Classification of Green Fluorescent Protein Brightness

Introduction

The goal of this project was to classify Green Fluorescent Protein (GFP) mutants into high brightness (class 1) and low brightness (class 0) categories. This binary classification task involved exploring features derived from amino acid sequences and implementing classification models using **Scikit-learn**.

Methods

Data Preprocessing

1. **Data Loading**: Training and testing datasets were fetched from Kaggle.
2. **Feature Engineering**:
    - Amino acid sequences were featurized using label encoding and amino acid properties (e.g., hydrophobicity, volume).
    - Features were extracted from multiple amino acid descriptors and mapped to numeric values.
    - Uneven sequences were padded, and the resulting feature matrix was standardized using StandardScaler.
3. **Feature Exploration**: Experiments included incorporating combinations of amino acid properties (e.g., D1, D2, VHSE1, Hydro) and their removal to evaluate their effect on classification performance.
4. **Model Evaluation**:
    - Data was split into training and validation sets (80:20 split).
    - Classification models tested included **Logistic Regression**, **Random Forest Classifier**, **Support Vector Classifier (SVC)**, and **K-Nearest Neighbors (KNN)**.
    - Metrics used for evaluation included **precision**, **recall**, **F1-score**, and **accuracy**.

Models Trained

1. **Logistic Regression** (Best performer): Achieved an F1-score of 0.87 on validation and 0.88073 on the Kaggle leaderboard.
2. **Random Forest Classifier**: Provided competitive results with an F1-score of 0.85 but slightly underperformed compared to Logistic Regression.
3. **SVC**: Showed good precision but lower recall compared to Logistic Regression.
4. **KNN**: Had the lowest accuracy and F1 scores, suggesting it was unsuitable for this dataset.

Results and Observations

Feature Engineering Insights

1. Including amino acid properties (e.g., Hydro, Vol, D1, D2, etc.) enhanced the model's performance, indicating the importance of biochemical features in predicting brightness.
2. Adding redundant or irrelevant descriptors (e.g., D3 through D10) negatively impacted model performance, reducing both precision and recall for all models.
3. Incorporating amino acid names did not affect the model's performance.

Key Findings

1. **Logistic Regression** consistently outperformed other models due to its ability to handle high-dimensional, structured datasets effectively.
2. Features related to hydrophobicity and volume were strong predictors of brightness levels in GFP.
3. Overfitting was mitigated by using standardized features and selecting meaningful descriptors.

Conclusion

Logistic Regression emerged as the best-performing model, achieving a Kaggle leaderboard score of **0.88073** and an F1 score of **0.87** on the validation dataset. Amino acid properties significantly influenced model performance, underscoring the importance of domain knowledge in feature engineering.