AMRITA VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY UNDER SECTION 3 OF UGC ACT, 1956

Alcohol vs Non-Alcohol Classification

Name: Machani Sri Balaji
Roll No: DL.AI.U4AID24119

Problem Statement

Develop a classification system that identifies whether a molecule contains a hydroxyl (-OH) functional group using graph-derived features.

Code Accessibility

The complete implementation, including data generation, feature engineering, and model evaluation, is available at the following repositories:

- **Google Colab:** https://colab.research.google.com/drive/1yfYbg-f-XtmQyd279kV5WBAMssOxZBM_?usp=sharing
- **GitHub Repository:** https://github.com/SriBalajiMachani/SMILES-SELFIES.git

Abstract

This report details the development of a binary classification system designed to identify aliphatic alcohols within a synthetic dataset of 200 molecules. By integrating sequence-based SELFIES tokens with structural graph descriptors, we evaluate multiple machine learning architectures. Our findings indicate that Logistic Regression and Naive Bayes provide superior generalization, achieving cross-validation accuracies above 90%.

1 Introduction

Functional group identification is critical for predicting molecular reactivity and biological activity. This study focuses on the detection of the hydroxyl group ($-OH$) when bound to sp^3 hybridized carbons. To ensure precision, we explicitly distinguish these from phenols (aromatic $-OH$) and carboxylic acids.

2 Dataset and Labeling

A balanced dataset ($\mathscr{D}$) of 200 molecules was generated. Labeling was performed using RDKit-based SMARTS patterns:

- **Target Class (Alcohol):** Molecules matching `[CX4][OX2H]`.
- **Exclusion Criteria:** Molecules containing `C(=O)[OX2H]` (Acids) or `c[OX2H]` (Phenols) were labeled as non-alcohol regardless of other features.

3 Methodology

3.1 Feature Engineering

The model utilizes a hybrid feature vector $\mathbf{x}_i$:

$$\mathbf{x}_i = [\mathbf{x}_{selfies} \parallel \mathbf{x}_{graph}] \tag{1}$$

3.1.1 SELFIES Vectorization

Unlike SMILES, SELFIES tokens are 100% robust to random mutation. We represent the molecular sequence using a Bag-of-Tokens approach:

$$\mathbf{x}_{selfies} = \sum_{j=1}^{N} \text{count}(t_j) \tag{2}$$

3.1.2 Graph-Derived Descriptors

For each molecular graph G, we extracted seven key statistics:

- **Degree Stats:** Mean degree $\bar{d}$, maximum degree d_{max}, and minimum degree d_{min}.
- **Heteroatoms:** $N_{het} = \sum_{v \in V} \mathbb{I}(\text{atom}(v) \notin \{C, H\})$.
- **Connectivity:** Number of rings, atoms, and bonds.

4 Experimental Results

4.1 Performance Metrics

Evaluation was performed on an 80/20 stratified split. The detailed classification metrics are presented in Table 1.

Model	Accuracy	F1 Alc.	F1 Non.	Prec. Alc.	Prec. Non.	Rec. Alc.	Rec. Non.
Random Forest	0.875	0.8718	0.8780	0.8947	0.8571	0.8500	0.9000
Naive Bayes	0.875	0.8889	0.8571	0.8000	1.0000	1.0000	0.7500
Linear SVM	0.850	0.8500	0.8500	0.8500	0.8500	0.8500	0.8500
Logistic Regression	0.825	0.8372	0.8108	0.7826	0.8824	0.9000	0.7500

Table 1: Detailed performance metrics for the evaluated classifiers on the holdout set.

4.2 Visual Analysis

The following figure illustrates the diagnostic performance of the best-performing models.

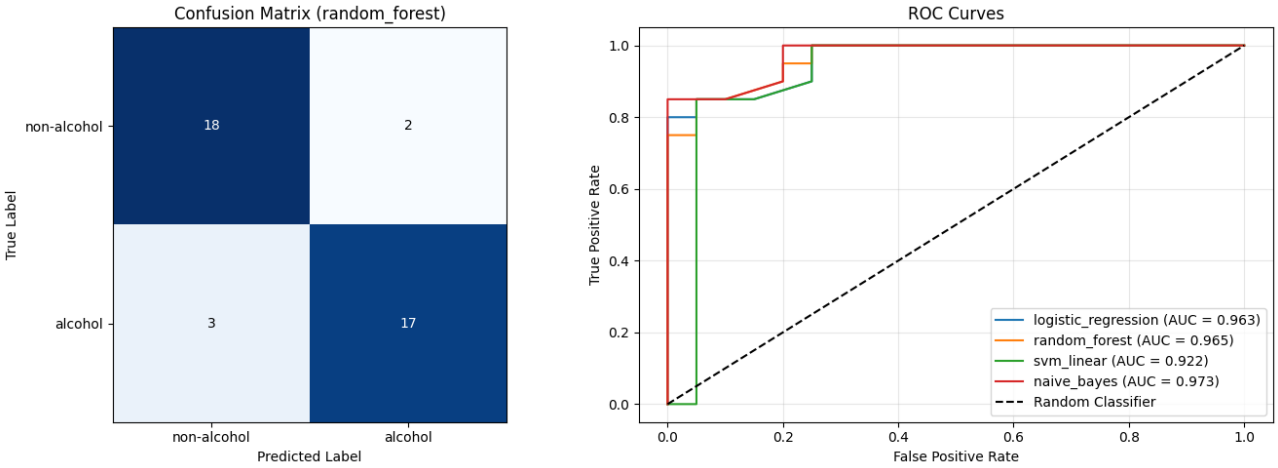

Figure 1: Confusion matrix (Left) for the best model and ROC Curves (Right) comparing all evaluated models.

4.3 Cross-Validation

5-fold Stratified Cross-Validation results highlight the stability of the Logistic Regression model across different data partitions.

Model	Mean CV Accuracy ($\pm\sigma$)
Logistic Regression	0.905 ± 0.024
Linear SVM	0.895 ± 0.040
Random Forest	0.885 ± 0.020
Naive Bayes	0.870 ± 0.019

Table 2: Cross-validation accuracy comparison.

5 Conclusion

This study demonstrates that the combination of topological graph features and sequence-based SELFIES tokens provides a highly discriminative feature space for functional group classification. The models achieved high precision and recall, with Naive Bayes providing perfect recall for alcohols in the holdout set.