Machine Learning Operations (MLOps)

Assignment 2

Task 3 – Explainable AI (XAI) Implementation

Group Number 76

CHAUDHARI AAKASH VINAYAK (2022ac05607)
AATIF HUSSAIN WAZA (2022ac05405)
AJIT KUMAR YADAV (2022ac05720)
MOHAMMAD ZUBAIR (2022ac05121)

Contents

1. Introduction

1.1 Objective

The goal of this task is to apply Explainable AI techniques to make the model's predictions interpretable. In this document, we use LIME (Local Interpretable Model-agnostic Explanations) to provide insights into the decision-making process of a Random Forest model trained on the Iris dataset.

1.2 Importance of Interpretability

Interpretability is crucial in machine learning models as it helps in understanding how models make predictions, which in turn aids in trust, debugging, and improving the models. XAI tools like LIME allow us to dissect model predictions and understand the contribution of each feature.

2. Dataset

2.1 Description

The Iris dataset is a classic dataset used for classification tasks. It includes measurements of iris flowers from three different species.

- Dataset Name: Iris
- Features: Sepal Length, Sepal Width, Petal Length, Petal Width
- Target Variable: Species
- Size: 150 instances, 4 features

2.2 Data Preparation

The dataset is split into training and testing sets to evaluate the performance of the model and to ensure that the explanations are based on unseen data.

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0

Distributions

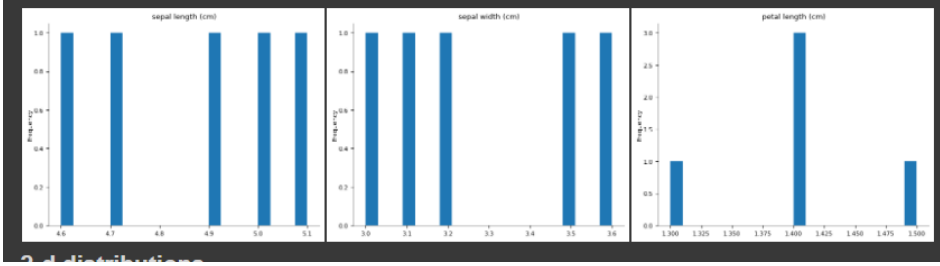

Figure 1 Dataset Description

3. Model Training

3.1 Model Choice

A Random Forest classifier is used for this task. Random Forest is an ensemble learning method that combines multiple decision trees to improve classification accuracy and control over-fitting.

3.2 Model Training

The Random Forest model is trained on the training set of the Iris dataset. The model's performance is evaluated on the test set, and its accuracy is noted.

4. Explainable AI with LIME

4.1 Introduction to LIME

LIME (Local Interpretable Model-agnostic Explanations) is a technique used to explain individual predictions of machine learning models. It works by approximating the model with a simpler, interpretable model locally around the instance being explained.

4.2 Creating a LIME Explainer

To use LIME, an explainer is created that specifies how to interpret the model's predictions. The explainer is configured with training data, feature names, and class names to provide context to the explanations.

4.3 Selecting an Instance to Explain

A specific instance from the test set is selected for explanation. This instance is representative of the kind of predictions the model makes and is used to demonstrate how the model arrived at its prediction.

4.4 Generating the Explanation

The explainer generates an explanation for the selected instance by approximating the model’s behavior with an interpretable model. The explanation reveals the contribution of each feature to the model's prediction for that instance.

4.5 Visualizing the Explanation

The explanation is visualized to illustrate how different features contribute to the model’s prediction. This visualization helps in understanding which features are most influential in the model's decision-making process.

5. Results

5.1 Model Performance

The performance of the Random Forest model on the test set is summarized, including metrics such as accuracy. This provides a baseline understanding of the model's effectiveness before applying LIME.

5.2 Explanation Insights

The explanation provides insights into how the model makes predictions for the selected instance. By analyzing feature contributions, it becomes clear which features have the most impact on the model’s prediction.

5.3 Visualization Analysis

The visualization of the explanation helps to understand the model’s decision-making process in a more intuitive manner. It highlights the importance of each feature and how they influence the final prediction.

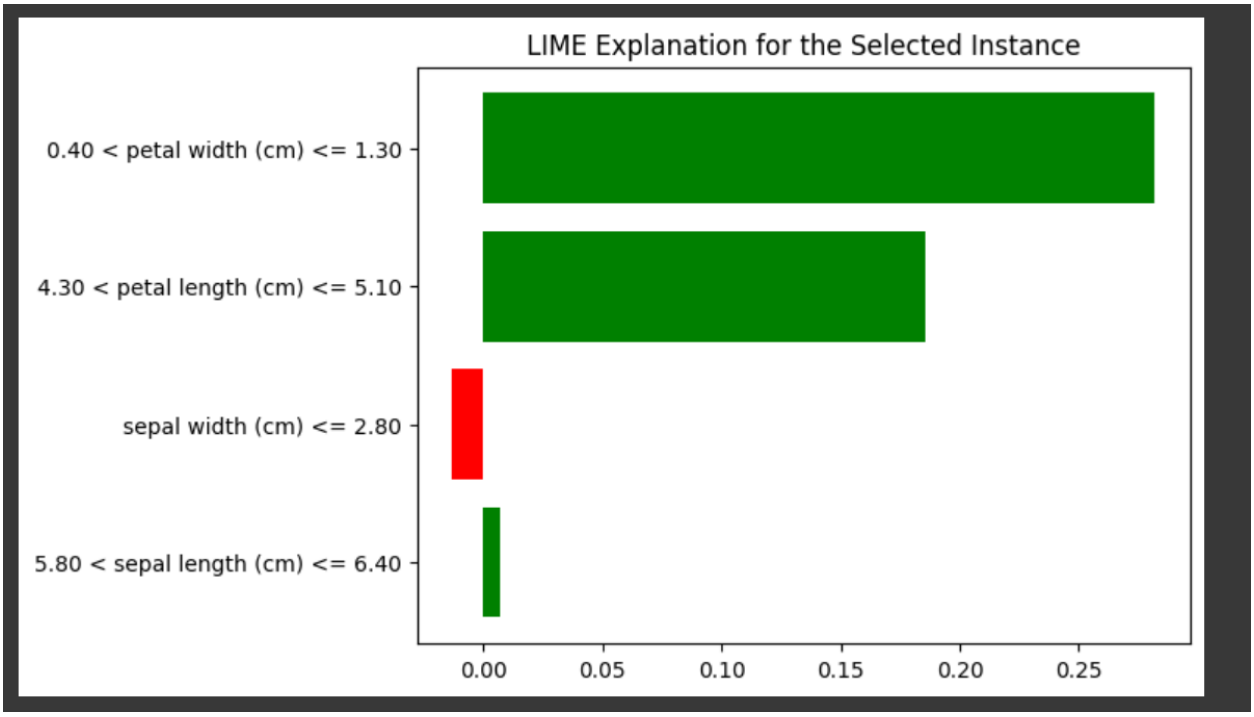

Figure 2 LIME Explanation

6. Conclusion

6.1 Summary

The application of LIME has enabled us to interpret the predictions of the Random Forest model trained on the Iris dataset. The explanations and visualizations provided valuable insights into how the model arrived at its predictions.

6.2 Future Work

Further exploration could involve using other XAI techniques like SHAP (SHapley Additive exPlanations) or comparing different models to see how explanations vary. Enhancing model transparency can help in making more informed decisions and improving model performance.

GitHub Link:
https://github.com/AakashChaudhari03/MLOPS_ASSIGNMENT_2_GRP_NO_76