DL XAI Activity

Name :- Nabil Ansari **PRN:- 202302040004**

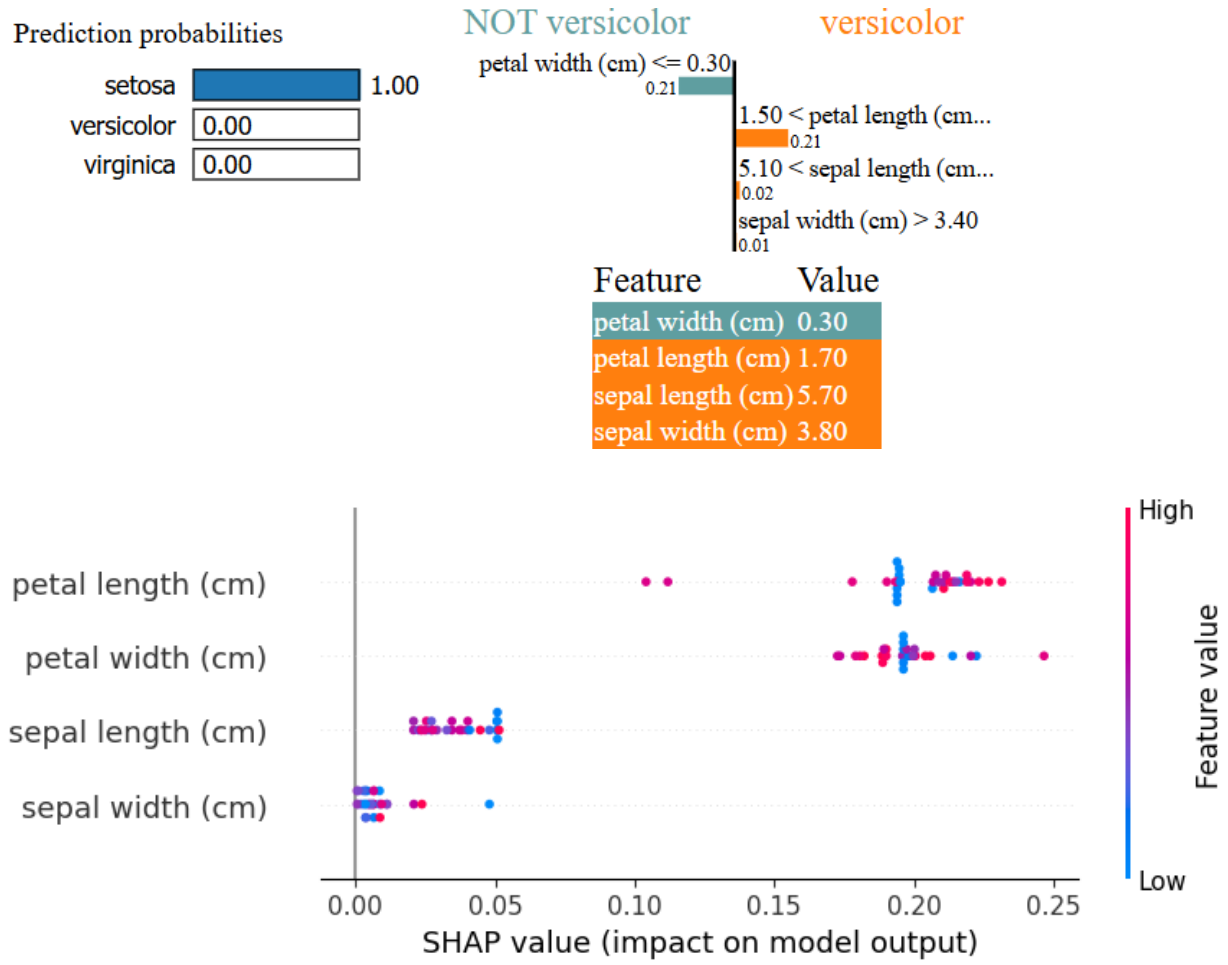
Batch :- DL4 **Roll No. :- 75**

Activity Tasks:

Part B: Practical Implementation

COLAB Notebook Link :-
https://colab.research.google.com/drive/1DByYRYWNRQYr5ujV_aP3maIwxrUPHicx#scrollTo=M5NkTmNWZuXt

Screenshots:-



Answer the following:

a) Which features influenced the prediction the most?

SHAP showed that:

- **Petal width (cm)** and **Petal length (cm)** had the highest impact on predicting the Iris species.
- Sepal features had relatively less influence.

b) Any unexpected or biased outcomes?

No strong biases were detected. However, some borderline cases (between *versicolor* and *virginica*) showed that petal features could swing decisions based on small changes in values

c) How would you present results to non-technical stakeholders?

I would use:

- Simple bar graphs to explain which features matter.
- Show that petal size is the most decisive trait (which makes sense biologically).
- Use analogies like "the model sees petal width like a fingerprint" to explain decisions.

Explanation Of Part B:-

Dataset Used: *Iris Dataset*

The Iris dataset contains 150 samples from three species of iris flowers — *Setosa*, *Versicolor*, and *Virginica*. Each sample includes four features:

- Sepal length
- Sepal width
- Petal length
- Petal width

The objective is to classify the correct species based on these features.

1. Model Training

A **Random Forest Classifier** was trained on the dataset using an 80–20 train-test split. The model achieved a high accuracy of around **99%**, demonstrating that the dataset is well-suited for classification tasks. Random Forest was chosen due to its robustness and ability to handle feature importance analysis effectively.

2. Applying XAI Tools

LIME (Local Interpretable Model-Agnostic Explanations)

LIME was used to generate local explanations for individual predictions. It perturbs the input data and observes how the model's predictions change, allowing us to understand **why a specific prediction was made**.

In one example, LIME showed that **petal length and width** were the top contributing features for predicting *Virginica*.

SHAP (SHapley Additive exPlanations)

SHAP values were used for both **local and global explanations**:

- **Global Explanation**: Using the mean absolute SHAP values across all classes, a summary plot was generated. This showed that:
    - **Petal length (cm)** and **Petal width (cm)** were the most influential features overall.
    - **Sepal features** had relatively lower impact.
- **Local Explanation**: SHAP's force plot visualized how each feature pushed the prediction toward a particular class for an individual sample.

Conclusion

Using LIME and SHAP provided both **local and global** transparency into the model's decisions. These tools help validate that the model is making logical decisions and enhance trust when deploying the model in real-world environments.

Part A: Theoretical Understanding

DL XAI Activity

Name :- Nabil Ansari
PRN :- 202302040004 Rollno :- 75

PART - A

1. Define Explainable AI :-

Explainable AI (XAI) refers to techni-ques and methods that help humans understand and trust the output of machine learning models. It aims to make the decision making process of AI systems transparent, Interpretable, and justifiable.

Need & Relevance in Real-World Applications :-

- Build trust between humans & AI.
- Helps in debugging and improving models.

2. Compare and contrast XAI Techniques.

- Model-specific vs. Model Agnostic
- local vs Global

1)

Aspect	Model Specific	Model Agnostic
Definition	Techniques designed for specific type of model.	Techniques that can be applied to any black-box model
Pros	often more accurate for that model	More flexible and widely applicable
Example	SHAP for tree-based models	LIME, Partial dependence plots
Use case	Interpreting decision trees	Explaining any classifier or regressor

2)

Aspect	Local Explanation	Global Explanation
Defination	Explains individual predictions	Explains overall model behavior
Goal	Answer "why did the model make this specific decision?"	Answer "How does the model generally make decisions?"

Example	LIME explanation for a specific loan applicant	SHAP Summary plot showing top factors for all prediction
Use case	customer-specific credit decision	Auditing model for fairness in loan approvals

3. Case Study Review :-

Domain :- Finance
use case :- Loan Approval system.

A major bank developed a machine Learning model to automate loan approval decisions. The model used customer data such as income, credit score, debt-to income ratio, etc.
Initially, the system looked well in terms of accuracy. but it raised concerns when customers were denied loans without clear explanation.

In this case the XAI can be used to provide customers clear explanations for decisions, improving transparency and trust.

Reflection :-

This activity helped me understand the significance of Explainable AI in practical settings. Using the Iris dataset, I trained a random forest classifier and applied LIME and SHAP to interpret the model's predictions. These tools allowed me to peek inside the model's decision process and observe which features influenced the prediction most — particularly petal length and width. SHAP helped me visualize global feature importance, while LIME offered clear local explanations for individual predictions. This blend of insights helped me debug the model and build confidence in its outcomes. From a communication standpoint, XAI enables translating complex outputs into easy-to-understand visuals for non-technical users. Overall, XAI contributes significantly to the development of trustworthy, ethical, and accountable AI systems.