**Essential Question #3 - AS**

*JD’s AI/ML Laboratory*

**Due: 07/11/2025, 11:59 PM EST**

**General Instruction**: Please **answer all questions** and provide relevant support from literature using **APA references**. You can answer the questions using bullet points with citations (preferably from peer-review journals). Keep in mind that most science competitions focus on participants having a very strong grasp of the literature. I would highly recommend printing out the relevant journal articles and do a deep dive. Using a highlighter and writing notes on the journal article can help you narrow down important content for future reference. I will be thorough in my review of your response and expect all the answers to be well thought-out and clearly explained with supporting literature.

1. **Explore various interpretability methods**

Shapely value: Shapley Value is additive and locally accurate. If you add up the Shapley Values of all the features, plus the base value, which is the prediction average, you will get the exact prediction value. This is a feature that many other methods do not have.

The plot shows the Shapley Value of each feature, representing the contribution that pushes the model outcome from base value to the final prediction. The red color means a positive contribution, and the blue color means a negative contribution.

1. **What is the difference between interpretability and explainability**

**Interpretability:** refers to the ability to understand the decision-making process of an AI model. An interpretable model is transparent in its operation and provides information about the relationships between inputs and outputs. An interpretable algorithm can be explained clearly and understandably by a human being. Interpretability is therefore important to ensure that users can understand and trust artificial intelligence models.

**Explainability:** pertains to the ability to explain the decision-making process of an AI model in terms understandable to the end user. An explainable model provides a clear and intuitive explanation of the decisions made, enabling users to understand why the model produced a particular result. In other words, explainability focuses on why an algorithm made a specific decision and how that decision can be justified.

1. Explore how to balance model performance with model interpretability

High ROC-AUC and recall (both >0.8) = good performance

High interpretability

1. Look for literature on model performance and interpretability
2. Train a machine learning model that would help you answer your scientific question
3. Draft PowerPoint with Background, Research Question, Hypothesis, research plan, EDA, Analysis Flowchart, ML model analysis, results, conclusions.
4. Prepare for oral presentation on Friday
5. Draft research paper outline

(see presentation)

References

Xu, L. (n.d.). Interpretability Methods in Machine Learning: A Brief Survey. Two Sigma. <https://www.twosigma.com/articles/interpretability-methods-in-machine-learning-a-brief-survey/>

A Data Odyssey. (2023, March 12). SHAP values for beginners | What they mean and their applications. YouTube. <https://www.youtube.com/watch?v=MQ6fFDwjuco>

giosmin, E. (2023, July 4). Interpretability vs explainability: Understanding the Differences and Importance in the World of Artificial Intelligence. XCALLY Motion. <https://www.xcally.com/news/interpretability-vs-explainability-understanding-the-importance-in-artificial-intelligence/>

Mariotti, E., Moral, A., & Gatt, A. (2023). Exploring the balance between interpretability and performance with carefully designed constrainable Neural Additive Models. Information Fusion, 99, 101882–101882. <https://doi.org/10.1016/j.inffus.2023.101882>