GalaxyMorph:

Automated Galaxy Morphology Classification Using Machine Learning

1. Introduction

1.1 Project Overview

This project aims to develop a machine learning-based system that classifies galaxies into morphological subclasses using photometric and structural features. Leveraging the Sloan Digital Sky Survey (SDSS) dataset, the model provides accurate predictions of galaxy types, assisting astronomers and researchers in large-scale galaxy classification tasks.

1.2 Objectives

- Collect and preprocess the SDSS galaxy dataset.
- Explore and visualize photometric and structural features.
- Train and optimize a Random Forest classifier for morphology prediction.
- Evaluate performance using accuracy, confusion matrix, and feature importance.
- Build a Flask-based web application for interactive predictions.

2. Project Initialization and Planning Phase

2.1 Define Problem Statement

Manual classification of galaxies is time-consuming, subjective, and infeasible for large surveys containing millions of objects. An automated system is required to classify galaxies based on measurable features, reducing human error and accelerating research.

2.2 Project Proposal (Proposed Solution)

We propose a supervised learning pipeline where galaxy features (magnitudes, fluxes, radii, etc.) are used to train a Random Forest model. The final system allows users to input feature values through a web interface and receive real-time subclass predictions.

2.3 Initial Project Planning

- **Dataset**: SDSS galaxy morphology catalog.
- **Tech Stack**: Python, Scikit-learn, Pandas, Flask, Matplotlib, Seaborn.
- **Phases**: Data exploration, feature engineering, model development, evaluation, integration.

3. Data Collection and Preprocessing Phase

3.1 Dataset Source

- **Source: Sloan Digital Sky Survey (SDSS).**
- **Classes: Galaxy morphological subclasses (e.g., STARFORMING, AGN, BROADLINE, etc.).**
- **Features: Magnitudes (u, g, r, i, z), fluxes, radii, PSF magnitudes, axis ratios.**

3.2 Data Quality Report

- **Shape: 100000 rows × 43 features.**
- **Missing Data: Checked; negligible missingness.**
- **Balance: Subclass distribution slightly imbalanced.**

3.3 Data Exploration and Preprocessing

- Distribution analysis of magnitudes, fluxes, and radii.
- Correlation heatmap between features.
- Pairplot of photometric features across subclasses.
- Label encoding for target classes.
- Standardization of input features.

4. Model Development Phase

4.1 Model Selection

- Random Forest Classifier chosen for its interpretability and robust performance.
- Hyperparameters: n_estimators=200, random_state=42.

4.2 Training & Validation

- Train-test split: 80/20 stratified.
- Features scaled using StandardScaler.
- Achieved accuracy: **87.8% on test data**.

4.3 Evaluation Metrics

- **Classification Report**: Precision, recall, F1-score for each subclass.
- **Confusion Matrix**: Visualized to assess misclassifications.
- **Feature Importance**: Identified top predictors such as psfMag_u, petroR50_g, u, z.

5. Integration and Web App Development

5.1 Flask Web App

- A lightweight Flask application enables real-time predictions.
- Users input galaxy features manually.
- The trained model processes the input and outputs the predicted subclass.

5.2 HTML Interface

- home.html: Input form with all feature fields.
- result.html: Displays predicted galaxy class.

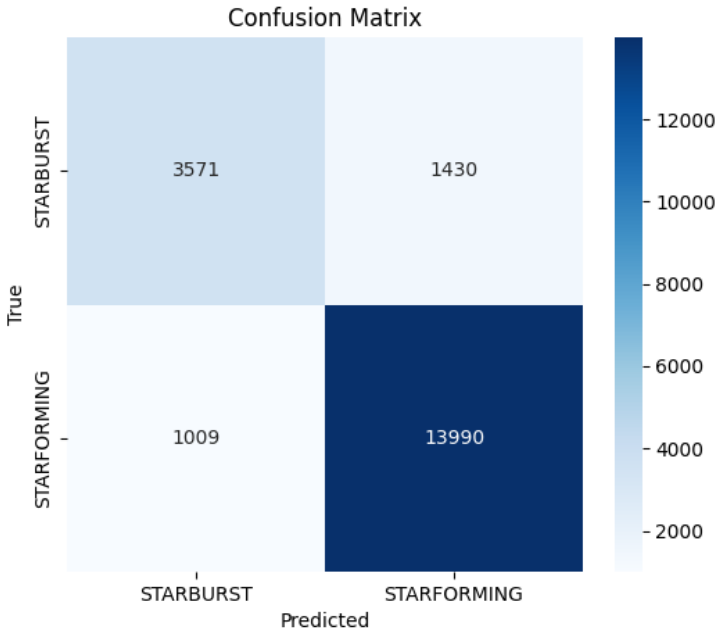
6. Testing

6.1 Model Prediction

- Tested on unseen 20% test set.
- Predictions matched actual subclasses with **~88% accuracy**.

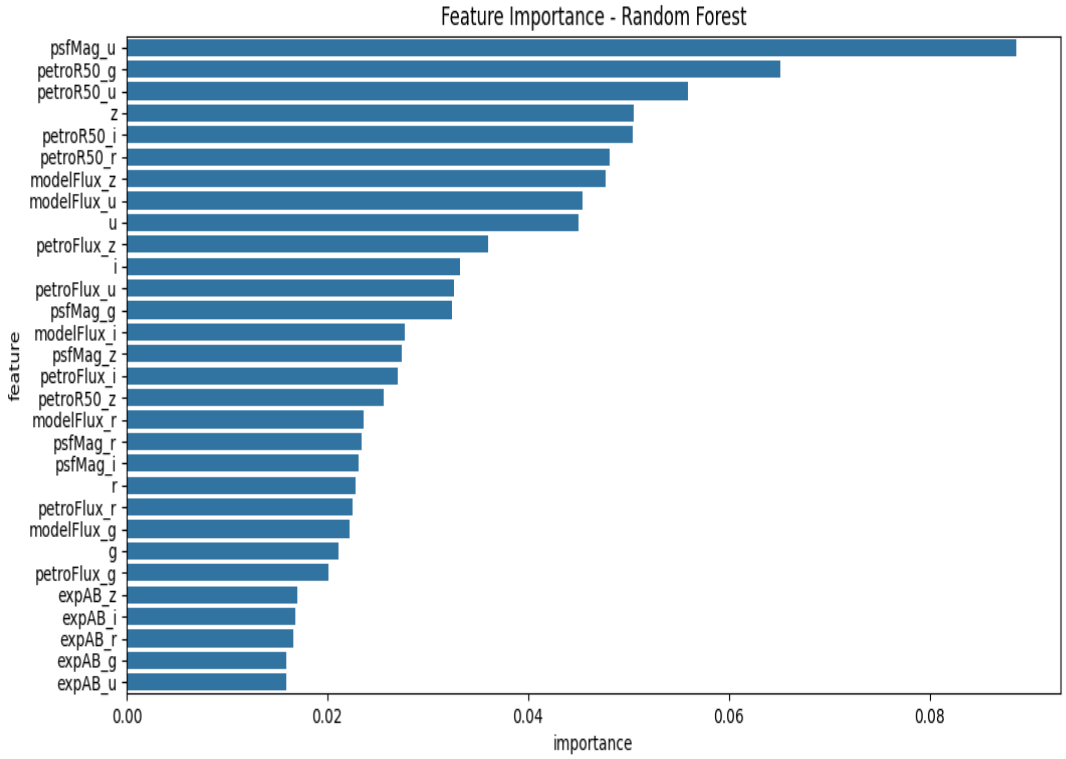
6.2 Confusion Matrix

- Correct classifications dominate the diagonal.
- Minor confusion observed between subclasses with similar feature ranges.



6.3 Feature Importance

- Top 10 predictors included photometric magnitudes and radii, highlighting their role in morphology classification.



6.4 Classification Report

- A detailed classification report was generated including:
    - **Precision**
    - **Recall**
    - **F1-score**
    - **Support**

Class	Precision	Recall	F1-Score	Support
STARBURST	0.78	0.71	0.75	5,001
STARFORMING	0.91	0.93	0.92	14,999
Accuracy			**0.88**	20,000
Macro Avg	0.84	0.82	0.83	20,000
Weighted Avg	0.88	0.88	0.88	20,000

7. Advantages & Disadvantages

Advantages:

- Automated and scalable for large datasets.
- Good accuracy with simple preprocessing.
- Feature importance provides interpretability.
- Web app is lightweight and user-friendly.

Disadvantages:

- Model limited to available subclasses in dataset.
- Requires manual feature input in web app.
- Prediction accuracy depends on quality of SDSS features.

8. Conclusion

GalaxyMorph successfully demonstrates automated galaxy classification using Random Forests. With ~88% accuracy and interpretability, this model provides a practical baseline for astroinformatics applications. The Flask app integration ensures accessibility for researchers and students.

9. Future Scope

- Explore deep learning models (e.g., CNNs) with imaging data.
- Improve handling of class imbalance.
- Deploy the web app on cloud (Heroku, AWS, GCP).
- Add visualization of input-output predictions in the UI.

10. Appendix

10.1 Source Code

- **Notebooks**: Data exploration, preprocessing, model training.
- **Scripts**: train_model.py, app.py.

10.2 GitHub / Demo Link

https://github.com/balapraharsha/GalaxyMorph