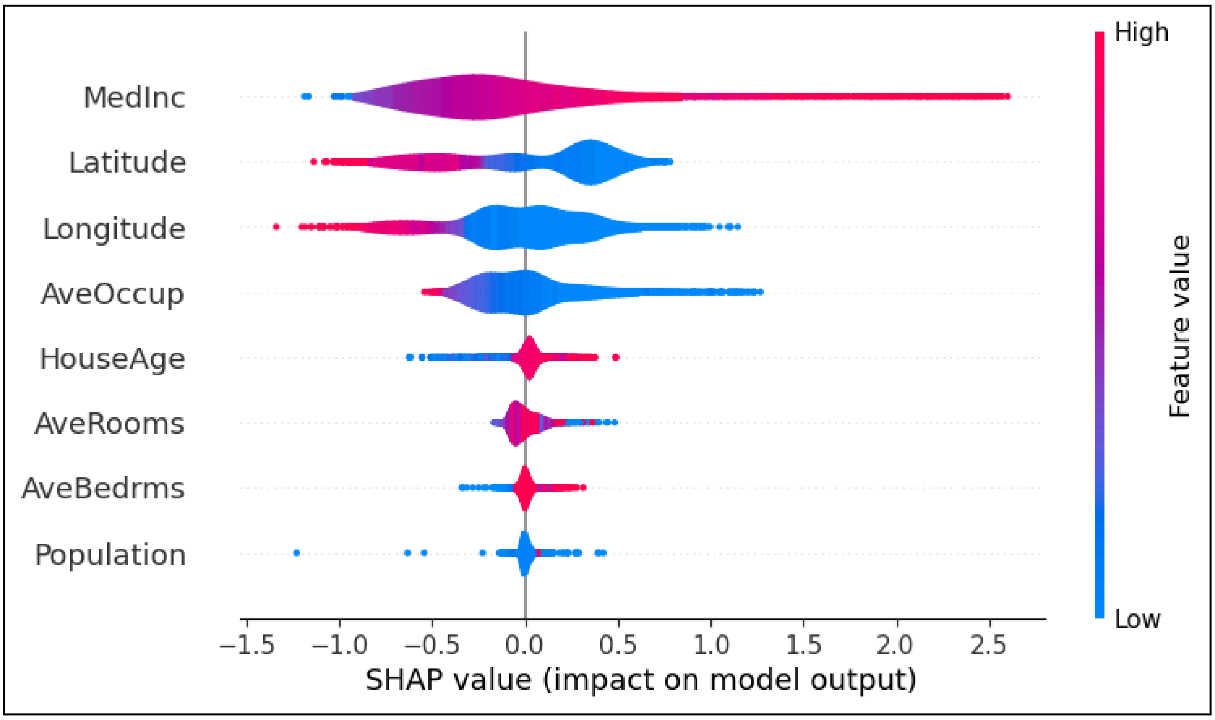
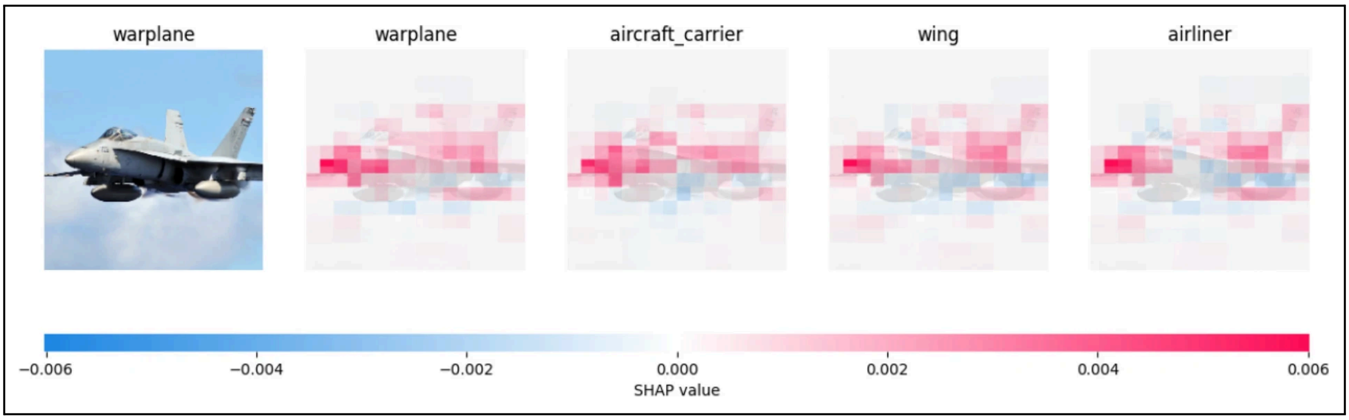
# SHAP Values

- What are they?
  - **Shapley values are a way to explain the output of any machine learning model**
  - Help with interpretability and explainability
- Each feature is assigned an "importance" value based on how much they contributed to the final prediction
- Based on game theory
- Features with a positive SHAP have a positive impact on the prediction
- Features with a negative SHAP have a negative impact on the prediction
- These values do not change even if parameters or hyperparameters are altered
  - They only measure the contribution of features
- SHAP values are model-agnostic, meaning they can be used to interpret any machine learning model, including:
  - Linear regression
  - Decision trees
  - Random forests
  - Gradient boosting models
  - Neural networks

** For us it could be helpful to determine if there is a feature that really does not impact the final prediction at all → this feature could then be removed



**Example with convolutional neural networks:**



### References

Abid Ali Awan. (2023, June 28). *An Introduction to SHAP Values and Machine Learning Interpretability*. Datacamp.com; DataCamp. https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability

SHAP. (2022). *An introduction to explainable AI with Shapley values — SHAP latest documentation*. Readthedocs.io. https://shap.readthedocs.io/en/stable/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html