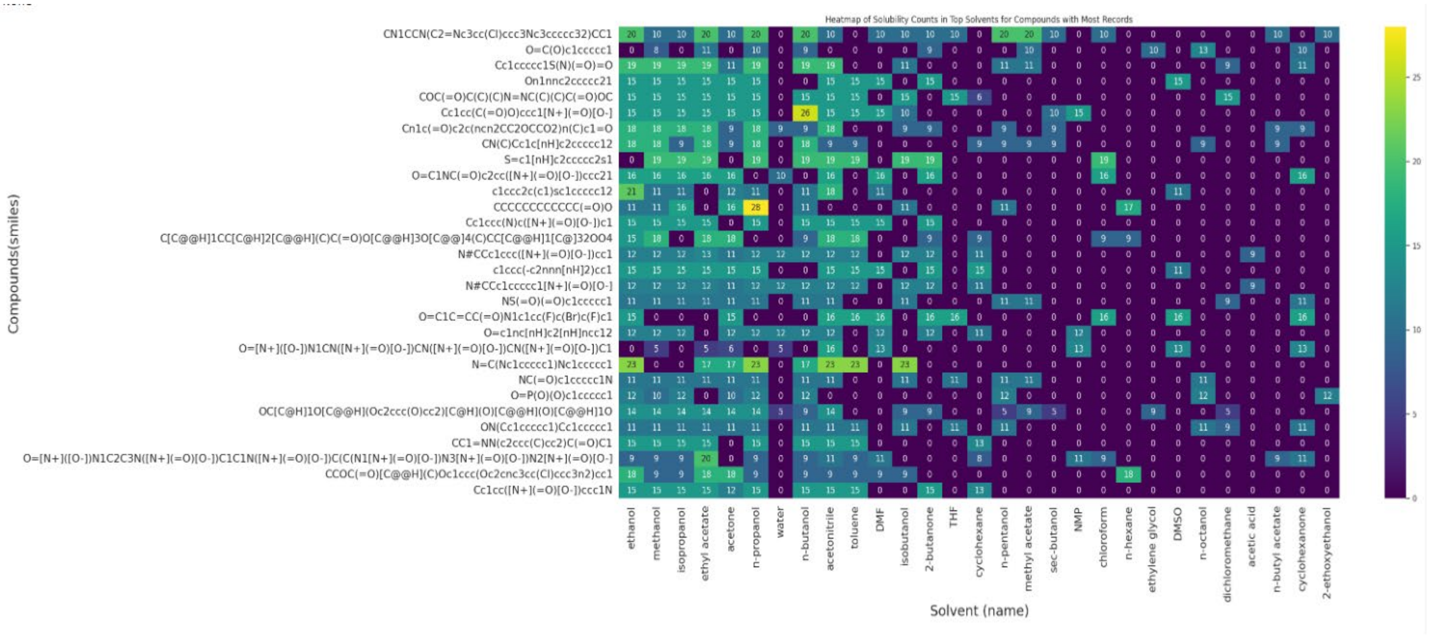
In contrast, the lowest solubility count, amounting to just 1 measurement, is linked to the solvent benzene at a temperature of 278.15 K. The limited occurrence of this combination suggests that benzene at 278.15 K is seldom utilized in solubility experiments. The significance of the low count prompts considerations regarding the experimental conditions, indicating that this particular combination might be less common or less suitable for the compounds under study.

*Prediction:*

The prevalence of **ethanol**, known as a universal solvent due to its molecular structure capable of dissolving both polar, hydrophilic, and nonpolar, hydrophobic compounds, is notable in solubility experiments. The consistent pairing of ethanol with **303.15 K** suggests a potential standardization, indicating a favored condition or convention in the experimental setup. Further exploration is recommended to uncover the specific advantages or reasons behind this frequently observed combination.

### 4.6 Heatmap of the number of solubilities measured in the 30 most popular solvents for the compounds with the largest number of records.

The presented heatmap serves as a powerful tool for unraveling intricate patterns in solubility measurements across the top 30 solvents, with a specific focus on compounds boasting the highest record counts.



*Analysis:*

In our analysis, we looked at how different solvents affect the solubility of compounds. The selected top-ten solvents, including ethanol, methanol, isopropanol, ethyl acetate, acetone, n-propanol, water, n-butanol, acetonitrile, and toluene, are considered essential for their widespread use. Across various temperatures, these solvents consistently exhibit high solubility for nearly all investigated compounds, as indicated by the extensive number of records.

We observe that 'Cc1cc(C(=O)O)ccc1N+[O-]' (sixth in records) consistently dissolved well in n-butanol in 26 trials at different temperatures. Similarly, 'CCCCCCCCCCCC(=O)O' (tenth in records) showed strong solubility in n-propanol across 28 trials.

The compound N=C(Nc1ccccc1)Nc1ccccc1 exhibited consistent high solubility in multiple solvents (ethanol, n-propanol, acetonitrile, toluene, and isobutanol) over 23 trials at various temperatures.

We can say from this analysis that solvent choice significantly influences compound solubility, with specific solvents like n-butanol and n-propanol consistently showing high solubility for certain compounds. Compound-specific behavior is evident, as exemplified by 'Cc1cc(C(=O)O)ccc1N+[O-]' and 'CCCCCCCCCCCC(=O)O.' Additionally, compounds like N=C(Nc1ccccc1)Nc1ccccc1 exhibit versatility, dissolving well in multiple solvents. However, water appears less effective for these 30 compounds. These trends hold consistently across different temperatures, providing valuable insights for future experiments and solvent selection.

## 4.6 Correlation matrix of solubility and the other attributes

The correlation matrix analysis sheds light on the relationships between solubility and the selected attributes.

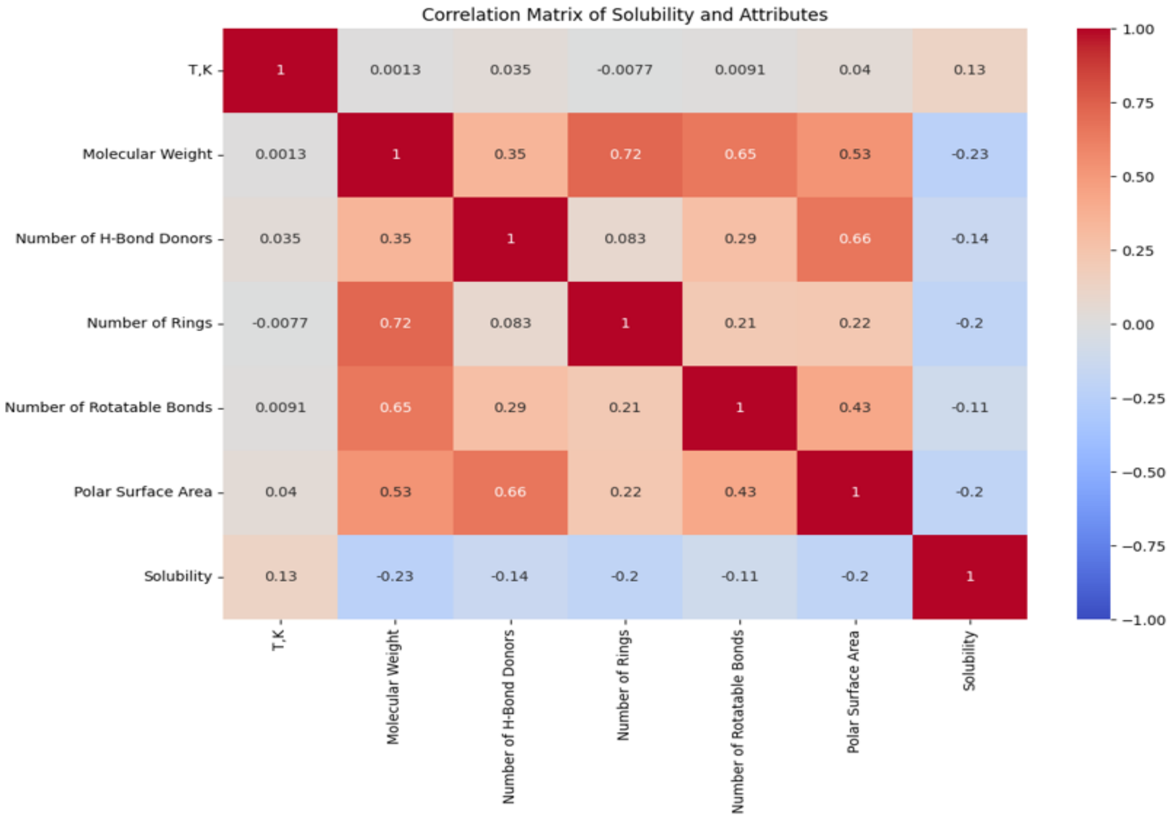


**Fig 4.6** Correlation matrix of solubility and the other attributes

*Analysis:*

In the analysis of various molecular factors influencing solubility, temperature (T, K) exhibits strong correlation with solubility in the dataset, suggesting that solubility may significantly change with variations in temperature. Molecular weight displays a negative correlation, indicating that lower molecular weight compounds tend to be more soluble, aligning with the general expectation that smaller molecules exhibit higher solubility. The number of H-bond donors correlates negatively with solubility, suggesting that compounds with more hydrogen bond donors may have lower solubility, while those with fewer donors might exhibit higher solubility. Similarly, the number of rings exhibits a negative correlation, implying that compounds with more rings might have lower solubility, possibly due to complex molecular structures with multiple rings contributing to reduced solubility. The number of rotatable bonds shows a weak negative correlation, suggesting that compounds with more rotatable bonds might have slightly lower solubility; however, the correlation is not very strong, indicating a nuanced relationship. Finally, polar surface area correlates negatively with solubility, supporting the idea that compounds with a larger polar surface area tend to have lower solubility, consistent with the notion that increased polar surface area could lead to decreased solubility.

### 4.7 Box Plot Insights: Solubility Trends in Ethanol and Water, Molecular Weight Variability, and Hydrogen Bond Donor Distribution

Analyzing the box plots (**Fig 4.7.1**) illuminates distinct differences and notable similarities in the solubility trends of ethanol and water concerning temperature. Ethanol's solubility consistently rises with increasing temperature, while water's solubility remains relatively constant. This disparity stems from variations in intermolecular forces, particularly the strength of hydrogen bonds. Ethanol's weaker bonds become more easily overcome at higher temperatures, leading to increased solubility, whereas water's robust hydrogen bonds remain resilient. Additionally, ethanol displays a broader solubility range compared to water, suggesting its capacity to dissolve a more diverse range of compounds. The boxplot shapes reveal that the spread of ethanol data points is wider at lower temperatures, possibly due to outliers or a broader range of solubilities. In contrast, the symmetrical shape of the water boxplot implies a more consistent distribution of solubility values across temperatures.

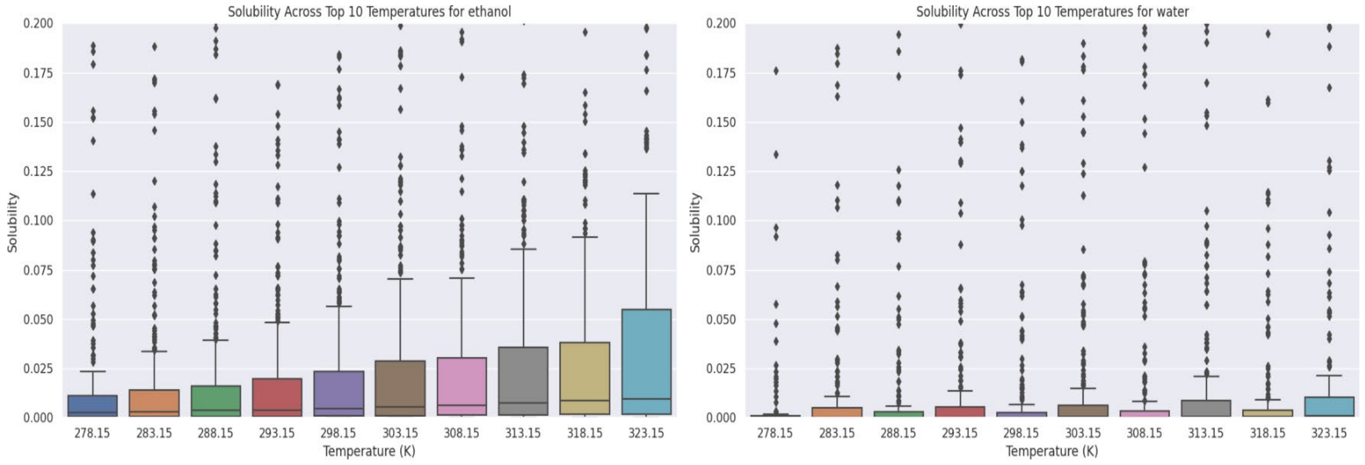


**Fig 4.7.1** BoxPlots:Solubility Across Top 10 Temperatures for Ethanol and Water

Turning to the box plot for molecular weight (**Fig 4.7.2)**, it offers insights into central tendencies, spread, variability, and the presence of outliers. The median of approximately 211.13 indicates an even distribution, while the interquartile range (IQR) suggests variability in molecular weights. The slight right skewness implies a prevalence of high molecular weight compounds. Outliers beyond the whiskers highlight significantly different molecular weights. Shifting focus to the box plot for the number of hydrogen bond donors (**Fig 4.7.3)**, the median of 1 suggests that half of the molecules have 1 or fewer donors, while the IQR of 1 indicates variability. The right skewness signifies a higher prevalence of molecules with more hydrogen bond donors. Additional box plots for the number of rings with median equals to 2 (**Fig 4.7.4)**, rotatable bonds with median equals to 2 (**Fig 4.7.5)**, and polar surface area with median equals to 63.32 (**Fig 4.7.6)**, contribute to a comprehensive understanding of solubility patterns and molecular characteristics. These analyses collectively provide valuable insights into the dataset, enhancing our comprehension of factors influencing solubility.

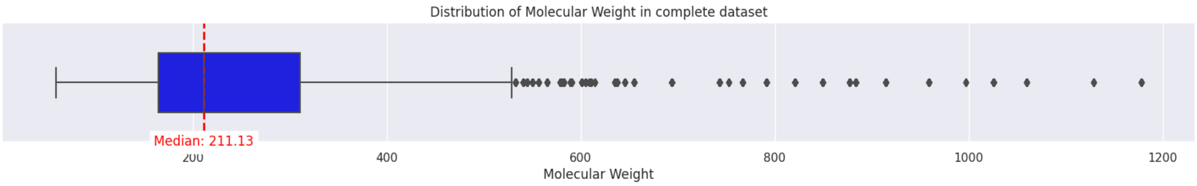
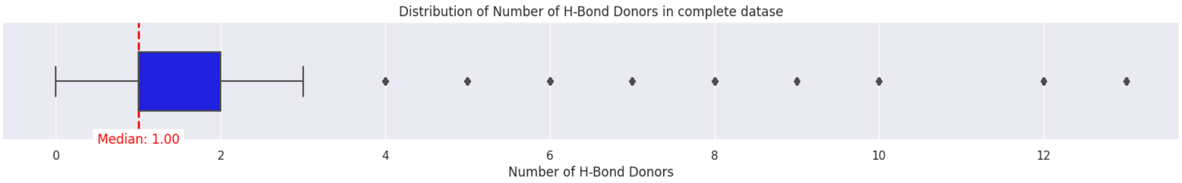


**Fig 4.7.2** BoxPlot:Distribution of Molecular Weight in the complete dataset



**Fig**

**4.7.3** BoxPlot:Distribution of Number of H-Bond Donors in the complete dataset

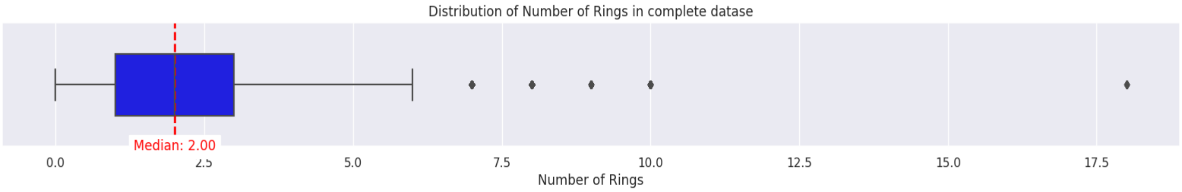


**Fig 4.7.4** BoxPlot:Distribution of Number of Rings in the complete dataset

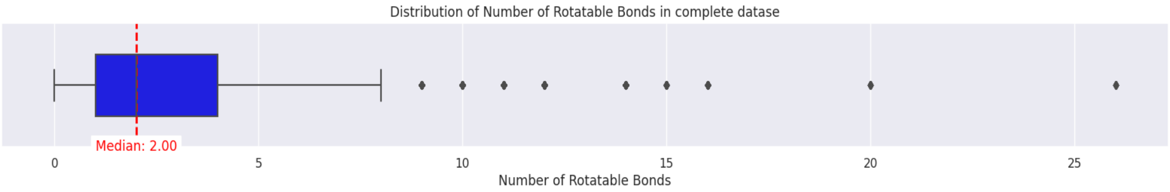


**Fig 4.7.5** BoxPlot:Distribution of Number of Rotatable Bonds in the complete dataset

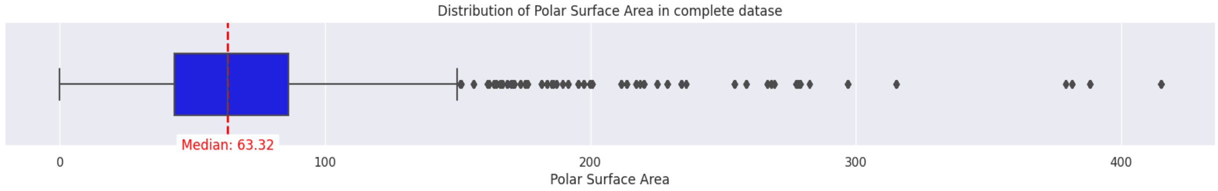


**Fig 4.7.6** BoxPlot:Distribution of Polar Surface Area in the complete dataset

## 5.Feature selection:

While feature selection can simplify models and potentially improve generalization, we chose to retain all features due to their established domain relevance in solubility prediction. Each attribute, from SMILES representation and solvent characteristics to molecular properties like size and polar surface area plays a crucial role in the solubility prediction.

Discarding any feature could inadvertently remove valuable information contributing to the intricate interplay between solute and solvent.

Additionally, considering our relatively small initial feature set of nine attributes, the risk of overfitting due a huge number of attributes is already mitigated.

# 6.Regression:

### 6.1. Dataset splitting strategy:

The technique of Stratified 5-Fold Cross-Validation is utilized on the entire dataset, it is essentially dividing the dataset into 5 folds in such a way that the proportion of values in the 'SMILES_Solvent_encoded' attribute is roughly the same in each fold. In each iteration, one-fold is used as the validation set, while the remaining folds collectively serve as the training set. Given this methodology,it avoids the need for a single split of the data into a fixed training and validation set. The Stratified Fold Cross-Validation is beneficial because it maximizes the usage of the available data for training and testing, providing a more reliable estimation of the model's performance on unseen data.

### 6.2. Evaluation criteria:

In machine learning, the coefficient of determination ($R^2$) and the Explained Variance Score (EVS) are common metrics used to evaluate the linear relationship between the true values and the predicted values of the model. An $R^2$ or an EVS of 1 means that the model perfectly predicts the target variable. In addition, mean absolute error (MAE) and mean squared error (MSE) were also commonly used as model performance metrics. They measure the mean of the absolute errors and the mean of the squared errors between the true values and the model predicted values on the data set with lower values indicating better model performance. They are defined as follows:

$$MAE = \frac{\sum_{i=1}^{n} \left|\hat{y}_i - y_i\right|}{n}$$

$$MSE = \frac{\sum_{i=1}^{n} \left(\hat{y}_i - y_i\right)^2}{n}$$

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

$$explained\ variance(y, \hat{y}) = 1 - \frac{Var(y - \hat{y})}{Var(y)}$$

### 6.3. Training models

In this study linear regression, Random Forest, K-Nearest Neighbors, Support Vector Machine were used to predict the solubility of compounds in various organic solvents and at different temperatures. These learning algorithms incorporate linear and nonlinear methods, as well as ensemble learning methods and traditional learning methods such as linear regression and support vector machines.

**Performances of different machine learning algorithms**

The Table showed the accuracies of the four machine learning models in the five-fold cross-validation test.

| ML Algorithms | $R^2$ | EVS | MSE | MAE |
|---|---|---|---|---|
| Linear Regression | 0.089360 | 0.089394 | 0.009851 | 0.062022 |
| Random Forest | 0.911278 | 0.910845 | 0.000964 | 0.011038 |
| KNN | 0.781659 | 0.781984 | 0.002358 | 0.021305 |
| SVM | -0.113508 | 0.087543 | 0.012043 | 0.089925 |
| XGBoost | 0.844752 | 0.844934 | 0.001677 | 0.022384 |

In analyzing the performance of various regression models on the dataset, it is evident that Random Forest and XGBoost stand out as top performers, where the model's prediction to the true values is higher. This superior performance may be attributed to several factors. Firstly, Random Forest and XGBoost excel in capturing complex and non-linear relationships within the data, making them well-suited for scenarios where the true relationship between features and the target variable is intricate. Additionally, these ensemble methods demonstrate robustness to overfitting, mitigating the risk of capturing noise and outliers in the data by building multiple trees with different subsets of data and features. In contrast, Linear Regression is inherently sensitive to outliers and may fail in the presence of influential data points or intricate feature interactions (complex relationships). Moreover, if the true relationship is non-linear, it might falter in capturing the inherent complexity, KNN, sensitive to noisy data, might yield suboptimal predictions, and SVM, assuming a linear relationship, may face challenges when the underlying relationship between features and the target is non-linear. This nuanced understanding of each model's strengths and limitations is crucial for selecting the most appropriate algorithm based on the characteristics of the data at hand.

# 7.Classification:

In the classification phase of our data mining project, we aimed to predict the solubility class of chemical compounds based on various features. For the sake of encoding the smiles and the smiles solvents we used the one hot encoding

## 7.1.One Hot Encoding

Since the 'SMILES' of the solute and the solvent are representations of chemical structures, and these structures typically don't have inherent ordering (they're considered nominal categories), one-hot encoding is a suitable strategy. By using one-hot encoding, we ensure that each element or character within the 'SMILES' strings is represented independently as a binary feature. This prevents the machine learning model from interpreting the encoded values as having any ordinal relationship or incorrect numerical order, which might occur if you used label encoding.

In the notebook we define **smiles_encoder** function that demonstrates the essence of one-hot encoding by converting SMILES strings into one-hot encoded matrices. This function creates a matrix where each row represents a unique character, and each column represents the position of the character in the string. The presence of a character in the string is indicated by a '1' in the corresponding position of the matrix, while the absence is represented by '0'.

## 7.2.Training models:

After thorough data preprocessing, including the creation of a categorical variable '*Solubility_Class'* with distinct class intervals, we divided our dataset using a *StratifiedShuffleSplit* technique. This ensured a balanced distribution of solubility classes in both the training+validation and test sets. The training+validation set was further split into training and validation sets to facilitate model development and fine-tuning. To address the multi-class classification task, we employed a range of classifiers, including Random Forest, k-Nearest Neighbors, Artificial Neural Network and XGBoost. A comprehensive evaluation approach was implemented, assessing model performance on training, validation, and test sets. Metrics such as accuracy and classification report were employed to gauge the models' effectiveness.

$$\text{Accuracy} = \frac{CorrectPredictions}{TotalSamples}$$

**Performances of different machine learning algorithms**

| Metric | Random Forest | KNN | ANN | XGBoost |
|---|---|---|---|---|
| Training Accuracy | 0.9355 | 0.9099 | 0.8337 | 0.9056 |
| Validation Accuracy | 0.8872 | 0.8806 | 0.8372 | 0.8930 |
| Testing Accuracy | 0.8852 | 0.8852 | 0.8323 | 0.8905 |
| Recall(macro avg) | 0.65 | 0.63 | 0.35 | 0.66 |
| F1-Score( macro avg) | 0.69 | 0.68 | 0.35 | 0.70 |
| Precision (macro avg) | 0.75 | 0.77 | 0.45 | 0.77 |

*Analysis:*

XGBoost appears to be the model that best fits the data. Here's a breakdown of the key points supporting this analysis:

- Consistently High Accuracy: XGBoost maintains high accuracy across training (0.9056), validation (0.8930), and testing (0.8905) sets, indicating it generalizes well to unseen data and isn't overfitting.
- Strong Recall, F1-Score, and Precision: XGBoost also performs well in terms of recall (0.66), F1-Score (0.70), and precision (0.77), balancing its ability to correctly identify true positives with its ability to avoid false positives.
- Comparative Performance: While Random Forest has slightly higher training accuracy, its validation and testing accuracies are lower than XGBoost. KNN and ANN have lower accuracies overall and weaker performance in recall, F1-Score, and precision.

To guarantee that the results of performance are accurate we evaluated the model with 5 fold_cross-validation to provide a robust assessment of each classifier's performance, revealing insights into their generalization capabilities. This comprehensive classification approach allows us to select the most suitable model for predicting solubility classes in our chemical compounds' dataset.

| K | Random Forest | KNN | ANN | XGBoost |
|---|---|---|---|---|
| 1 | 0.90211335 | 0.89759846 | 0.86993276 | 0.89740634 |
| 2 | 0.90921318 | 0.90988568 | 0.84254011 | 0.90268037 |
| 3 | 0.90056682 | 0.90123931 | 0.85685465 | 0.89461043 |
| 4 | 0.86694207 | 0.86982419 | 0.84542223 | 0.86502066 |
| 5 | 0.81938707 | 0.80949179 | 0.78316841 | 0.8137189 |
| Mean Accuracy | 0.8796 | 0.8776 | 0.8396 | 0.8747 |
| Std Accuracy | 0.0335 | 0.0366 | 0.0298 | 0.0332 |
| Precision( Macro avg) | 0.72 | 0.70 | 0.65 | 0.72 |
| Recall ( Macro avg) | 0.63 | 0.62 | 0.37 | 0.62 |
| F1 Score ( Macro avg) | 0.67 | 0.65 | 0.38 | 0.66 |

*Analysis:*

Based on the 5-fold cross-validation results, Random Forest appears to be the model that best fits the data overall. Here's a breakdown of the key factors:

Mean Accuracy:

- Random Forest has the highest mean accuracy (0.8796) across all folds, indicating its consistent performance across different data splits.
- XGBoost (0.8747) is very close behind, suggesting a similar level of generalization.
- KNN (0.8776) and ANN (0.8396) have lower mean accuracies.

Accuracy Standard Deviation:

- Random Forest has a slightly lower standard deviation in accuracy (0.0335) compared to XGBoost (0.0332), implying slightly more consistent performance across folds.

Precision, Recall, and F1-Score:

- Random Forest and XGBoost have essentially equivalent scores for precision (0.72), recall (0.63), and F1-score (0.67), demonstrating a good balance between correctly identifying true positives and avoiding false positives.
- KNN and ANN have lower scores in these metrics.
- Comparative Performance: While XGBoost was initially favored based on single evaluation metrics, Random Forest's slightly higher mean accuracy and consistency across folds in cross-validation suggest it might be a better choice for this specific dataset.

### 7.3. Conclusion of the analysis:

The evidence strongly suggests that either Random Forest or XGBoost would be the most suitable choice for the solubility dataset. Here's a summary of their strengths and considerations:

Random Forest:

- Strong Performance: Exhibits the highest mean accuracy and consistency across 5-fold cross-validation.
- Interpretability: Offers better feature importance insights for understanding model behavior.
- Less Sensitive to Outliers: Handles outliers well due to its ensemble nature.

XGBoost:

- High Accuracy: Achieves accuracy very close to Random Forest, demonstrating excellent predictive power.
- Computational Efficiency: Often handles large datasets more efficiently than Random Forest.
- Fine-Grained Control: Allows for more control over the model's complexity and learning process through its hyperparameters.

## 8.Conclusion

In conclusion, our project represents a significant advancement in chemical compounds solubility screening using data mining and machine learning. Leveraging the "BigSolDB" dataset with RDKit, we crafted a robust dataset for predicting solubility in various organic solvents across a wide temperature range. Through meticulous data cleaning and exploration, we uncovered key insights into the complex relationships between solubility, molecular attributes, and experimental conditions.

Our exploratory data analysis (EDA) has unveiled crucial insights into solubility trends and experimental conditions. The dataset exhibits a stark bias towards the top 50 solvents, predominantly featuring ethanol, methanol, and isopropanol. Notably, temperature, particularly at 303.15 K with ethanol, plays a pivotal role in solubility. Smaller molecular weight compounds tend to be more soluble, and there's a strong positive correlation between temperature and solubility, except for water. Feature correlations suggest the importance of certain molecular properties, guiding our considerations for model training. To address data imbalance, we employed stratification to ensure a well-balanced distribution of the data. Model selection will prioritize handling imbalanced data, and interpretability measures will ensure robust decision-making, guarding against spurious correlations and biases.

The Random Forest and XGBoost machine learning models have demonstrated notable efficacy in predicting solubility values, showcasing superior performance compared to K-Nearest Neighbors (KNN) and Support Vector Machines (SVM). The success of Random Forest and XGBoost suggests a non-linear relationship within our dataset. To enhance the robustness of our model evaluation, we employed Stratified Cross-validation using the solvent attribute rather than the solubility attribute. This strategic choice ensures an even distribution of solvents across each fold. This approach aligns with the primary objective of our project, which is to predict the solubility of a compound across various solvents.

Our data mining project to predict the solubility of chemical compounds in different solvents and temperatures has the potential to make a significant impact across various fields, particularly in chemistry and beyond. Here's a breakdown of the potential benefits:

Impact on Chemistry:

- Accelerated Drug Discovery: Predicting solubility can guide the design of new drugs with optimal bioavailability and delivery properties, speeding up the drug discovery process and potentially bringing life-saving medications to patients faster. Here PFIZER is a prominent example of well-known company actively seeking to improve the solubility of their products:

  PFIZER Challenge: Developing new drugs with poor water solubility, hindering their absorption and bioavailability in the body

- Improved Chemical Processes: Solubility prediction can optimize reaction conditions, solvent selection, and crystallization processes in the chemical industry, leading to increased efficiency, reduced waste, and cost savings.

- Fundamental Understanding: By analyzing the relationships between molecular properties, solvent interactions, and temperature effects on solubility, your project can contribute to a deeper understanding of solution thermodynamics and intermolecular forces in chemistry.

Impact on Other Fields:

- Environmental Science: Predicting the solubility of pollutants and contaminants in water and soil can inform environmental cleanup strategies, assess potential risks, and develop remediation technologies.
- Agriculture: Understanding the solubility of fertilizers and pesticides in different soil conditions can optimize agricultural practices, improve crop yields, and minimize environmental impact.
- Food Science: Solubility prediction can guide the development of new food formulations, improve stability and shelf life of food products, and optimize extraction processes for food ingredients. In addition, Food companies want ingredients that dissolve easily and consistently, but some face limited solubility challenges. Our project can be a valuable tool to solve these problems and these are examples of companies that it facing solubility challenge
    - Nestlé: Needs fast-dissolving ingredients in instant coffee and drinks, but flavors and sweeteners might be tricky. This project can predict the best solvent and temperature for optimal solubility.
    - Kraft Heinz: Wants stable sauces and dressings with specific textures, but thickeners and starches can be solubility hurdles. Our tool can help them find the right conditions for smooth, shelf-stable solutions.
    - Danone: Aims to enrich yogurt with vitamins and minerals, but dairy can limit their solubility.  can predict how to encapsulate or modify them for better dispersion and absorption.

## Bibliography:

- Tayyebi, A., Alshami, A., Rabiei, Z., Yu, X., Ismail, N., Talukder, M. J., & Power, J. (2023). Prediction of organic compound aqueous solubility using machine learning: a comparison study of descriptor-based and fingerprints-based models. *Journal of Cheminformatics*, *15*(1). https://doi.org/10.1186/s13321-023-00752-6

- Boobier, S., Hose, D. R. J., Blacker, A. J., & Nguyen, B. (2020). Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-19594-z

- Garzon, L. C. A., & Martínez, F. (2004b). Temperature dependence of solubility for ibuprofen in some organic and aqueous solvents. *Journal of Solution Chemistry*, *33*(11), 1379–1395. https://doi.org/10.1007/s10953-004-1051-2

- Krasnov, L., Mikhaylov, S., Fedorov, M. V., & Sosnin, S. (2023). BigSolDB: Solubility Dataset of Compounds in Organic Solvents and Water in a Wide Range of Temperatures. *Unknown*. https://doi.org/10.26434/chemrxiv-2023-qqslt