Predicting Aqueous solubility using machine learning Techniques

Sonu Kumar MT23144

Bhavya Gupta 2021458

CHI PROJECT

Introduction

Aqueous solubility is important steps for drug discovery , Metarial Science, Environmental studies becouse each chemical compound solubility depends on its Descriptors using various machine learning models and ensemble techniques.

Motivation:-

Building Machine Learning Model that accurately predicts the solubility of chemical compounds. This high-precision approach is designed to address the critical challenges in drug discovery.

Flowchart

Steps:-

Start -> Load Dataset -> Preprocess Data -> Feature Engineering -> Train-Test Split -> Define Model, Linear Regression, Random Forest ,Gradient Boosting, XGBoost, SVR -> Train Models Individually

-> Evaluate Performance

-> Optimize Parameters -> Ensemble Models

-> Generate Predictions -> Analyze Results

-> Submit Predictions -> End

DataSets

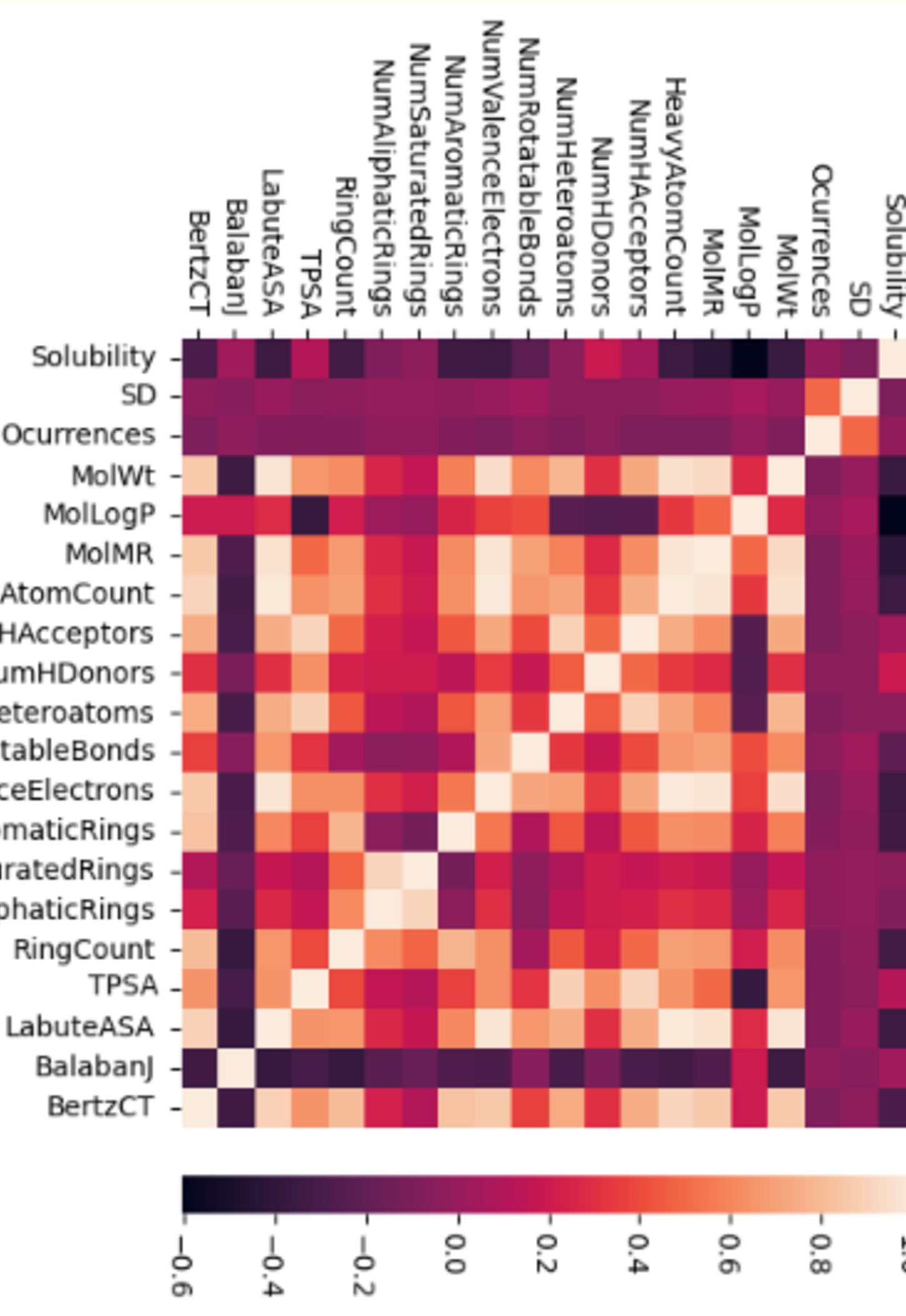
Analysis:-

- Name: AqSolDB (Aqueous Solubility Database).
- Size: 19,262 compounds.
- Source: Compiled from 9 publicly available datasets.
- source ID of compound (first letter indicates source)
- description of compounds (Name,InChI,InChIKey,SMILES)
- curated solubility value
- standard deviation and number of occurances of multiple solubility values
- reliability group other 2D descriptors calculated using RDKit
- DOI: https://doi.org/10.1038/s41597-019-0151-1
- Citation Paper: Nature Scientific Data - Citation Ahmad Elawady, Ahmed Elkerdawy, and George Iskander. Aqueous Solubility Predictioin. https://kaggle.com/competitions/aqueous-solubility-predictioin, 2023. Kaggle
- Reproducible code: Code Ocean - Citation Ahmad Elawady, Ahmed Elkerdawy, and George Iskander. Aqueous Solubility Predictioin. https://kaggle.com/competitions/aqueous-solubility-predictioin, 2023. KaggleReferences
- Boosting the predictive performance with aqueous solubility dataset curation
- AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set

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.

Correlation Matrx



Methodology

Data Pre-processing

- Remove Non Numeric Column like smiles
- Categorical Variable Encoding like Group
- Normalization feature like LabuteASA, TPSA
- Descriptors Extracted: Molecular weight, LogP, TPSA
- Polynomial Features and Interaction Terms
- Feature Selection based on corr matrix and RFE
- Train-Test Split
- The dataset was divided into training (80%) and testing (20%) sets using train_test_split.

Methodology

Model Development

Individual Models:

Trained individual models, including Linear Regression, Random Forest, Gradient Boosting, XGBoost, Decision Trees, and SVR.

Ensemble Techniques:

Stacking Regressor: Combined multiple base models (e.g., Random Forest, SVR) with a meta-model (Gradient Boosting) for final predictions.

Voting Regressor: Aggregated predictions from top-performing models (e.g., Random Forest, XGBoost, Gradient Boosting).

Bagging Regressor: Employed Decision Trees as base estimators to improve robustness.

Hyperparameter

Best paramaters for random forest

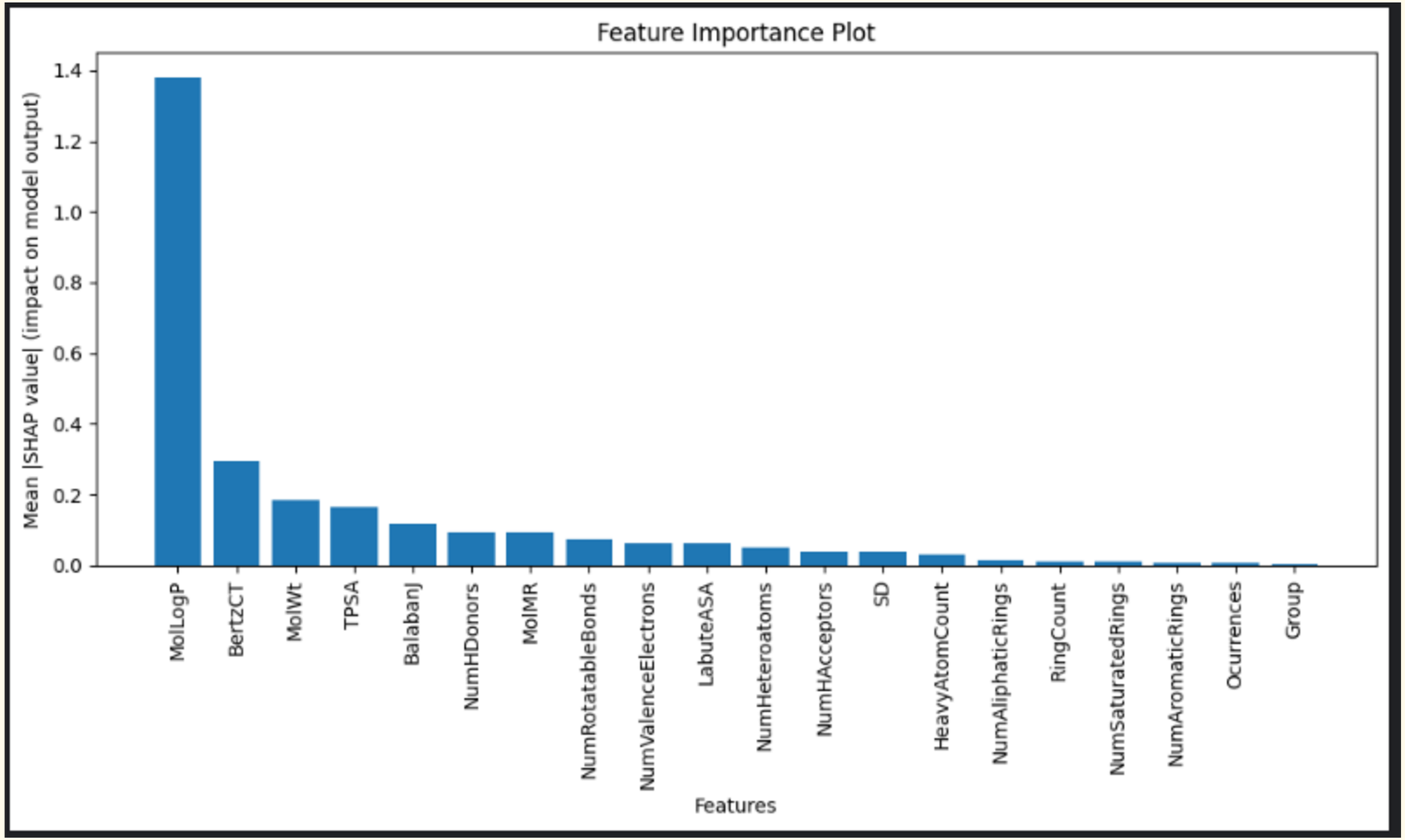
- Parameter Grid:
- n_estimators: [100, 200, 500]
- max_depth: [5, 10, 15]
- min_samples_split: [2, 5, 10]
- min_samples_leaf: [1, 2, 4]
- Best Parameters:
- n_estimators: 500
- max_depth: 10
- min_samples_split: 10
- min_samples_leaf: 2
- Outcome: Improved performance and generalizability.

Evaluation Metrics

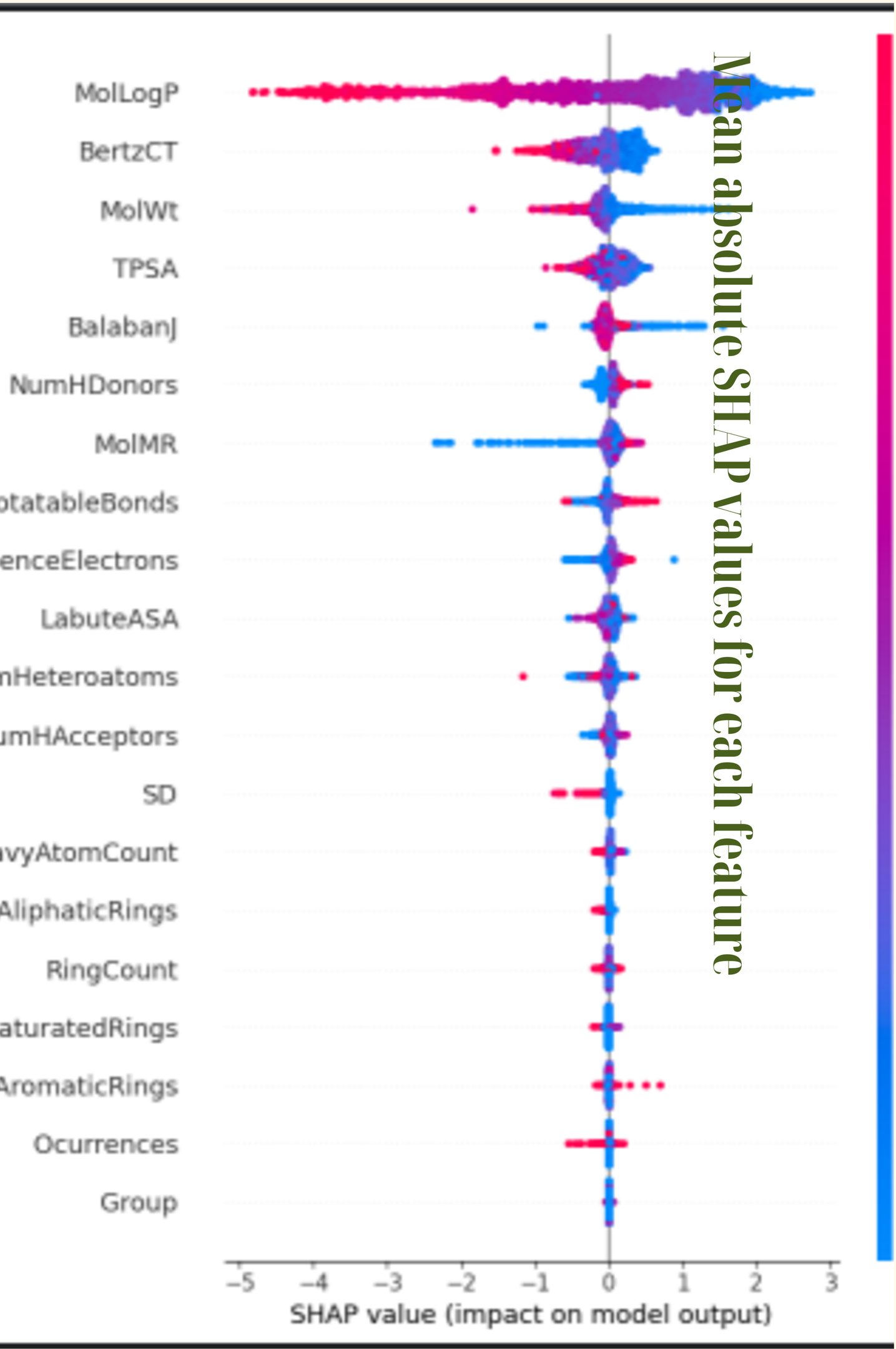
Model/Approach	MSE	R^2
Stacking Regressor (Approach 1)	**1.1573**	0.7867
Random Forest Regressor (Individual)	1.2877	0.7810
Gradient Boosting Regressor (Individual)	1.3573	0.7498
XGBoost Regressor (Individual)	1.2340	0.7725
Support Vector Regressor (SVR)	2.7686	0.4896
Bagging Regressor (Decision Tree)	1.1879	0.7810
Voting Regressor (All Models)	1.3410	0.7528
Top 3 Voting Regressor	1.1681	0.7847
RFE with Random Forest	1.3573	0.7666
Gradient Boosting Regressor (Boosting)	1.2663	0.7498

SHAP VALUE

Mean absolute SHAP values for each feature



SHAP VALUE


MolLogP
BertzCT
MolWt
TPSA
BalabanJ
NumHDonors
MolMR
otatableBonds
enceElectrons
LabuteASA
nHeteroatoms
umHAcceptors
SD
avyAtomCount
AliphaticRings
RingCount
aturatedRings
AromaticRings
Ocurrences
Group
−5
−4
−3
−2
−1
0
1
2
3
SHAP value (impact on model output)
Mean absolute SHAP values for each feature

Results

Stacking Regressor

MSE: 1.1573, R^2: 0.7867

Combining multiple models (Linear Regression, Random Forest, Gradient Boosting, etc.) significantly improved accuracy, indicating complementary strengths of the base models.

Recursive Feature Elimination (RFE)

MSE: 1.2663, R^2: 0.7666

RFE successfully identified the most relevant features, but the performance was slightly lower than stacking.

Results

Ensemble Voting Regressor

MSE: 1.2761, R^2: 0.7648

Voting among Random Forest, Gradient Boosting, and XGBoost provided a balance between predictions.

Bagging Regressor

MSE: 1.1879, R^2: 0.7810

Bagging performed well due to its robustness to overfitting, especially with Decision Trees.

Top 3 Model Voting Ensemble

MSE: 1.1681, R^2: 0.7847

Using the top three models optimized predictive accuracy further.

Trends and Anomalies

Ensemble methods consistently outperformed individual models, emphasizing the importance of combining complementary strengths.
Gradient Boosting and Random Forest excelled in handling complex, non-linear relationships.
SVR underperformed ($R^2 \approx 0.49$), likely due to its sensitivity to scaling and insufficient tuning for high-dimensional data.

Conclusions

Ensemble techniques (stacking and voting) were the most effective approaches for solubility prediction.
Feature engineering, particularly using molecular descriptors, was crucial for enhancing model performance.

Challenges

Degree of Difficulty

Handling high-dimensional molecular data required sophisticated preprocessing and feature selection techniques.

Ensemble model tuning demanded careful experimentation to optimize hyperparameters.

Innovation and Creativity in Approaches

Novel derived features like MolWt_TPSA_Ratio and polynomial interactions added predictive value.

Integration of RDKit descriptors offered domain-specific insights.

Complexity of Tasks

Implementing and comparing multiple ensemble strategies required extensive experimentation and validation.

Combining domain knowledge (chemistry) with machine learning techniques increased complexity.

Future Research

Analysis:-

- use transformer based model such as ROBERTA and BERT MODEL
- Advance technique for feature engineering
- explore the application of neural networks and deep learning, specifically Graph Neural Networks, to handle molecular data directly
- effective use of machine learning in chemical informatics, offering extensive opportunities for future enhancements and proving its worth as a valuable tool in the scientific community's toolkit.

References

- ***RDKit for Molecular Descriptors*:*AqSolDB Dataset*:**
- *Palchem, A. et al. "AqSolDB: The largest publicly available dataset of aqueous solubility."*
- *Nature Scientific Data (2019).*
- *DOI: https://doi.org/10.1038/s41597-019-0151-1.*
- *RDKit: Open-Source Cheminformatics Software.*
- *Link: https://www.rdkit.org/.*
- ***Comparison of Descriptor-Based and Fingerprints-Based Models*:**
- *Tayyebi, A., Alshami, A. S., Rabiei, Z., Yu, X., Ismail, N., Talukder, M. J., & Power, J.*
- *"Prediction of organic compound aqueous solubility using machine learning:*
- *a comparison study of descriptor-based and fingerprints-based models." Frontiers in Chemistry (2022).*
- *DOI: https://doi.org/10.3389/fchem.2022.1055542.*

Thank You