Automated Iris Species Classification: A CRISP-DM Approach

Koluguri Ruchitha Reddy

September 2023

Abstract

The Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology is to classify iris flowers into distinct species based on sepal and petal measurements. The study showcases the systematic journey from problem definition to model deployment, emphasizing the structured approach's efficacy in achieving accuracy and efficiency. Through data understanding, preparation, modeling, and evaluation, the paper demonstrates the potential of a Decision Tree classifier in achieving promising results. The insights offered by the classification report and discussions on deployment contribute to a valuable resource for data mining practitioners, underlining the importance of systematic analysis in data science and machine learning.

1 Introduction

Comprising measurements of sepals and petals from three distinct species of the iris flower, this dataset presents an inviting yet challenging classification problem. As we venture into the depths of the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, the Iris dataset offers a familiar terrain to demonstrate the comprehensive steps from business understanding to deployment. Through this journey, not only do we aim to classify these beautiful flowers accurately, but we also seek to highlight the systematic approach that CRISP-DM brings to any data mining project, ensuring robustness, clarity, and efficiency.

2 CRISP-DM Methodology

2.1 Business Understanding

In this phase, the primary focus is on understanding the problem at hand and defining clear objectives. It involves identifying business goals, requirements, and constraints. By the end of this phase, you should have a well-defined understanding of what needs to be achieved and why.

2.2 Data Understanding:

Data Understanding involves exploring and familiarizing yourself with the data. This phase includes data collection, data description, and initial data quality assessment. It helps in identifying the data's structure, quality, and any potential issues that need to be addressed.

2.3 Data Preparation:

Data Preparation is the phase where data cleaning, transformation, and preprocessing take place. This step ensures that the data is in a suitable format for modeling. Data is cleaned of inconsistencies, missing values are handled, and features are prepared for analysis.

2.4 Modeling

In the Modeling phase, various modeling techniques are applied to the prepared data. This phase includes selecting appropriate algorithms, training models, and assessing their performance. It's where you build and evaluate models to solve the specific problem defined in the Business Understanding phase.

2.5 Evaluation:

The Evaluation phase focuses on assessing the quality and performance of the models developed in the previous phase. This includes model testing and validation to ensure that it meets the project's objectives. Evaluation helps determine if the models are suitable for deployment.

2.6 Deployment

Deployment is the final phase where the selected models are integrated into the operational environment. This phase involves creating the necessary infrastructure for making predictions on new data and monitoring the model's performance in real-world use.

3 Implementation

3.1 Understanding the Objective

Objective: Classify iris flowers into one of the three species based on their sepal and petal dimensions. This could be useful for automated plant identification systems or botanical research.

3.2 Data Understanding

Data Overview: The Iris dataset consists of 149 entries with the following columns:

Sepal Length (5.1) Sepal Width (3.5) Petal Length (1.4) Petal Width (0.2) Species (Iris-setosa)

3.3 Data Preparation

Renamed the columns for clarity. Visualized the data distributions for each species to better understand the relationships between the features.

3.4 Modeling

Model Selection: Given the nature of the dataset and the task (classification), we used a Decision Tree classifier for demonstration purposes. Model Training: The Decision Tree classifier achieved an accuracy of approximately 91.11

3.5 Evaluation:

Classification Report:

Iris-setosa: Precision 1.00, Recall 1.00, F1-score 1.00 Iris-versicolor: Precision 0.91, Recall 0.77, F1-score 0.83 Iris-virginica: Precision 0.80, Recall 0.92, F1-score 0.86

3.6 Deployment

Deployment is the final phase where the selected models are integrated into the operational environment. This phase involves creating the necessary infrastructure for making predictions on new data and monitoring the model's performance in real-world use.

4 Conclusion

The journey through automated iris species classification via CRISP-DM has been an insightful exploration of the data mining process. Starting with a clear business objective of classifying iris flowers, the data understanding phase revealed a clean dataset and guided meaningful data preparation. Modeling with a Decision Tree showcased the dataset's potential with a 91.11 accuracy. The evaluation phase provided insights into precision and recall, and the deployment steps were outlined. This systematic approach not only achieved accurate classification but also underscored the importance of structured data mining in real-world problem-solving. It serves as a strong foundation for future data mining endeavors in the realm of data science and machine learning.

Acknowledgements.

Please place your acknowledgments at the end of the paper (just before the list of references), preceded by an unnumbered run-in heading (i.e. 3rd-level heading).

References

1. https://www.kaggle.com/datasets/vikrishnan/iris-dataset
2. PyCaret. (2020). An open-source, low-code ML library in Python.
3. https://github.com/ruchithareddy269/Dm-Assignment-3