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

The Breast Cancer dataset is numerical, with high feature correlations, such as between *concave points mean* and *area mean*. These correlations played a crucial role in model performance.

1. **Random Forest:**
    - Achieved an **accuracy of 97.1%** on the 80/20 split, leveraging its ability to handle redundant and correlated features.
    - Accuracy decreased to **96.1% (50/50)** and **94.7% (20/80)** as training data reduced, reflecting its reliance on sufficient data to learn complex patterns.
    - The hyperparameters (n_estimators: 200, max_depth: None, min_samples_split: 5) suggest a deep forest with high flexibility, ideal for this dataset.
2. **SVM:**
    - Outperformed Random Forest slightly, achieving **98.2% accuracy** on the 80/20 split with a **linear kernel** (c=0.1).
    - Accuracy remained consistent at **97.8%** across 50/50 and 20/80 splits, indicating robustness even with reduced training data.
    - The linear kernel effectively captured decision boundaries, while the regularization parameter (c=0.1) balanced bias and variance.
3. **Logistic Regression:**
    - Performed well, achieving **97.3% (80/20)** and peaking at **98.6% (50/50)**. However, accuracy dropped slightly to **97.6% (20/80)**.
    - Regularization (c=1) played a key role in preventing overfitting, and scaling features improved performance for this linear model.
    - Despite being a simpler model, Logistic Regression was competitive due to the strong linear separability of the dataset.

Overall: SVM was the best-performing model for Breast Cancer, excelling in both accuracy and robustness across splits. However, Logistic Regression demonstrated surprising strength given its simplicity.

Car Evaluation Dataset

This categorical dataset was transformed using ordinal encoding and had a significant class imbalance (70% "unacceptable" vs. 30% "acceptable"). The imbalance was addressed through oversampling.

1. **Random Forest:**
    - Performed consistently across partitions, with an accuracy of **91.0% (80/20)**, **91.3% (50/50)**, and **88.1% (20/80)**.
    - Hyperparameters (n_estimators: 100, max_depth: 10) indicate that a moderately deep forest with fewer trees was sufficient to capture categorical relationships.
    - Its ability to handle imbalances and non-linear patterns made it highly effective.
2. **SVM:**

- Achieved the highest accuracy of **96.8% (80/20)** using an **rbf kernel** (c= 10, gamma = auto).
- Accuracy remained high at **96.5% (50/50)** but dropped to **91.6% (20/80)** as training data decreased.
- The kernel effectively mapped the categorical features into higher dimensions, but the model was sensitive to data imbalance.

3. **Logistic Regression:**
    - Struggled, with accuracy hovering around **63.5% (80/20)**, **63.8% (50/50)**, and **63.5% (20/80)**.
    - The linear decision boundary was inadequate for this dataset, as relationships between features and labels were highly non-linear.

Overall: SVM emerged as the best performer for Car Evaluation due to its ability to model non-linear relationships, while Random Forest provided strong, consistent performance across splits. Logistic Regression was unsuitable for this dataset due to its linear nature.

Higher Education Dataset

This dataset contained mixed features (numerical and categorical) and required both ordinal and one-hot encoding. Feature interactions, such as GPA and parental education, influenced model performance.

1. **Random Forest:**
    - Achieved the highest accuracy of **79.3% (80/20)** but dropped to **71.2% (50/50)** and **60.3% (20/80)** as training data decreased.
    - Hyperparameters (n_estimators: 200, max_depth: None) show the need for a robust forest to capture mixed feature interactions.
    - Its versatility in handling both categorical and numerical data was key.
2. **SVM:**
    - Underperformed, with an accuracy of **55.2% (80/20)**, peaking at **65.7% (50/50)** but dropping again to **53.4% (20/80)**.
    - The linear kernel (C = 1, gamma = scale) struggled to capture complex feature relationships. Even with RBF kernels, accuracy remained limited due to the mixed nature of the dataset.
3. **Logistic Regression:**
    - Achieved moderate performance with **65.5% (80/20)**, **64.3% (50/50)**, and **66.4% (20/80)**.
    - Regularization (C=0.01) helped manage feature interactions, but due to its linear nature, it was hard to fully capture the dataset's complexity.

Overall: Random Forest was the best choice for Higher Education due to its ability to model complex feature interactions. SVM and Logistic Regression were less effective, struggling with the dataset's mixed structure.

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/*

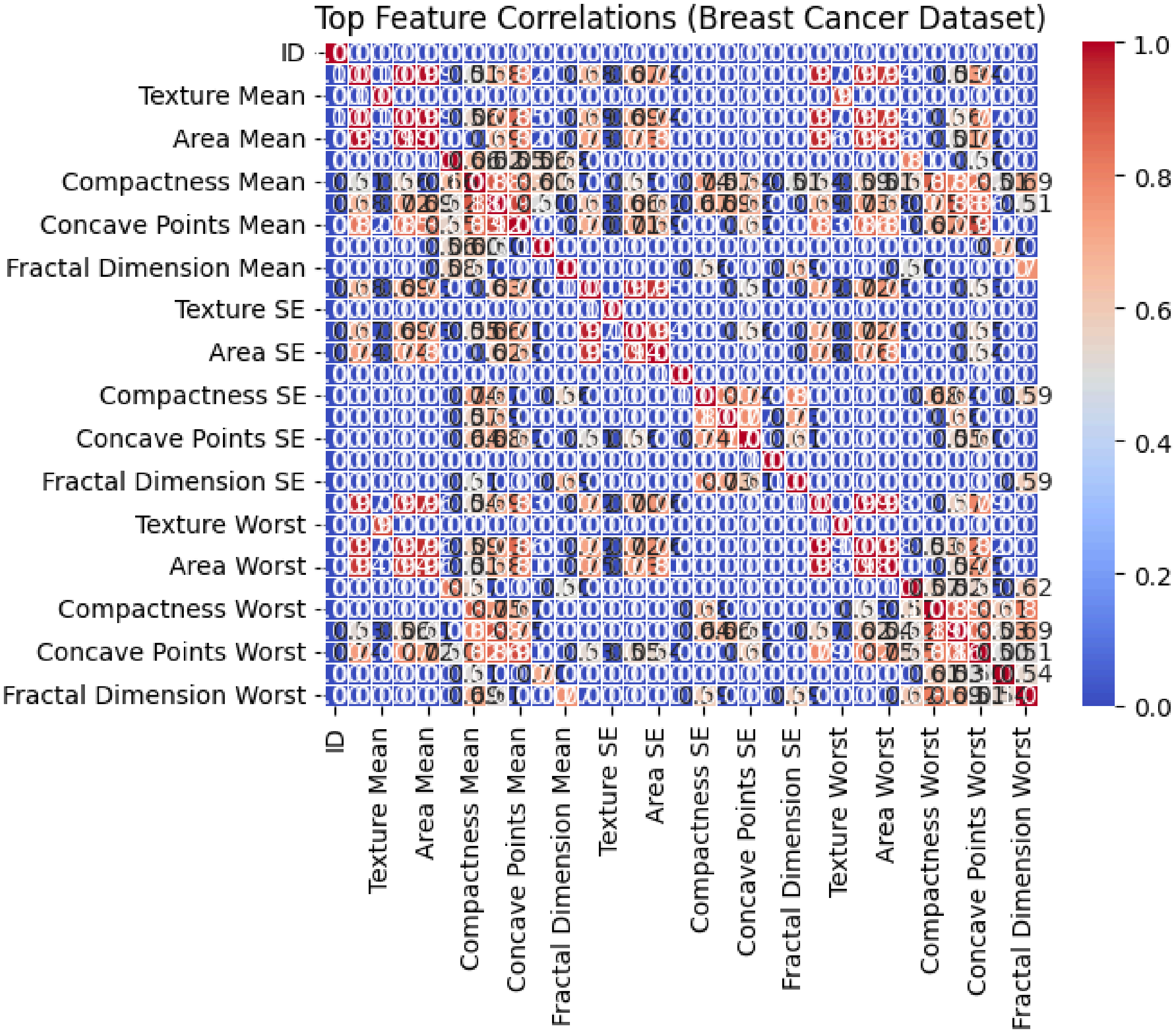
Top Feature Correlations (Breast Cancer Dataset)
ID
Texture Mean
Area Mean
Compactness Mean
Concave Points Mean
Fractal Dimension Mean
Texture SE
Area SE
Compactness SE
Concave Points SE
Fractal Dimension SE
Texture Worst
Area Worst
Compactness Worst
Concave Points Worst
Fractal Dimension Worst
1.0
0.8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

Figure 1

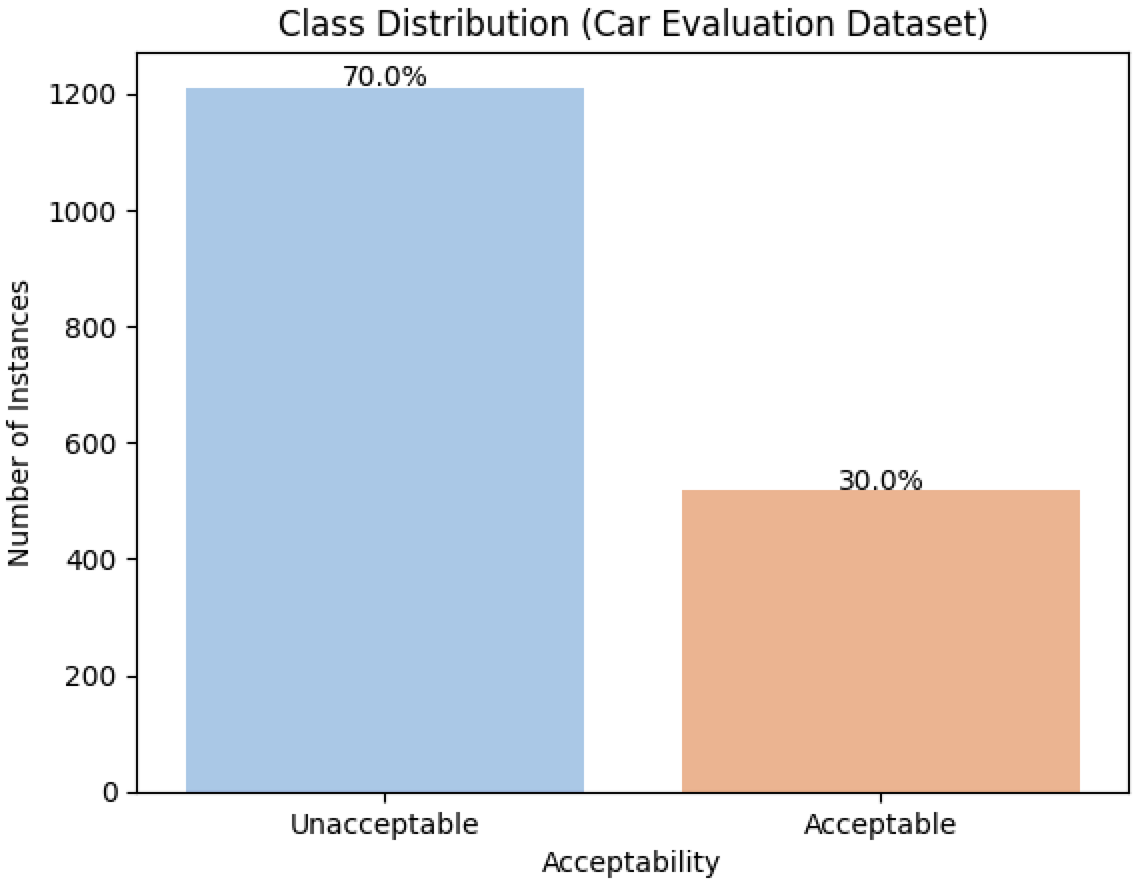
Class Distribution (Car Evaluation Dataset)
70.0%
30.0%
Number of Instances
0
200
400
600
800
1000
1200
Unacceptable
Acceptable
Acceptability

Figure 2

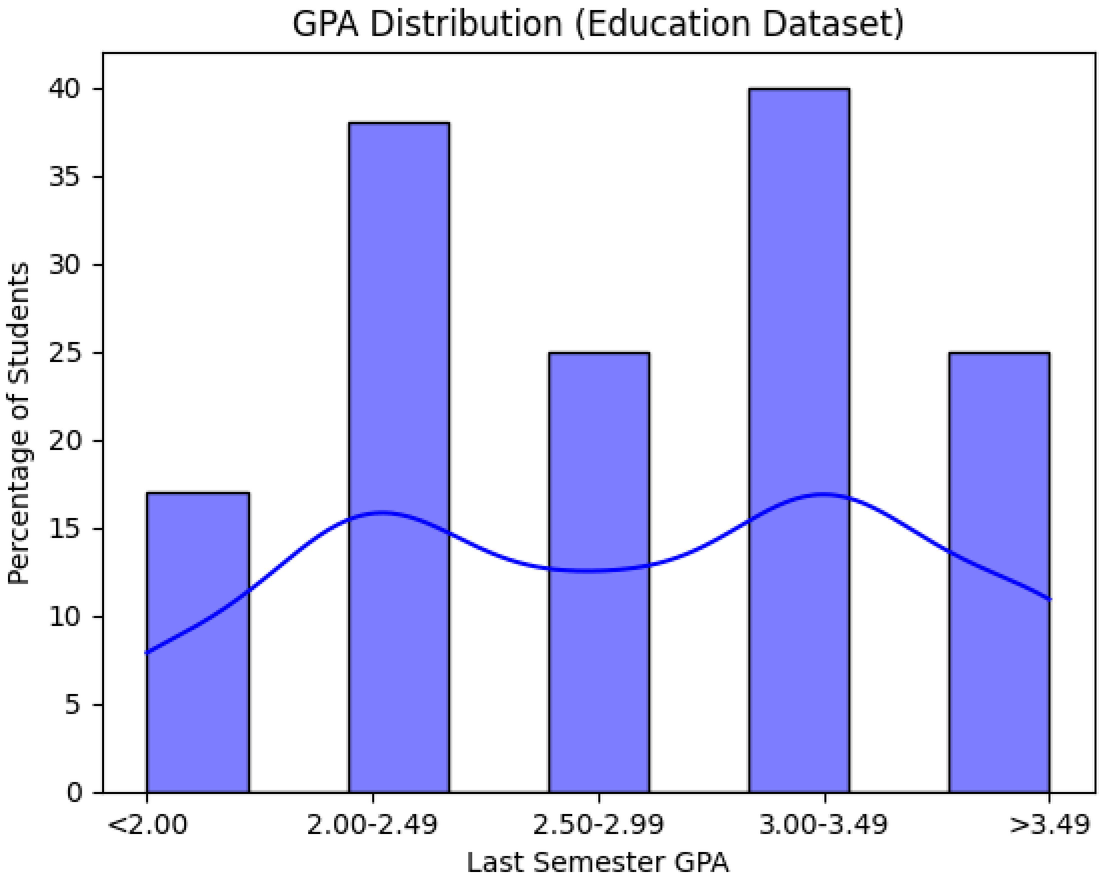
GPA Distribution (Education Dataset)
Percentage of Students
0
5
10
15
20
25
30
35
40
<2.00
2.00-2.49
2.50-2.99
3.00-3.49
>3.49
Last Semester GPA

Figure 3