Project Proposal: Predicting Solubility of Chemical Compounds Using Machine Learning Approaches

Abstract

Predicting the aqueous solubility of chemical compounds in water influences drug design and many types of chemical research.it application in drug discovery,enviournmental science and other field it help us to identigy that chemical compund is solubale or not in water which is important. Due to the benefits of using application techniques with machine learning combinations together with collaboration with the AqSolDB dataset, this project constructs predictive models of solubility using the generation of molecular descriptors with the help of the advanced regression models as well as calculations with the RDKit software. It even validates a prediction by testing it on models like Random Forest, Gradient Boosting, and Stacking Regressors with metrics MSE and R^2..

1. Keywords

Aqueous solubility, Molecular Descriptors, RDKit, AqSolDB, min max scaling,feature scaling

2. Background and Motivation

From among a very wide range of topics, from drugs to environmental science, and from chemical engineering to applied physics, one relevant property has turned out to be solubility in an aqueous medium. Motivation is the building robust model to give accurate result with high performance Indeed, even somewhat better predictions for solubility might save significant efforts and time connected with expensive experiments. The current project uses the readily available database AqSolDB together with tools like RDKit in cheminformatics to predict solubility applying various machine learning models.

Previous studies, such as Tayyebi et al. (2022), have highlighted the potential of descriptor-based and fingerprint-based machine learning models for solubility prediction. This project builds on this foundation, comparing diverse modeling strategies and optimizing performance using advanced techniques like feature engineering, ensemble learning, and hyperparameter tuning.

3. Objectives

1. **Primary Objective**: Develop and optimize machine learning models for predicting solubility of chemical compounds.
2. Compare the performance of various machine learning approaches:
    - Linear regression
    - Random Forest
    - Gradient Boosting
    - XGBoost
    - Support Vector Regression (SVR)

- Ensemble models (Stacking, Voting, Bagging)

3. Evaluate the impact of molecular descriptors generated using RDKit on model performance.
4. Identify key features influencing solubility predictions through feature engineering and selection techniques.

4. Methodology

4.1 Dataset

- **Source**: AqSolDB 【Palchem et al., 2019】.
- **Size**: 9,982 entries with 26 features, including molecular weight (MolWt), polar surface area (TPSA), and solubility as the target variable.
- **Tools**: RDKit for molecular descriptor computation, Python for preprocessing and model implementation.

4.2 Data Preprocessing

1. Removed non-numeric columns (e.g., ID, Name, SMILES).
2. Encoded categorical features (Group) using LabelEncoder and OneHotEncoder.
3. Normalized numerical features using StandardScaler and MinMaxScaler.
4. Engineered additional features such as MolWt_TPSA_Ratio and polynomial interaction terms.

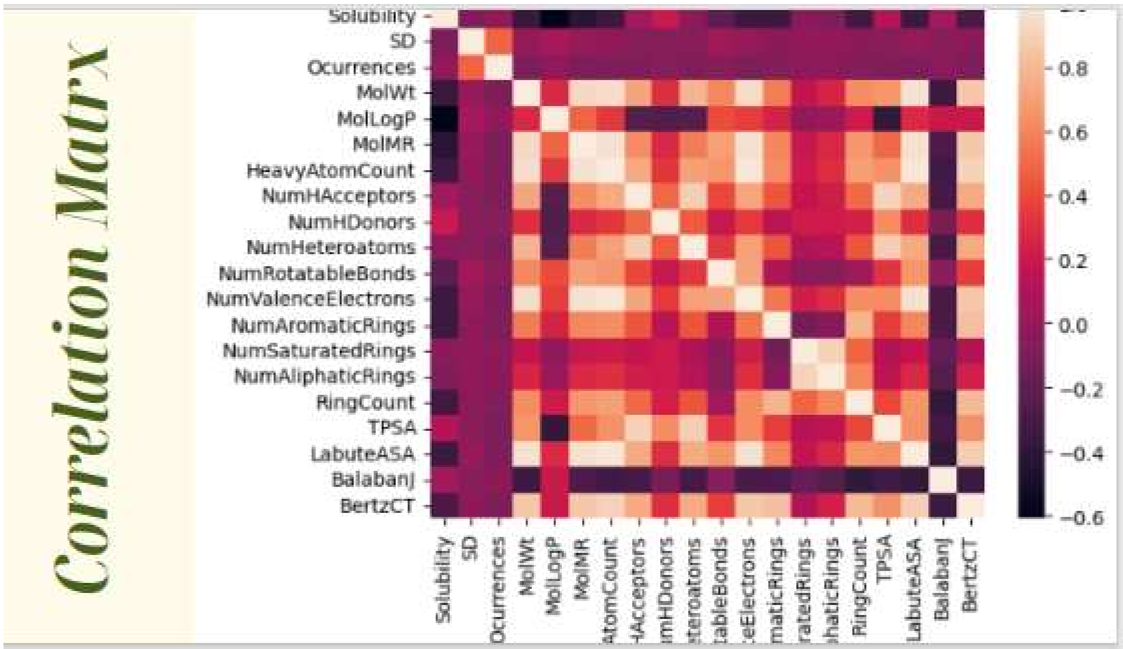
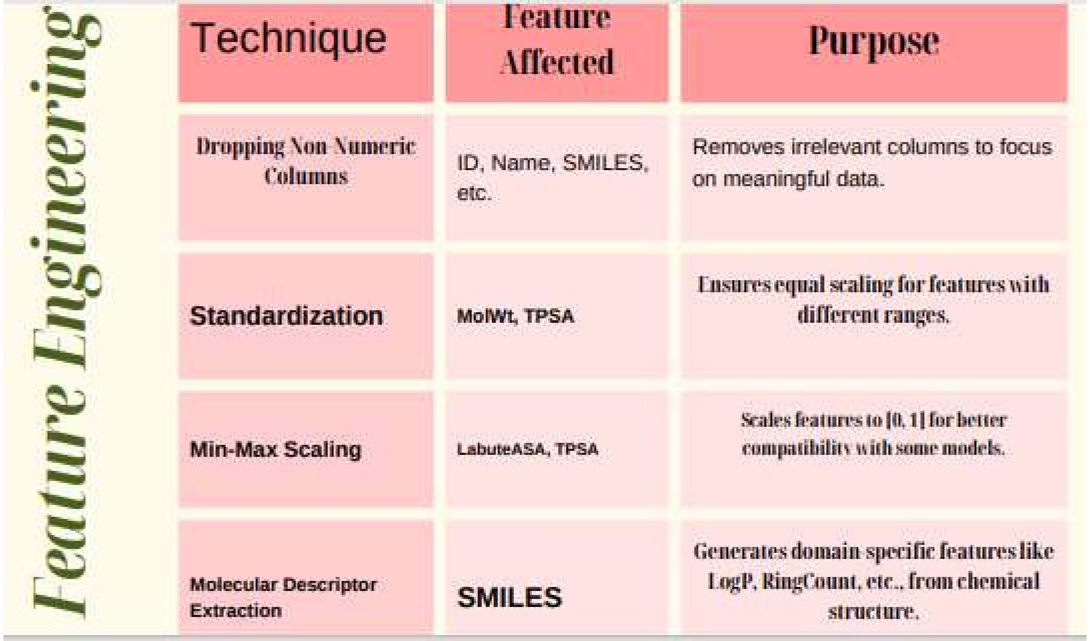


Figure 1

4.3 Feature Engineering

- Molecular descriptors (e.g., MolWt, TPSA, LogP) calculated using RDKit.
- Recursive Feature Elimination (RFE) was used to select the top 10 most impactful features.



Feature Engineering

Technique	Feature Affected	Purpose
Dropping Non-Numeric Columns	ID, Name, SMILES, etc.	Removes irrelevant columns to focus on meaningful data.
Standardization	MolWt, TPSA	Ensures equal scaling for features with different ranges.
Min-Max Scaling	LabuteASA, TPSA	Scales features to [0, 1] for better compatibility with some models.
Molecular Descriptor Extraction	SMILES	Generates domain-specific features like LogP, RingCount, etc., from chemical structure.

Figure 2

4.4 Machine Learning Models

1. **Baseline Models**:
   - Linear Regression
   - Decision Trees
2. **Advanced Models**:
   - Random Forest
   - Gradient Boosting
   - XGBoost
   - Support Vector Regression (SVR)
3. **Ensemble Techniques**:
   - Stacking Regressor with Gradient Boosting as the meta-model.
   - Voting Regressor combining Random Forest, Gradient Boosting, and XGBoost.
   - Bagging Regressor with Decision Trees and SVR.

4.5 Hyperparameter Optimization

- **GridSearchCV** was used to tune Random Forest parameters:
  - `n_estimators`: [100, 200, 500]
  - `max_depth`: [5, 10, 15]
  - **Best Parameters**: `n_estimators=500`, `max_depth=10`, `min_samples_split=10`, `min_samples_leaf=2`.

4.6 Evaluation Metrics

- Mean Squared Error (MSE)
- Coefficient of Determination (R^2)

5. Results and Evaluation

Insights:

- Stacking Regressor achieved the best performance (MSE: 1.1573, R^2: 0.7867).
- Random Forest and Voting Regressor demonstrated high generalizability.
- SVR underperformed due to its sensitivity to feature scaling.

6. Discussion

6.1 Trends

- Ensemble techniques consistently outperformed individual models, with stacking showing the best overall performance.
- Feature selection reduced dimensionality without compromising model accuracy, validating the importance of descriptors like TPSA, MolWt, and LabuteASA.

6.2 Anomalies

- SVR struggled with solubility predictions, likely due to high-dimensional feature space and non-linear relationships.
- Gradient Boosting's performance was slightly lower than Random Forest and XGBoost, potentially due to overfitting or limited hyperparameter tuning.

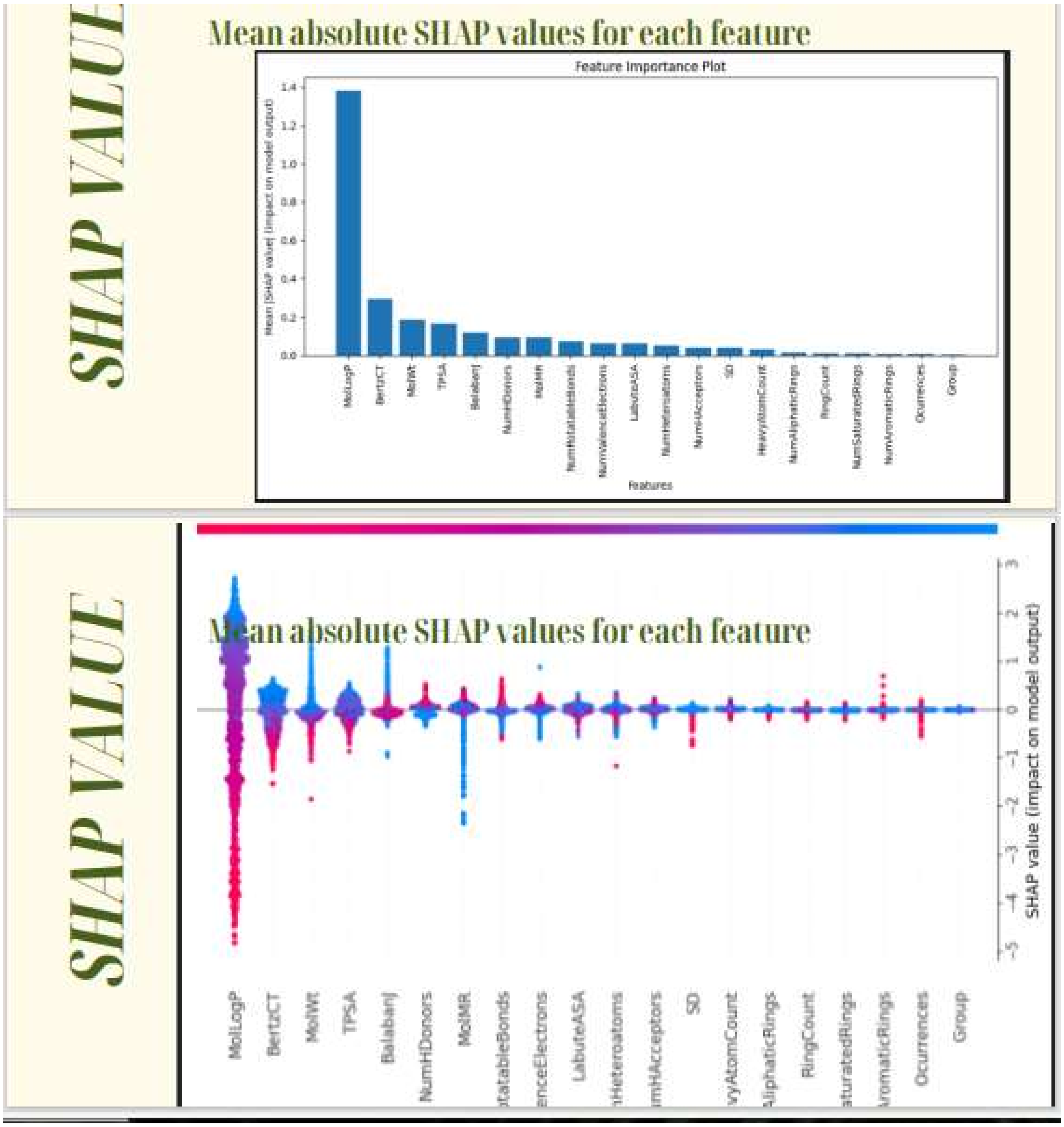


Figure 3

6.3 Conclusions

- Ensemble learning approaches, especially stacking and voting regressors, are highly effective for solubility prediction.
- RDKit descriptors play a critical role in capturing molecular properties influencing solubility.

7. Future Scope

7.1 Follow-Up Studies

1. **Fingerprint-Based Features**:
    - Explore synergies between descriptor-based and fingerprint-based models (e.g., Morgan, MACCS).
2. **Quantum Chemical Properties**:

    - Incorporate properties like dipole moment and polarizability using quantum chemistry tools.
3. **Deep Learning Models**:
    - Investigate Graph Neural Networks (GNNs) or CNNs for direct processing of molecular graphs or SMILES.

7.2 Improvement Directions

1. **Interpretability**:
    - Use SHAP or LIME to gain insights into feature contributions.
2. **Advanced Optimization**:
    - Experiment with Bayesian optimization for hyperparameter tuning.
3. **Uncertainty Quantification**:
    - Predict uncertainty in solubility estimates using probabilistic models.
4. **Real-World Integration**:
    - Collaborate with experimental chemists to validate predictions and deploy models via cloud-based APIs.

8. Conclusion

- Stacking Regressor proved the most effective, achieving the highest R^2 of 0.7867.
- Advanced feature engineering and RDKit descriptor computation significantly enhanced model performance.
- The study demonstrates machine learning’s potential to accelerate solubility predictions, aiding drug discovery and chemical research.

References

1. Palchem, A., et al. (2019). **AqSolDB: The largest publicly available dataset of aqueous solubility.** *Nature Scientific Data*. DOI: 10.1038/s41597-019-0151-1.

2. RDKit: **Open-Source Cheminformatics Software.** Link: https://www.rdkit.org/.

3. Tayyebi, A., Alshami, A. S., Rabiei, Z., Yu, X., Ismail, N., Talukder, M. J., & Power, J. (2022). **Prediction of organic compound aqueous solubility using machine learning: a comparison study of descriptor-based and fingerprints-based models.** *Frontiers in Chemistry*. DOI: 10.3389/fchem.2022.1055542.

4. Chatgpt I have used in this project as reference we have mentioned