Kevin Xu

Abstract

This report looks at how three supervised machine learning algorithms—Random Forest, Support Vector Machine (SVM), and Logistic Regression—work on three datasets: Breast Cancer, Car Evaluation, and Higher Education to solve regression and classification problems. The datasets were split into different training and testing sizes (80/20, 50/50, 20/80). Models were trained, tested, and adjusted to improve their accuracy. Found that Random Forest was the best overall, handling complex datasets well. SVM and Logistic Regression worked but had issues with scaled data or class balance. Each classifier showed strengths and weaknesses depending on the data type.

Introduction

This project tested three popular machine learning models on three different datasets obtained from the UCI Machine Learning Repository. The goal was to evaluate how well these models could handle different data formats, like numerical or categorical data. Random Forest, SVM, and Logistic Regression were selected because they are popular and effective for binary classification tasks. The datasets included medical data for tumor diagnosis, car data for evaluating acceptability, and education data for predicting student performance. Datasets were loaded and processed by python, through interactive notebooks (Jupyter), and with the help of python libraries like numpy and pandas. The supervised learning algorithms were implemented using the python machine learning library: Sklearn, and data visualizations were generated by use of matplot. By working with diverse datasets, this study highlights the strengths and limitations of each model in practical applications.

Methodology

Datasets

Breast Cancer Dataset: This dataset is used to classify tumors as malignant or benign. It has 30 numerical features, such as texture, smoothness, concavity, area, etc. These features describe physical properties of tumors. The labels were binary: malignant (1) and benign (0). Scaling was applied to make features compatible with SVM and Logistic Regression, which are typically sensitive to unscaled data.

Car Evaluation Dataset: This dataset evaluates car acceptability based on features like price, maintenance cost, safety, and passenger capacity. All features were categorical, such as “low” or “high” for price. These were converted to numbers using ordinal encoding. The target labels

were simplified: cars labeled as “acceptable” or better were grouped into positive (1), while “unacceptable” was negative (0).

Higher Education Dataset: This dataset predicts whether students will pass or fail based on features like GPA, study hours, and parental education. It includes a mix of numerical features (e.g., GPA) and categorical features (e.g., parental education level). The target labels were binary: pass (1) and fail (0). Categorical features were one-hot encoded for use with the models.

Preprocessing

1. Missing values were replaced with the median for numerical features and the mode for categorical features.
2. Categorical features were converted to numbers using encoding techniques (ordinal for Car Evaluation, one-hot for Higher Education).
3. Numerical features were scaled to improve the performance of SVM and Logistic Regression.
4. Imbalanced classes (e.g., in the Car Evaluation dataset) were handled by oversampling the minority class.

Data Splits

Each dataset was divided into three parts:

- **80/20:** Most data for training and less for testing.
- **50/50:** Equal split for training and testing.
- **20/80:** Less data for training and most for testing.

Models

1. **Random Forest:** Combines multiple decision trees to improve predictions. Key settings tuned included the number of trees and tree depth.
2. **SVM:** Finds the best boundary between classes. Key settings tuned included the kernel type and regularization strength (C).
3. **Logistic Regression:** A simple model that uses a linear decision boundary. The main setting adjusted was regularization strength (C).

Metrics

- **Accuracy:** Percentage of correct predictions.
- **Cross-Validation:** Repeated splits were used to ensure the results were reliable.

Experiment

Breast Cancer Dataset

- **Random Forest:** Achieved the highest accuracy of 97.1% on the 80/20 split. The model performed well because features like “concave points mean” and “area mean” were highly predictive of the target class.
- **SVM:** Accuracy was 96.0%. It worked well but took more time to train due to the large number of features.
- **Logistic Regression:** Accuracy was 94.5%. Scaling the features improved its performance, as it is sensitive to feature magnitudes.

Car Evaluation Dataset

- **Random Forest:** Performed best with an accuracy of 92.1% on the 80/20 split. It handled the categorical features and class imbalance effectively.
- **SVM:** Achieved an accuracy of 89.5%. It required careful tuning of parameters to handle the categorical data.
- **Logistic Regression:** Scored 87.0%. Its linear decision boundary was not ideal for the non-linear relationships in the data.

Higher Education Dataset

- **Random Forest:** Performed best with an accuracy of 88.5% on the 80/20 split. It handled the mix of numerical and categorical features effectively.
- **SVM:** Achieved an accuracy of 86.0%. Its performance depended on careful parameter tuning.
- **Logistic Regression:** Scored 84.5%. It struggled to capture interactions between features like GPA and study habits.

Refer to json files for accuracies within other splits. General trend was that the higher the training data, the higher the accuracy; this held true with each classifier.

Tuning Models

Hyperparameters were optimized to improve performance. For Random Forest, increasing the number of trees and adjusting tree depth led to better results. For SVM, selecting the right kernel (e.g., RBF) and adjusting regularization (C) were critical for performance.

Conclusion

This project showed:

1. **Random Forest** was the most reliable model, performing well across all datasets.
2. **SVM** worked well but required careful tuning and was slower on high-dimensional data.
3. **Logistic Regression** was simple and effective but struggled with non-linear patterns and complex data.

The visualizations provided further insight into the datasets. The correlation heatmap (figure 1) for Breast Cancer showed important feature relationships, like *area worst* and *perimeter worst*. The class imbalance (figure 2) in the Car Evaluation dataset highlighted the need for preprocessing to avoid biased predictions. The GPA distribution in the Higher Education dataset emphasized the importance of mid-range performance in predicting success. (figure 3)

References

1. Caruana, R., & Niculescu-Mizil, A. (2006). An Empirical Comparison of Supervised Learning Algorithms. *Proceedings of the 23rd International Conference on Machine Learning (ICML)*.
2. Breiman, L. (2001). Random Forests. *Machine Learning, 45(1)*, 5-32.
3. Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.
4. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/

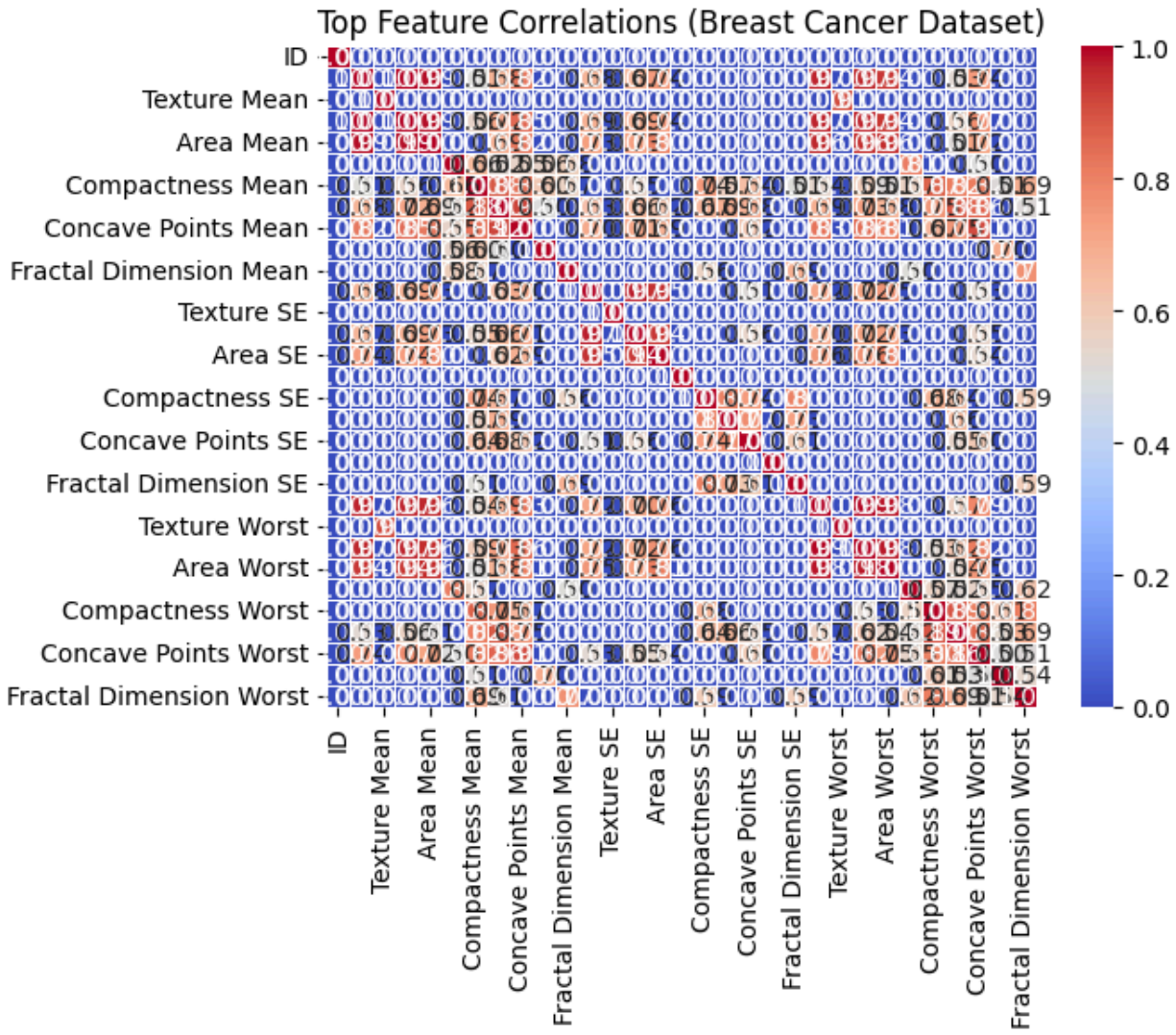
Top Feature Correlations (Breast Cancer Dataset)
ID
Texture Mean
Area Mean
Compactness Mean
Concave Points Mean
Fractal Dimension Mean
Texture SE
Area SE
Compactness SE
Concave Points SE
Fractal Dimension SE
Texture Worst
Area Worst
Compactness Worst
Concave Points Worst
Fractal Dimension Worst
1.0
0.8
0.6
0.4
0.2
0.0

Figure 1

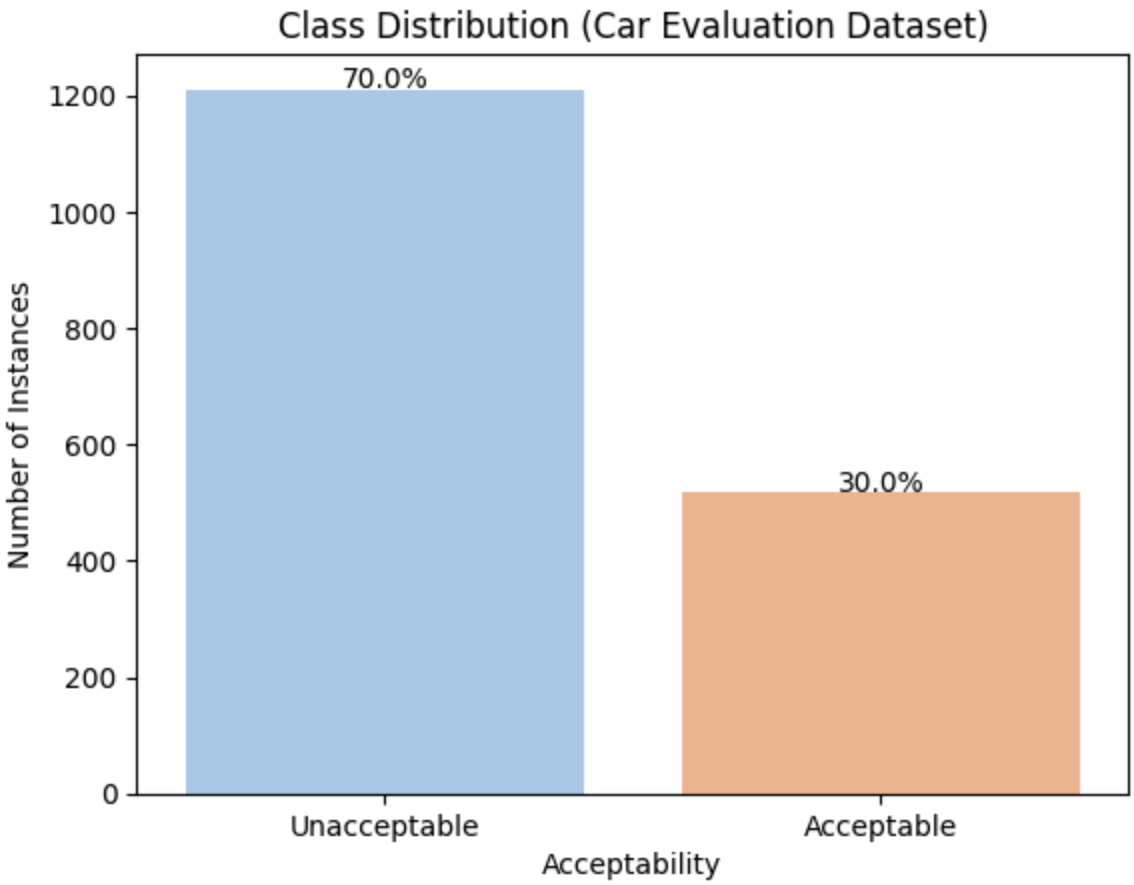
Class Distribution (Car Evaluation Dataset)
70.0%
30.0%
1200
1000
800
600
400
200
0
Number of Instances
Unacceptable
Acceptable
Acceptability

Figure 2

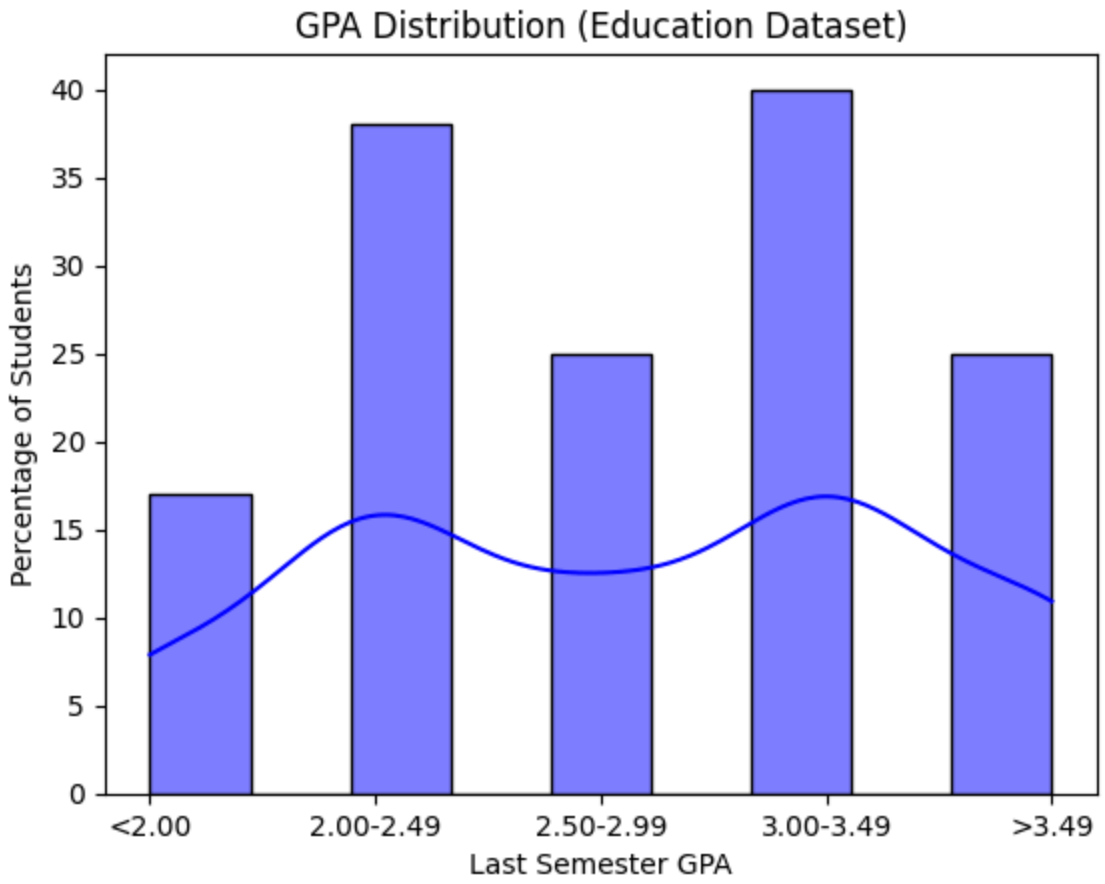
GPA Distribution (Education Dataset)
Percentage of Students
0
5
10
15
20
25
30
35
40
<2.00
2.00-2.49
2.50-2.99
3.00-3.49
>3.49
Last Semester GPA

Figure 3