A SEMMA-Based Analysis and Modeling of the Iris Dataset

Siri Batchu
Department of Software Engineering
San Jose State University

October 21, 2024

Abstract

This paper presents an analysis of the Iris dataset using the SEMMA (Sample, Explore, Modify, Model, Assess) methodology, a systematic approach often used in data mining. We employed the Random Forest algorithm for classification and assessed the model's performance. The Iris dataset, consisting of sepal and petal measurements of three Iris species, serves as a benchmark for classification tasks. Results show high classification accuracy, demonstrating the efficiency of the SEMMA approach and the suitability of Random Forest for this dataset.

1 Introduction

Data mining has become essential in transforming raw data into actionable insights. One effective methodology for structuring this process is SEMMA (Sample, Explore, Modify, Model, and Assess), developed by the SAS Institute. This paper applies SEMMA to the Iris dataset, which consists of 150 samples from three species of Iris flowers (Iris-setosa, Iris-versicolor, and Iris-virginica), each described by four features: sepal length, sepal width, petal length, and petal width.

2 Methodology: SEMMA Framework

2.1 Sample

The Iris dataset is relatively small, consisting of 150 observations, 50 from each species. No subsampling was necessary due to the manageable size of the dataset.

2.2 Explore

Exploratory Data Analysis (EDA) was performed to understand the relationships between features and the target variable. Summary statistics revealed clear differences in petal length and width across species, while pairplots demonstrated clear separability of species based on petal measurements. Missing data were not present, eliminating the need for imputation.

2.3 Modify

Modifications to the dataset were minimal since the data was already clean. However, the categorical species column was transformed into numerical values for modeling. Feature standardization was applied to ensure that all features had equal weight in the model, preventing any one feature from disproportionately influencing the classification process.

2.4 Model

The Random Forest classifier was selected as the model due to its robustness and ability to handle both categorical and numerical data. The dataset was split into training (70%) and testing (30%) sets. After fitting the model on the training data, predictions were made on the test set.

2.5 Assess

The model was evaluated using classification accuracy, precision, recall, and F1-score. A confusion matrix was generated to visualize model performance across the three species.

3 Results and Discussion

3.1 Exploration Results

Summary statistics indicated that petal length and petal width were the most discriminative features. Pairplots confirmed that while sepal features slightly overlapped between species, petal features provided clearer boundaries.

3.2 Model Performance

The Random Forest model achieved an accuracy of 97.8% on the test set. The classification report in Table 1 demonstrates high precision and recall across all species.

The confusion matrix shows minimal misclassification between Iris-versicolor and Iris-virginica, which is expected given their closer proximity in feature space.

Table 1: Classification Report for Iris Species Prediction

Species	Precision	Recall	F1-Score
Iris-setosa	1.00	1.00	1.00
Iris-versicolor	0.94	1.00	0.97
Iris-virginica	1.00	0.94	0.97

3.3 Discussion

The high classification accuracy reflects the distinctiveness of the species in terms of their petal and sepal dimensions. While Iris-setosa is perfectly separable, there is slight overlap between Iris-versicolor and Iris-virginica, which was accurately handled by the Random Forest algorithm. This confirms the strength of ensemble models in handling complex feature interactions.

4 Conclusion

The SEMMA methodology offers a structured approach to data analysis that ensures each step of the process is performed thoroughly. In this study, SEMMA was applied to the Iris dataset, demonstrating the efficacy of the methodology and the Random Forest model in classifying Iris species. The high accuracy achieved shows that even simple datasets, when processed systematically, can yield valuable insights.

Future work could explore other machine learning models and their comparative performance on this dataset, or extend the analysis to more complex datasets, further leveraging the SEMMA process.

References

[1] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 179-188.

[2] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

[3] SAS Institute Inc. (2007). *Data Mining Using SAS Enterprise Miner: A Case Study Approach*.