Predicting Toxicity: Accelerating Drug Discovery for CRY1 Targets

DATA 2206 – Professor Sam Plati

Ctrl Alt Data
Rohan Saldanha – 100976167
Jayanth Hassan Murali – 100994668
Ujjwal Vasava – 100976100
Akshaya Bhalikha Dhanasekaran – 100936892
Bhavanish S Nair – 100936855
Mustafa Khaja Masood Khaja – 100923081
Grifith Pereira – 100991416

Background

Problem Statement:

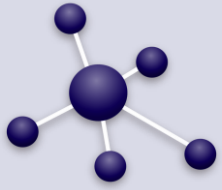
- Circadian rhythm disruptions cause severe health issues like cancer, mood disorders, and sleep disturbances.
- Traditional drug discovery is slow, costly, and prone to high failure rates due to toxicity risks. The model's performance will be evaluated using Recall, Precision, F1-score and Accuracy, to maximize the recall(for health-care purposes) and keep the precision high.

Objective:

- Develop a machine learning model to predict molecule toxicity, enabling faster and safer drug discovery.
- Identify molecular features critical for regulating the circadian rhythm.

Significance:

- Accelerates drug discovery by identifying toxic compounds early.
- Reduces costs and enhances efficiency in pharmaceutical research.
- Supports safer and more effective treatments for circadian rhythm disorders.



Data Overview

Data Source:

- Dataset containing 171 molecules with chemical descriptors.

Data Type:

- Structured dataset with numerical features representing molecular properties.

Key Variables:

- Features: Molecular descriptors (e.g., "MDEC-23," "C1SP2").
- Target: Toxicity classification (toxic or non-toxic).
- Data quality enhancements: SMOTE applied for class balance, outliers handled using Tukey's method.

Optimal number of features: 11
Optimized Model
Feature Importances

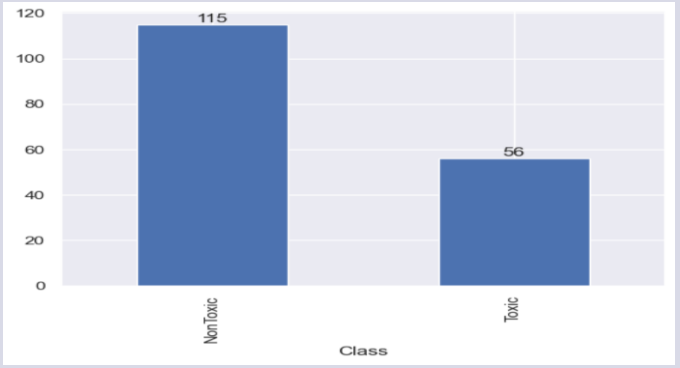
MDEC-23	**0.26**
MATS2v	0.0
ATSC8s	0.09
VE3_Dt	**0.17**
CrippenMR	0.1
SpMax7_Bhe	**0.12**
SpMin1_Bhs	0.0
C1SP2	**0.16**
GATS8e	0.0
GATS8s	0.0
SpMax5_Bhv	0.1

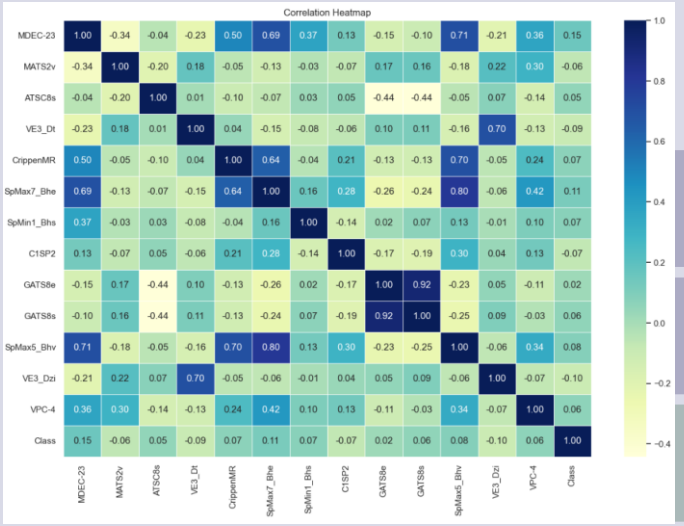
Exploratory Data Analysis

Class Imbalance: ‘Non-Toxic’ class dominates the dataset.

High Correlations: There were strong positive correlations between **GATS8e and GATS8s**, **SpMax7_Bhe and SpMax5_Bhv, and VE3_Dt and VPC-4** features**.**

Normality: Heterogeneous distribution data, were some features were normally distributed, some are right-skewed and some are left-skewed.





Modeling Approach

We evaluated several machine learning models, Logistic Regression, Decision Tree, and Random Forest, to classify molecules as either toxic or non-toxic based on their chemical properties. These models were selected to compare the effectiveness of linear versus tree-based approaches. Each model was trained and tested using consistent preprocessing and evaluation strategies to ensure fair performance assessment. Our objective was to determine the most accurate and reliable model for toxicity prediction.

Rationale for Selection:

1. Logistic Regression

Selected as a baseline model due to its simplicity and interpretability. Logistic Regression helps determine how well a linear model can separate toxic and non-toxic compounds using a probabilistic approach.

2. Decision Tree

Chosen for its capability to model non-linear relationships and interpret feature importance. Decision Trees are easy to visualize and provide clear decision paths, making them useful for understanding the behavior of the toxicity classification.

3. Random Forest

Implemented to enhance accuracy and reduce overfitting. By averaging multiple decision trees, Random Forest tends to generalize better and is more robust to noise and variance in the dataset.

Modeling Performace

Results for Logistic Regression: Accuracy of 49%

Logistic Regression performed the weakest among the three models with an accuracy below 50%. It struggled to distinguish toxic from non-toxic molecules, especially underperforming in identifying toxic samples (low precision and recall for class 1). This highlights that a linear approach may not be well-suited for capturing the complexity of the chemical features involved.

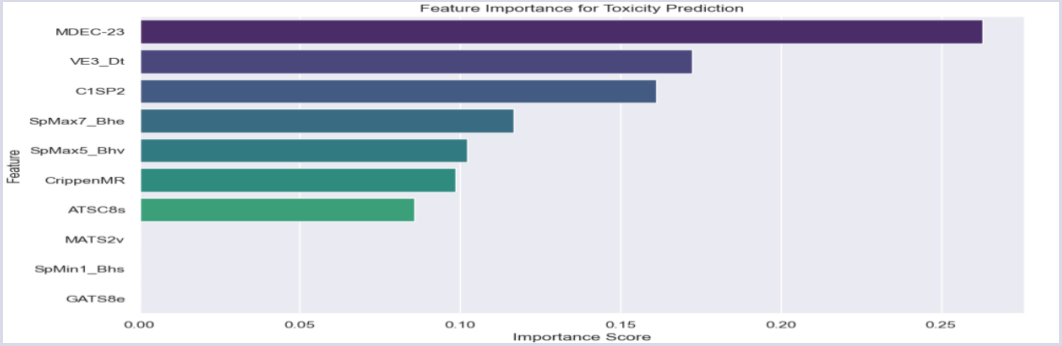
Results for Decision Trees: Accuracy of 80% | Recall = 91%

The Decision Tree model significantly outperformed Logistic Regression. With a high recall (0.91) for toxic samples, it effectively captured most of the toxic compounds, which is critical in safety-sensitive applications. However, its slightly lower precision for the toxic class (0.62) indicates some false positives. Overall, it offers a balanced performance and strong potential for real-world deployment.

Results for Random Forest: Accuracy of 77%

The Random Forest model provided consistent results with a strong overall accuracy of 77%. It maintained a good balance between precision and recall for both classes, making fewer errors than Logistic Regression and slightly more balanced than the Decision Tree. While not achieving the highest recall like the Decision Tree.

Feature Importance Graph



1. MDEC-23 (Highest importance = 0.26)

This is the most influential feature in determining molecular toxicity. It likely captures a key structural or electronic property affecting toxic behavior.

2. VE3_Dt and C1SP2 (0.17 – 0.16)

Both are highly relevant in contributing to accurate predictions. VE3_Dt may reflect molecular topology, while C1SP2 could relate to specific atom connectivity patterns.

3. SpMax7_Bhe, SpMax5_Bhv, and CrippenMR

These mid-tier features still play meaningful roles in the model's decision-making, possibly capturing molecular shape or polarity characteristics.

4. ATSC8s, MATS2v, SpMin1_Bhs, GATS8e (Lower importance)

These features contribute less to the overall model performance but may provide complementary information that improves prediction reliability.

Evaluation Metrics

Accuracy: Measures overall correct predictions. Decision Tree: ~80% , Random Forest: ~77%

Precision: This reflects the proportion of predicted toxic molecules that are truly toxic. Higher precision for class 0 in both models indicates fewer false positives.

Recall: A priority metric to avoid missing toxic compounds

Decision Tree: Recall for toxic class (1) is 0.91, meaning it correctly identified 91% of actual toxic samples.

Random Forest: Recall for class 1 is lower at 0.64, suggesting more toxic compounds were missed.

F1 Score: Balances precision and recall

Macro average F1: Decision Tree = 0.79, **Random Forest** = 0.73.

Confusion Matrix:

Decision Tree: Correctly predicted 19 non-toxic and 10 toxic samples; only 1 false negative (missed toxic) and 5 false positives.

Random Forest: More balanced but less effective—4 false positives and 4 false negatives, indicating lower reliability in detecting toxic cases.

```
[[18  6]
 [ 1 10]]
```

	precision	recall	f1-score	support
Non-Toxic	0.95	0.75	0.84	24
Toxic	0.62	0.91	0.74	11
accuracy			0.80	35
macro avg	0.79	0.83	0.79	35
weighted avg	0.85	0.80	0.81	35

```
[[20  4]
 [ 4  7]]
```

	precision	recall	f1-score	support
0	0.83	0.83	0.83	24
1	0.64	0.64	0.64	11
accuracy			0.77	35
macro avg	0.73	0.73	0.73	35
weighted avg	0.77	0.77	0.77	35

Evaluation Metrics

Key Findings:

- Tuned Decision Tree outperformed Random Forest in accuracy and recall.
- Recall was prioritized to reduce the chance of missing true toxic compounds.
- Features such as "MDEC-23" and "C1SP2" significantly influenced model predictions.
- Confusion matrices highlight Decision Tree's strength in minimizing false negatives, which is critical in toxicity detection.

Implications

Practical Impact:

- Faster Screening: Accelerates R&D by automating toxicity predictions.
- Cost & Resource Efficiency: Prioritizes high-risk compounds and reduces manual testing.
- Regulatory Compliance: Systematically flags toxic substances, ensuring safety standards.

Business & Scientific Benefits:

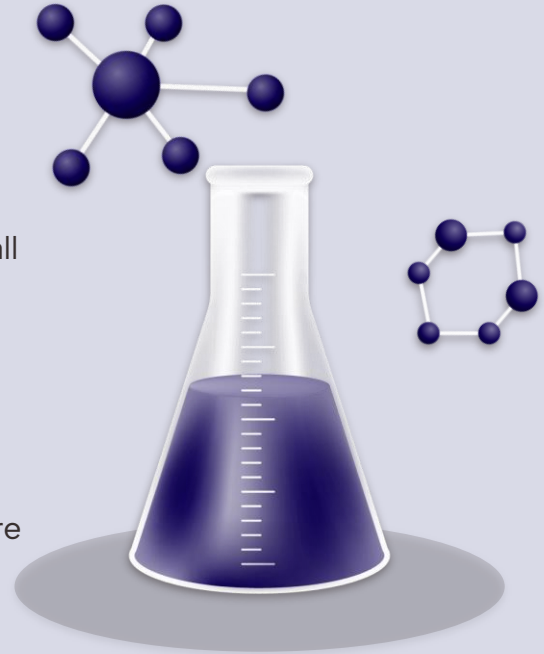
- Reduced Risk: Higher recall mitigates missed toxic compounds.
- Enhanced Decision-Making: Highlights crucial molecular attributes guiding toxicity.
- Scalability: Adapts to larger datasets or deep learning for fewer false negatives.

Recommended Next Steps:

- Data Enrichment: Increase toxic sample representation by gathering more minority-class instances.
- Advanced Methods: Explore ensemble approaches or deep learning techniques to further enhance recall.
- Explainability: Employ SHAP or LIME to provide clearer insights into model decision-making.

Recommendations

- **Prioritize top features ("MDEC-23", "C1SP2") for focused model tuning**
  Focusing on these features can reduce model complexity, improve interpretability, and speed up training without losing predictive power
- **Integrate advanced models**
  Gradient Boosting like XGBoost can handle non-linear patterns and small feature interactions better than simple trees
- **Explore alternatives to SMOTE**
  In order to address class imbalance, SMOTE generates synthetic data blindly. In contrast, ADASYN – Adaptive Synthetic Sampling makes improvements by concentrating on cases that are difficult to classify, improving minority class learning, and lowering false negatives, which are essential for toxicity detection



Conclusion

Summary

- For the best accuracy and interpretability, the Decision Tree was used together with Recursive Feature Elimination (RFE) and hyperparameter tuning. this model is an excellent place to start, particularly in high-stakes areas like toxicity prediction.

Final Thought

- Machine Learning can transform toxicity detection—faster, safer, and cost-effective with real-world impact in drug discovery and chemical safety. With improvements, such models can be production-ready for real-world use, potentially saving lives and reducing costs.