LEIHASREE A
Panimalar Engineering College

Machine Learning Engineer Virtual Internship-Project Documentation

Project Title: SDSS Galaxy Classification using Machine Learning

Introduction

Project Overview

The rapid increase in astronomical data has created a need for automated galaxy classification systems. This project uses machine learning techniques to classify galaxies based on features from the Sloan Digital Sky Survey (SDSS) dataset. By analyzing important galaxy attributes, the model predicts galaxy subclasses efficiently and accurately, reducing manual effort and supporting better understanding of galaxy behavior and evolution.

Problem Statement

Manual classification of galaxies from SDSS data is time-consuming and inefficient. This project uses machine learning to automatically classify galaxies into subclasses.

Project Objectives

To analyze and preprocess SDSS galaxy data for classification.

To build and compare multiple machine learning models.

To identify the best-performing model for galaxy subclass prediction.

To deploy the selected model using a simple web application.

Dataset Description

Dataset Name: SDSS Galaxy Dataset

Source: Sloan Digital Sky Survey (SDSS)

Target Variable: subclass

Classes: Starforming Galaxy and Starbursting Galaxy

Features Used (Top 10 Selected Features):
The following ten features were selected using the SelectKBest feature selection technique to improve model performance and reduce dimensionality:

specobjid,modelFlux_i,modelFlux_z,petroRad_u,petroRad_g,petroRad_i,petroRad_r,petroRad_z,petroR50_z,redshift

Project Milestones

1.Dataset Exploration & Understanding (EDA)

2.Data Preprocessing

3.Feature Selection

4.Model Training

5.Model Evaluation

6.Best Model Selection

7.Model Deployment

8.Testing & Validation

Dataset Exploration & Understanding(EDA)

Objective:
Understand the dataset structure, features, and target variable.

Key Activities:
1.Load the SDSS Galaxy Dataset
2.Inspect data types and check for missing values
3.Analyze basic statistics of features
4.Visualize distributions and patterns

Outcome:
1.Clear understanding of data quality and structure
2.Insights into feature distributions and target balance

Data Preprocessing

Objective:
Prepare the dataset for modeling by handling inconsistencies and scaling features.

Key Activities:
1.Handle missing or duplicate values
2.Encode categorical variables (if any)
3.Apply feature scaling (StandardScaler)
4.Balance classes using techniques like SMOTE

Outcome:
1.Cleaned, scaled, and balanced dataset ready for training
2.Reduced chances of bias due to class imbalance

Feature Selection

Objective:
Select the most relevant features from the dataset that contribute significantly to predicting the galaxy subclass.

Key Activities:
1.Identified highly correlated features.
2.Dropped irrelevant or redundant features.
3.Selected the top 10 features for model training.

Outcome:
A reduced and optimized feature set that improves model efficiency and performance while maintaining predictive accuracy.

Model Training

Objective:
Train multiple machine learning models using the selected features to classify galaxies into starforming or starbursting.

Key Activities:
1.Split dataset into training and testing sets.
2.Applied scaling using StandardScaler.
3.Trained models using Decision Tree, Logistic Regression, and Random Forest algorithms.

Outcome:
Trained models ready for evaluation, each with learned patterns to classify galaxies.

Model Evaluation

Objective:
Assess the performance of the trained models to identify the most accurate and reliable one.

Key Activities:
1.Calculated metrics such as accuracy, precision, recall, and F1-score.
2.Generated confusion matrices for each model.
3.Compared training and testing performance to check for overfitting.

Outcome:
Performance insights for each model, highlighting strengths and weaknesses to guide model selection.

Best Model Selection

Objective:
Select the model with the highest performance for deployment.

Key Activities:
1.Compared evaluation metrics of Decision Tree, Logistic Regression, and Random Forest models.
2.Chose the model with the best balance of accuracy, precision, and recall on test data.

Outcome:
Random Forest identified as the best-performing model, ready for deployment.

Model Deployment

Objective:
Deploy the selected model so it can make predictions on new data via a web interface.

Key Activities:
1.Saved the trained Random Forest model using pickle.
2.Built a Flask web application with routes for home, input, and output pages.
3.Connected the web app to the trained model to accept user input and display predictions.

Outcome:
A functional web application capable of classifying galaxies into “starforming” or “starbursting” based on input features.

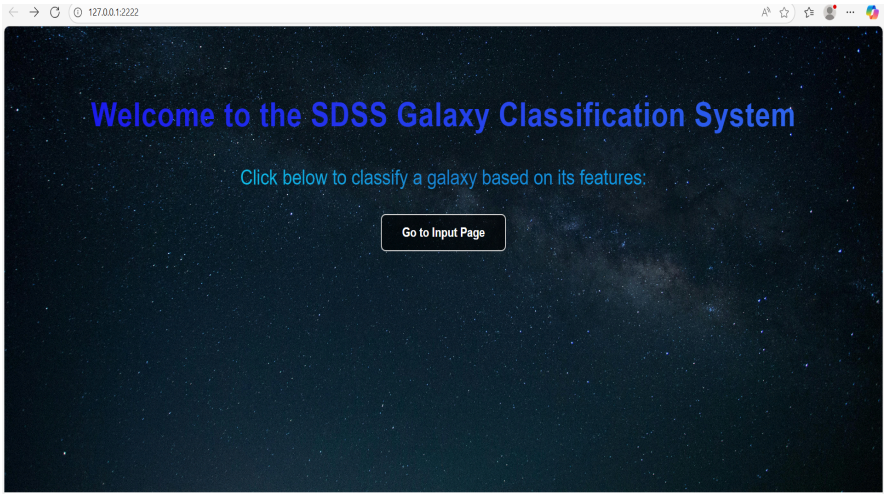
Testing & Validation

Objective:
Validate the deployed model to ensure its predictions are accurate and consistent with expectations.

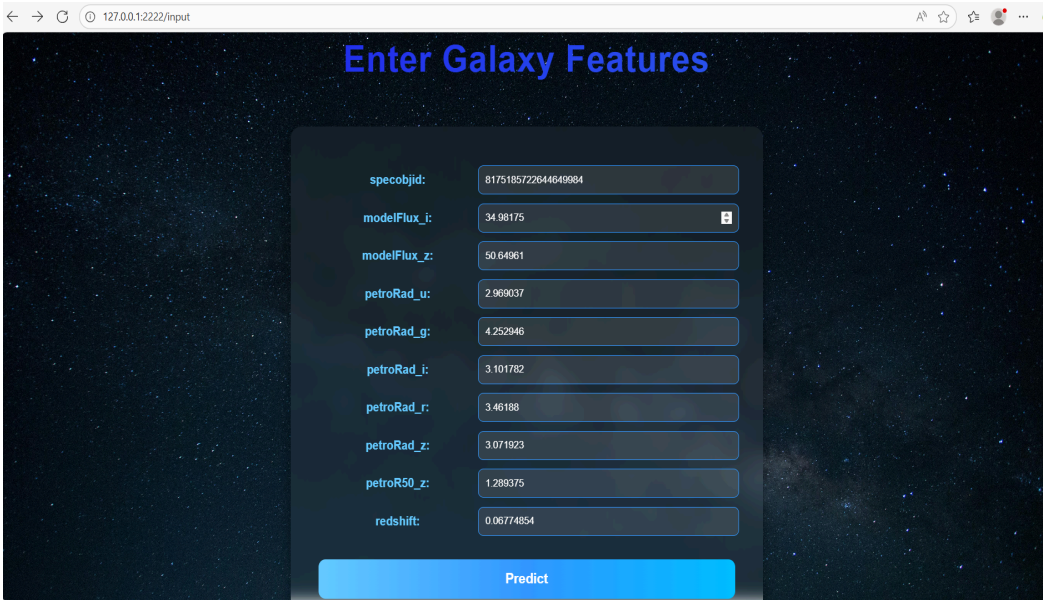
Key Activities:
1.Tested the model on sample inputs through the web application.
2.Verified prediction outputs against known data.
3.Checked for usability, responsiveness, and correctness of the web interface.

Outcome:
The deployed application provides reliable predictions and correctly classifies galaxy types based on user-provided features.

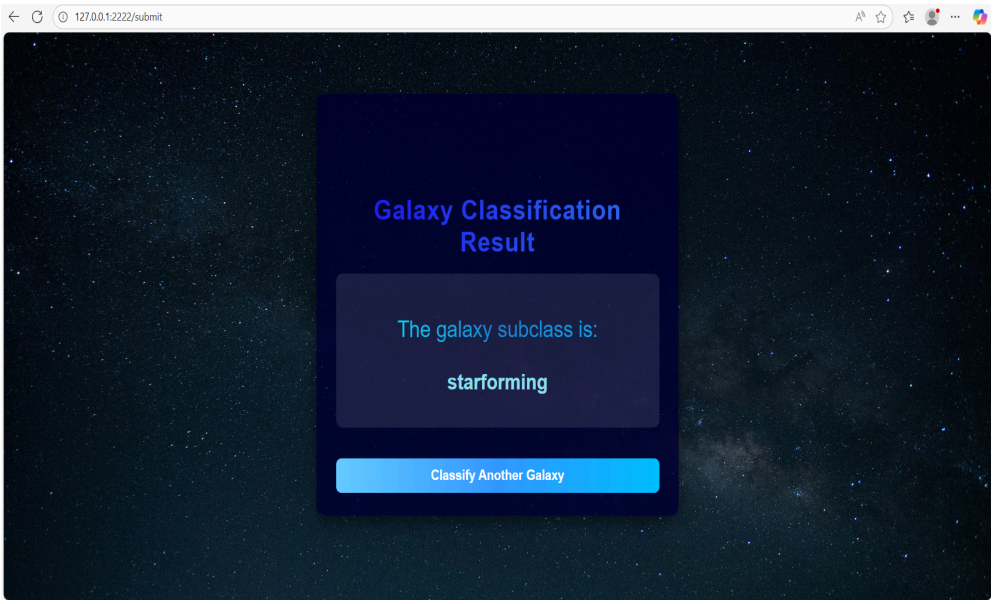
System Implementation Results



HOME PAGE



INPUT PAGE



OUTPUT PAGE

Results and Performance analysis

The Random Forest model was used to classify galaxies into starforming and starbursting classes using the SDSS dataset. The model showed good accuracy and consistent prediction results based on the selected features.

The web application successfully integrates the trained model, allowing users to input galaxy features and receive reliable classification results in real time.

Overall, the system demonstrates effective performance and practical usability.

Project Resources

Project Demonstration Video:
https://drive.google.com/file/d/1Mo8okcyk7keIKJ0f7CgQW-LhHxQk3pYq/view?usp=sharing

GitHub Repository Link:
https://github.com/leihasree-a/Galaxy_Classification_Project

Conclusion

A machine learning–based galaxy classification system was successfuly developed using SDSS data. The Random Forest model delivers reliable predictions and is deployed through a simple web interface, making the solution practical, scalable, and easy to use.