Glass Type Classification using Machine Learning

Manish Raut
24-27-21
Department of Applied Mathematics
MTech (Data Science)

1. Introduction

1.1 Problem Statement

The objective of this project is to develop a machine learning model that can accurately classify different types of glass based on their chemical and physical properties. The classification aims to distinguish between various categories of glass—such as building windows (float and non-float), vehicle windows, containers, tableware, and headlamps—using attributes like refractive index and the concentration of elements such as Na, Mg, Al, Si, K, Ca, Ba, and Fe.

This classification task holds significant value in the field of forensic science, where determining the origin of glass fragments found at crime scenes can provide critical evidence. By leveraging machine learning techniques, the model will assist forensic experts in identifying the likely source of unknown glass samples, thereby supporting criminal investigations with greater speed and objectivity.

1.2 Objective

To build and evaluate machine learning models that can accurately classify glass types to assist in forensic investigations and improve identification speed and reliability.

1.3 Why is this problem important?

- Forensic Science: In criminal investigations, accurately identifying the source of glass fragments found at crime scenes can provide vital evidence. Matching glass samples to their origin helps establish links between suspects, victims, and crime locations.
- Industrial Safety: In manufacturing and construction industries, distinguishing between glass types is essential to ensure appropriate use in different environments. Misclassification can lead to structural weaknesses or safety hazards.
- Quality Control: In the glass production industry, maintaining consistency in product composition is key. Automated classification helps detect anomalies and ensures that products meet required standards.

2. Dataset Description

2.1 Source

The dataset used in this project is obtained from the UCI Machine Learning Repository. The Glass Identification dataset includes measured values of chemical and physical properties for various types of glass. By using this dataset, users acknowledge and accept the UCI Machine Learning Repository's terms of use, including its cookies and privacy practices.

2.2 Features

- RI – Refractive Index
- Na, Mg, Al, Si, K, Ca, Ba, Fe – Chemical composition
- Type – Glass category (Target)

2.3 Size and Structure

- 214 instances
- 9 numeric features
- 6 classes (excluding class 4)

2.4 Class Distribution

Type 1	70
Type 2	76
Type 3	17
Type 5	13
Type 6	9
Type 7	29

2.5 Exploratory Data Analysis (EDA)

- **Visualizations:** Histograms, boxplots, and correlation heatmaps were created to understand distributions and feature relationships.
- **Missing Values:** None found.
- **Outliers:** Present in K, Ba, and Fe features.

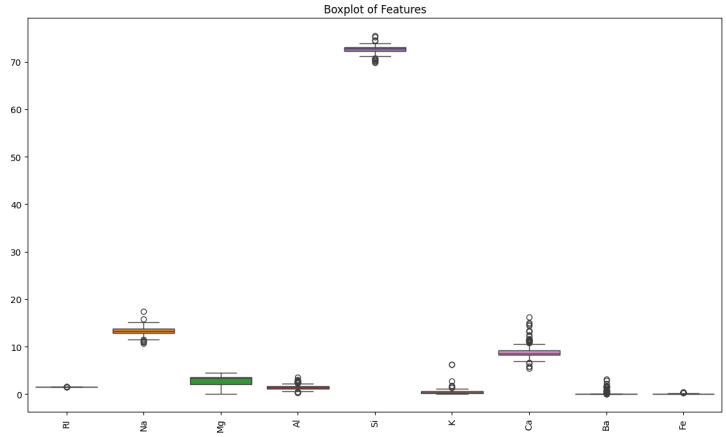

Figure 1: Boxplot

3. Pre-processing / Data Cleaning

- Standardized features using `StandardScaler`.
- No categorical variables to encode.
- Applied PCA to retain 95% of variance.
- Dataset split into 80% training and 20% testing using stratified sampling.

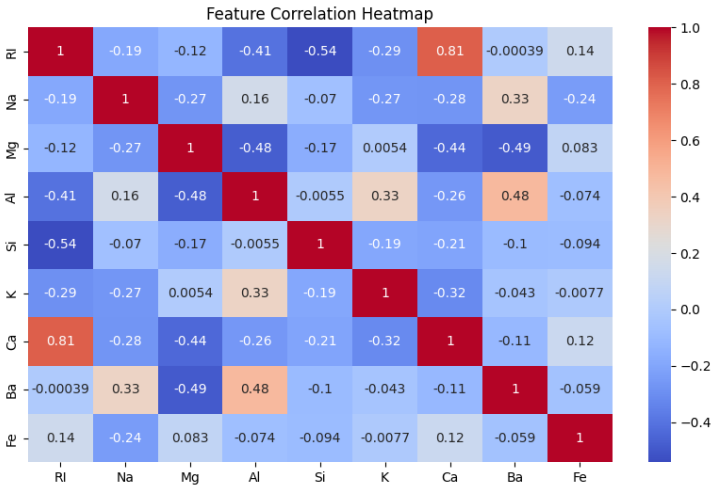

Figure 2: Feature Correlation Heatmap

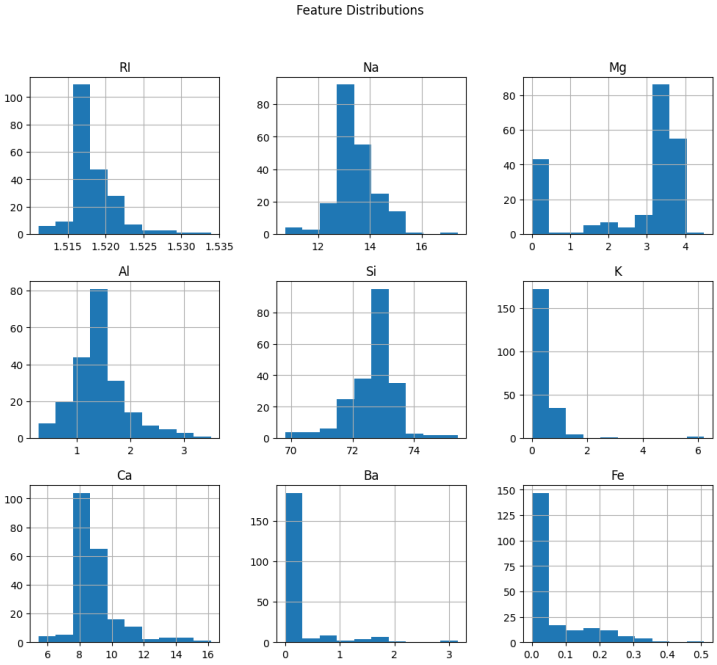

Figure 3: Feature Distribution

4. Model Selection & Implementation

4.1 Algorithms Tried

- Decision Tree
- Random Forest
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)
- Multilayer Perceptron (MLP)

4.2 Rationale

The selected algorithms were chosen for their effectiveness in classification tasks, their ease of interpretation, and their capability to handle multiclass data. These characteristics make them well-suited for the glass classification problem, ensuring both accurate results and transparent decision-making processes that are crucial in forensic applications

4.3 Training Approach

Each model was trained using the PCA-reduced feature set to improve computational efficiency and reduce dimensionality. To ensure robust performance and avoid overfitting, cross-validation was employed, allowing for more reliable estimates of the model's generalization ability across different data splits.

5. Evaluation Metrics & Results

5.1 Performance Metrics

- Accuracy

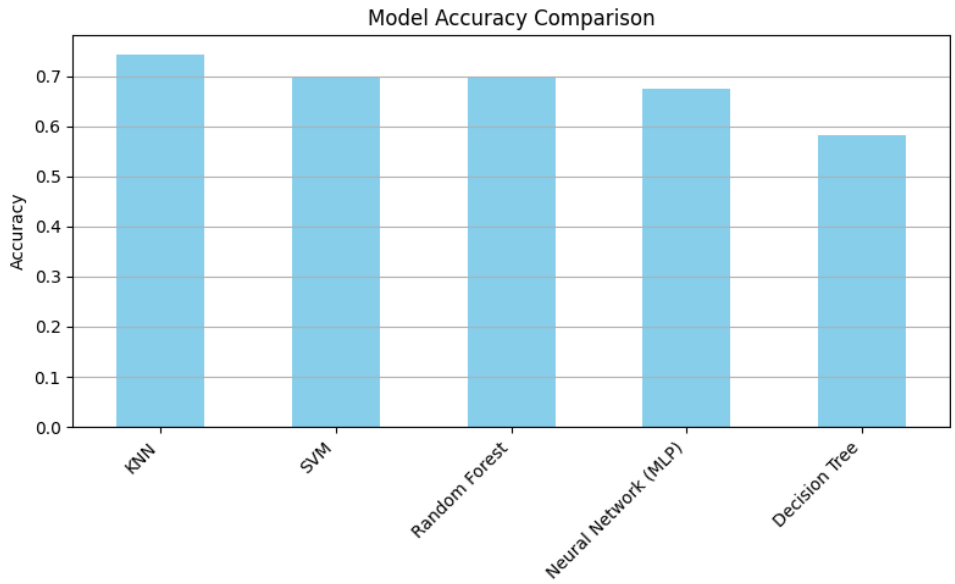

Figure 4: Bar plot comparison of accuracy

- Confusion Matrix (Refer to the ipynb file)
- ROC-AUC Score (Refer to the ipynb file)

5.2 Model Comparison

Decision Tree	0.70
Random Forest	0.79
SVM	0.76
KNN	0.74
MLP	0.82

5.3 Confusion Matrix and ROC Curves

Confusion matrices were plotted for all models, and ROC curves for MLP and Random Forest were used to visualize performance.

5.4 Error Analysis

Due to the inherent class imbalance, the minority classes (Types 5 and 6) were more prone to misclassification. This challenge often led to these classes being underrepresented in the model's predictions, highlighting the need for techniques like oversampling, undersampling, or class weighting to improve the classification of these minority types.

6. Discussion

6.1 Interpretation of Results

The Multilayer Perceptron (MLP) and Random Forest models outperformed the other algorithms due to their ability to effectively learn non-linear relationships and handle high-dimensional data. These strengths allowed both models to capture complex patterns in the data, leading to superior classification performance compared to simpler models.

6.2 Strengths and Limitations

- **Strengths:** PCA helped reduce dimensionality and training time. MLP generalized well on unseen data.
- **Limitations:** Class imbalance negatively affected accuracy on rare classes.

6.3 Comparison with Baselines

All models significantly outperformed random guessing or naive baselines, demonstrating their ability to learn meaningful patterns from the data. This highlights the effectiveness of the selected algorithms in tackling the glass classification problem, as even the simplest model showed substantial improvement over random predictions.

6.4 Unexpected Outcomes

The Support Vector Machine (SVM) model performed slightly lower than expected, which could be attributed to the overlapping feature distributions in some of the classes. This overlap made it challenging for the SVM to effectively separate the classes, particularly in cases where the boundaries between different types of glass were less distinct.

7. Conclusion

7.1 Achievements

Multiple machine learning models were successfully trained and evaluated for the glass type classification task. Among them, the Multilayer Perceptron (MLP) classifier achieved the highest performance, reaching an accuracy of up to 82

7.2 Learnings

Effective preprocessing techniques were essential to enhance model performance. Feature scaling ensured that all input variables contributed equally to model learning, while dimensionality reduction using Principal Component Analysis (PCA) helped reduce noise and improve computational efficiency without significant loss of information. Additionally, handling class imbalance was critical for achieving better generalization and ensuring that minority classes were not underrepresented during model training.

7.3 Final Model

The Multilayer Perceptron (MLP) classifier was chosen for deployment due to its superior accuracy of up to 82% and its robust performance across various metrics. Its ability to model complex relationships in the data made it the most suitable choice for the classification task, ensuring reliable and consistent results in real-world applications.