Kenan Stredic
Shapely Values Summary

1. Introduction to Shapley Values:
   - Shapley values are introduced as an approach from cooperative game theory for explaining machine learning models.
   - They are praised for their desirable properties and are commonly used to provide explanations for model predictions.
   - The tutorial aims to help readers understand how to compute and interpret Shapley-based explanations of machine learning models using the shap Python package.

2. Explaining Linear Regression Model:
   - The tutorial begins with explaining a simple linear regression model trained on the California housing dataset.
   - It demonstrates how to interpret the model coefficients and explains the limitations of using coefficients as a measure of feature importance.
   - Partial dependence plots are introduced as a more comprehensive way to understand feature importance in a linear model.

3. Explaining Generalized Additive Regression Model:
   - Generalized additive regression models (GAMs) are introduced as an extension of linear models where the effects of features are not constrained to be linear.
   - InterpretML's explainable boosting machines are used to train a GAM model, and Shapley values are computed to explain the model's predictions.

4. Explaining Non-Additive Boosted Tree Model:
   - A non-additive boosted tree model (XGBoost) is trained on the California housing dataset, and Shapley values are computed to explain the model's predictions.
   - The tutorial demonstrates how to interpret Shapley values for complex models like XGBoost.

5. Explaining Linear Logistic Regression Model:
   - A linear logistic regression model is trained on the adult census dataset to predict income class probabilities.
   - The tutorial highlights the difference in explaining probabilities compared to explaining linear regression models.

6. Explaining Non-Additive Boosted Tree Logistic Regression Model:
   - A non-additive boosted tree logistic regression model (XGBoost classifier) is trained on the adult census dataset, and Shapley values are computed to explain the model's predictions.
   - Various visualization techniques, including bar plots, beeswarm plots, and heatmaps, are used to interpret the Shapley values.

7. Dealing with Correlated Features:
   - Techniques for dealing with correlated input features are discussed, including hierarchical clustering to group similar features and adjusting clustering cutoffs to find optimal feature groupings.

Kenan Stredic
Shapely Values Summary

8. Explaining a Transformers NLP Model:
- The tutorial demonstrates how Shapley values can be applied to explain complex models with structured inputs, using a BERT sentiment analysis model trained on the IMDB reviews dataset.
- Token maskers are used to explain the model's predictions on text data, and various visualization techniques are employed to interpret the Shapley values.