Interpretability and Explainability in AI

Ankita Jamdade

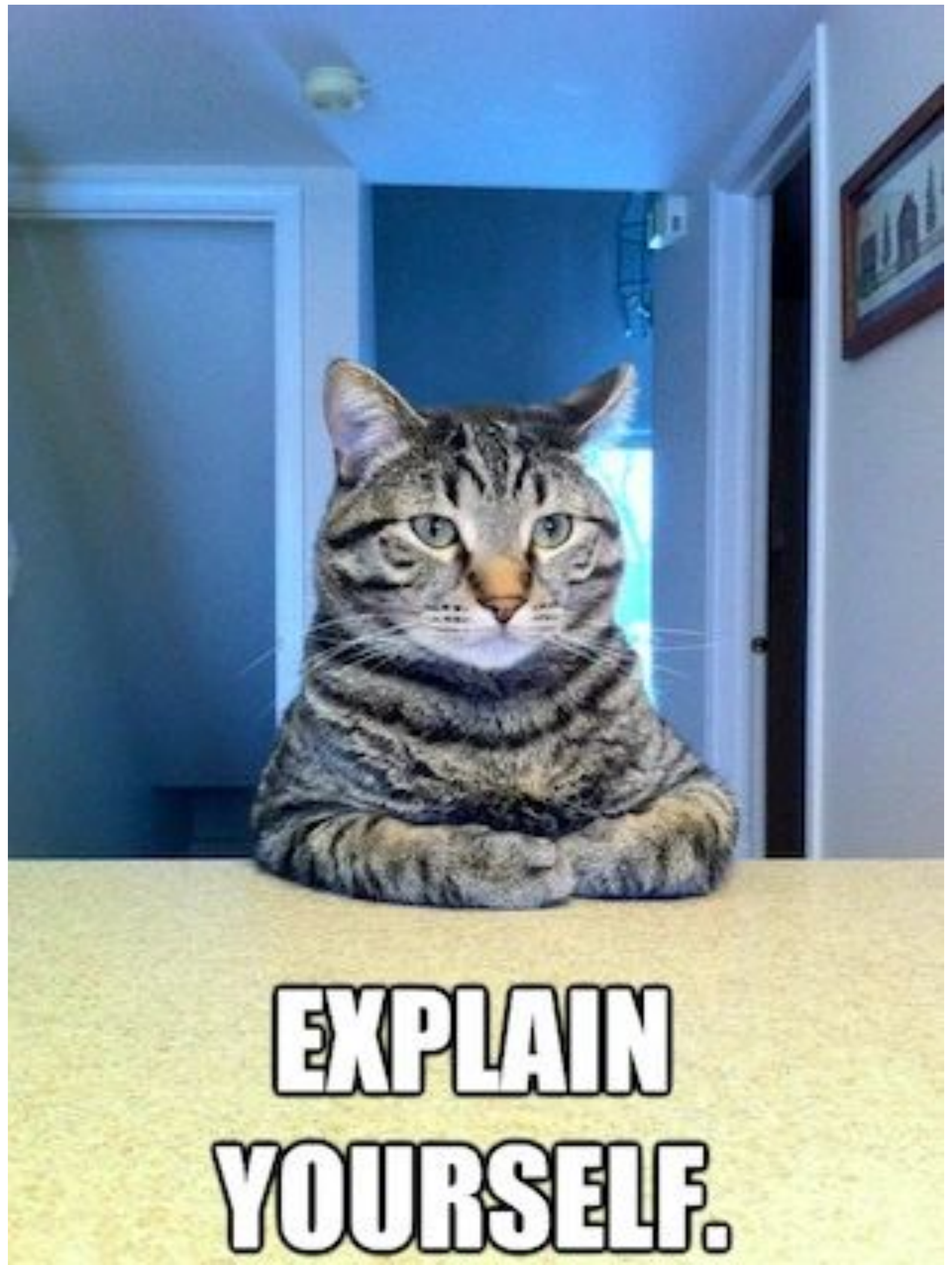
-Software/Machine Learning, AltaML
-AI Education Intern and Machine Learning Associate, Vector Institute
-Master of Data Science and AI, University of Waterloo

Reference: https://www.xcally.com/news/interpretability-vs-explainability-understanding-the-importance-in-artificial-intelligence/

Agenda

1. Introduction to Interpretability and Explainability
2. Key Concepts
3. Tools and Techniques
4. Ethical Considerations
5. Python Demo

Introduction to Interpretability and Explainability



Explainability refers to the degree to which the behaviour of a model can be explained in human-understandable terms.

Hard to know how a result was arrived at, but you mostly know why.

Interpretability refers to the ease with which a human can understand the cause of a decision made by a machine learning model.

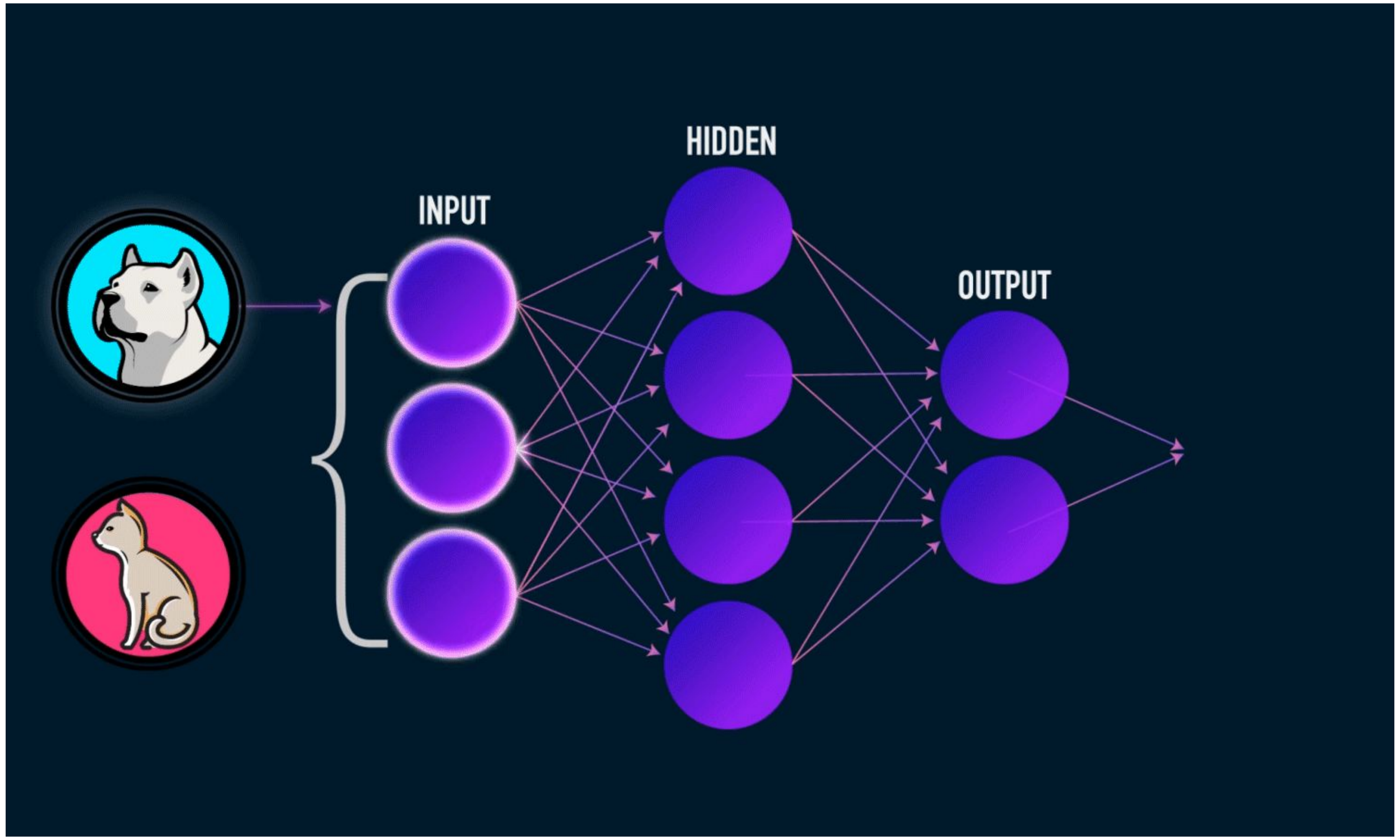
Easy to see how the algorithm arrived at its conclusion but not why each step of the decision process was created.

The terms *interpretability* and *explainability* are commonly interchangeable.

Reference: https://medium.com/swlh/explainable-vs-interpretable-ai-an-intuitive-example-6baf8fc6d402

Explainable Machine Learning

Explainable models are functions that are too complicated for a human to understand (Black-box models)



ANN trained on a given task to classify images of cats or dogs

- Difficult to explain the relationships between the input features and the response (output).
- More neurons and layers - more difficult to explain and identify what functions affect the output.

Reference: https://medium.com/@shagunm1210/the-explainable-neural-network-8f95256dcddb

Simple use-case: Wolf in the snow

Siberian Husky (Dog)

Wolf

Explainable Machine Learning

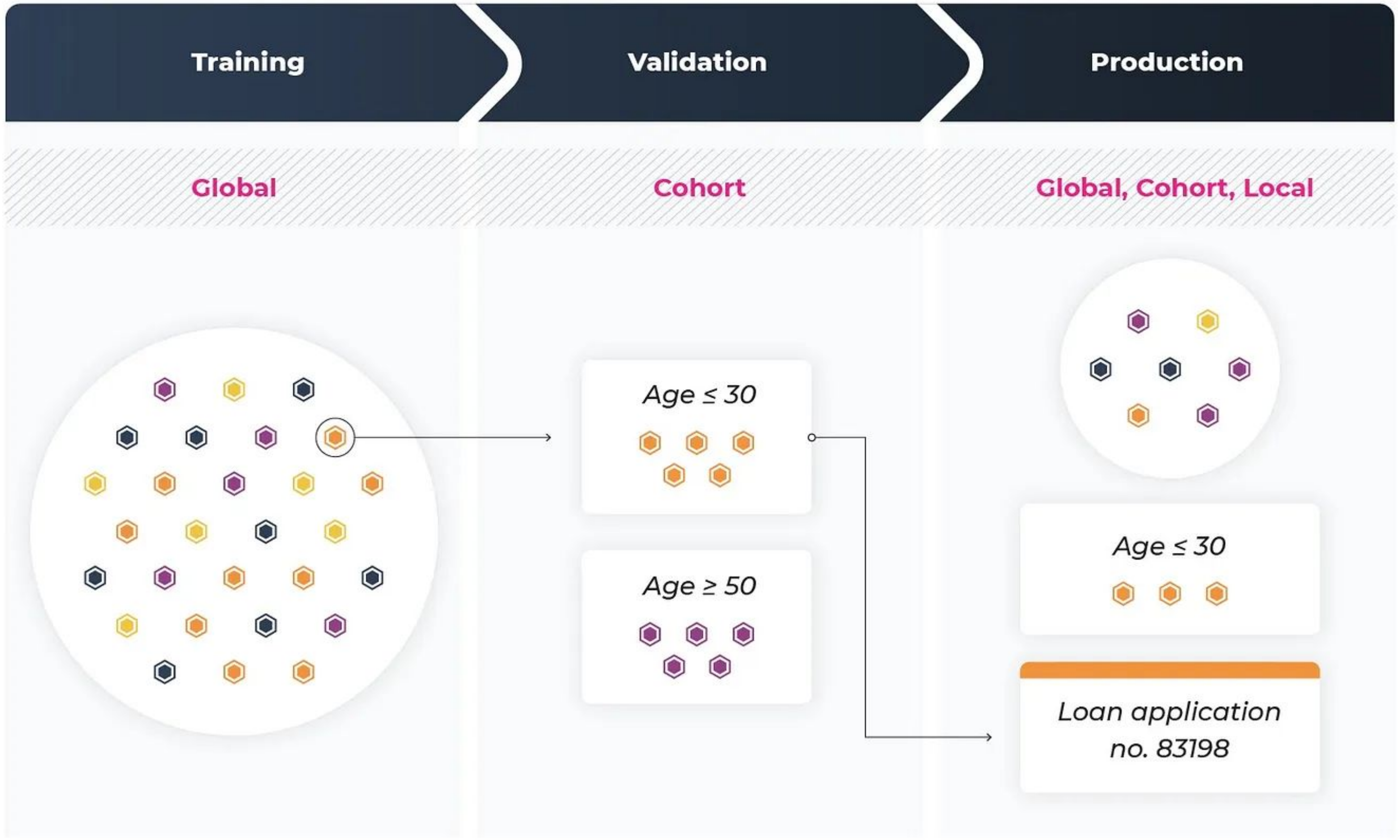
6

Global Explainability: Importance of feature contribution on the model predictions **over all of the data.**

Cohort Explainability: Importance of feature contribution on the model predictions **over a subset of the data.**

Local Explainability: Importance of feature contribution on the model predictions **over a data point.**

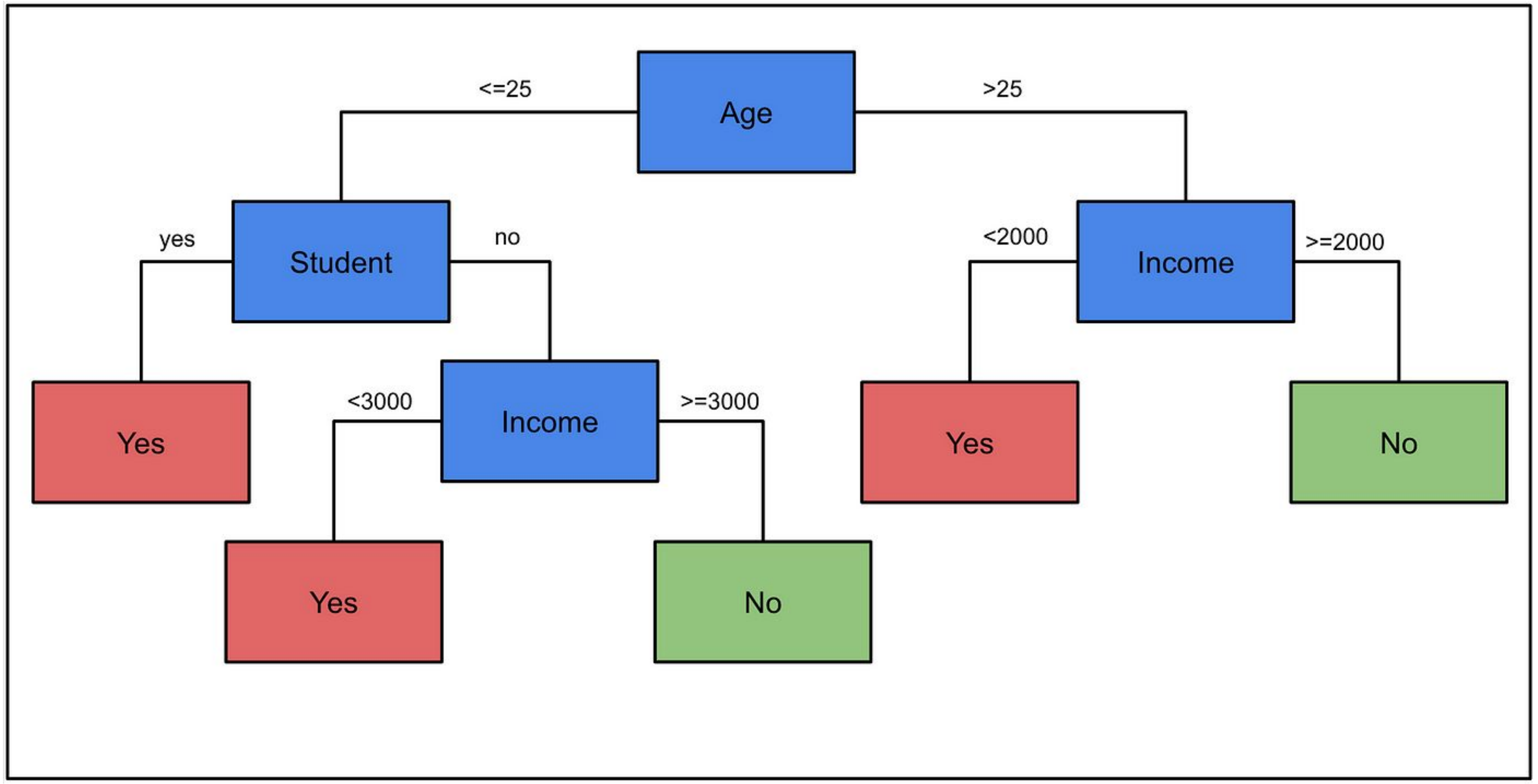
Explainability across the ML lifecycle



Reference: https://towardsdatascience.com/a-look-into-global-cohort-and-local-model-explainability-973bd449969f

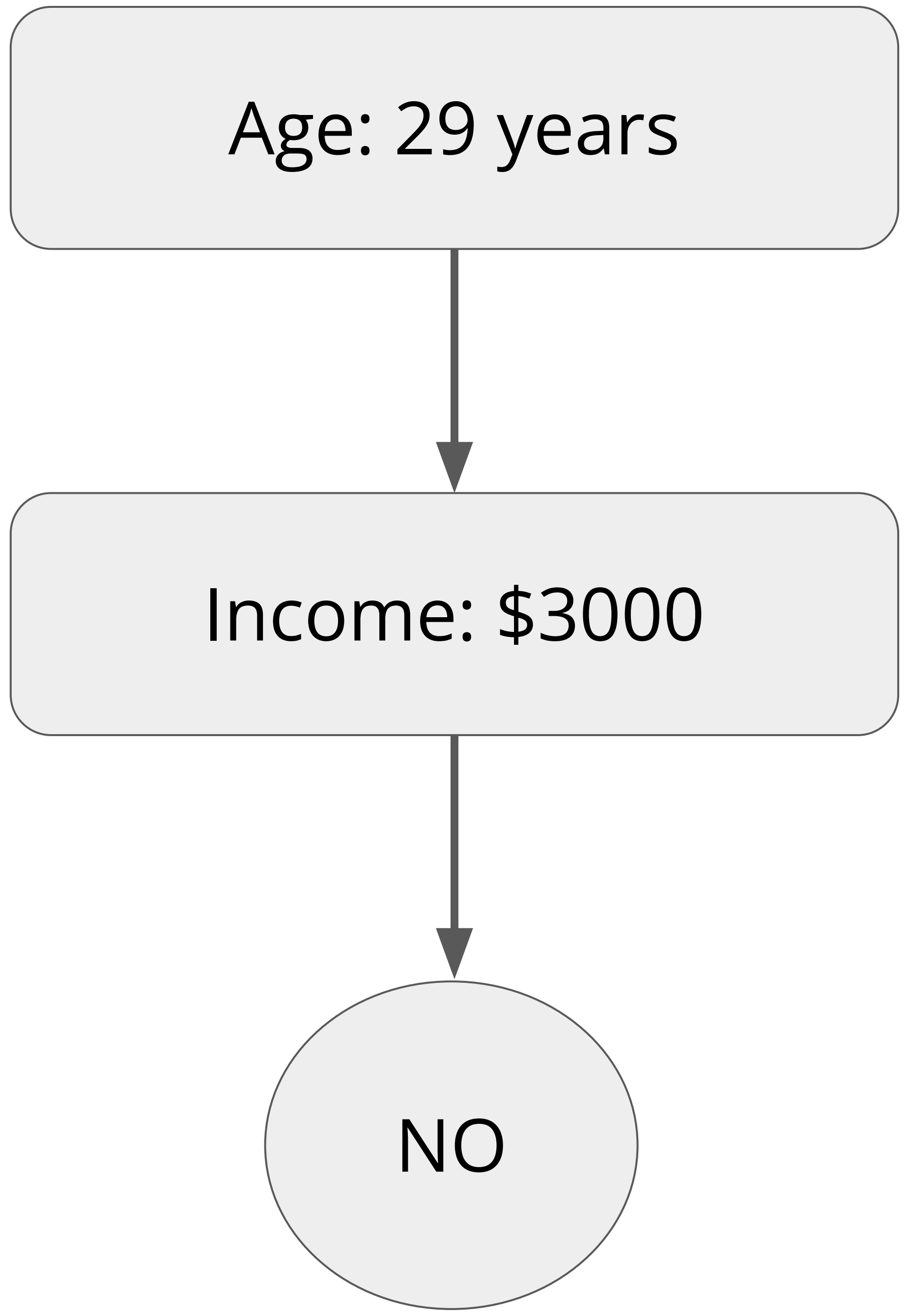
Interpretable Machine Learning

Interpretable models can be understood by a human without any other aids/techniques.



Decision Tree for Loan Default Predictions

Predicting whether loan will be defaulted



Reference: https://towardsdatascience.com/interperable-vs-explainable-machine-learning-1fa525e12f48

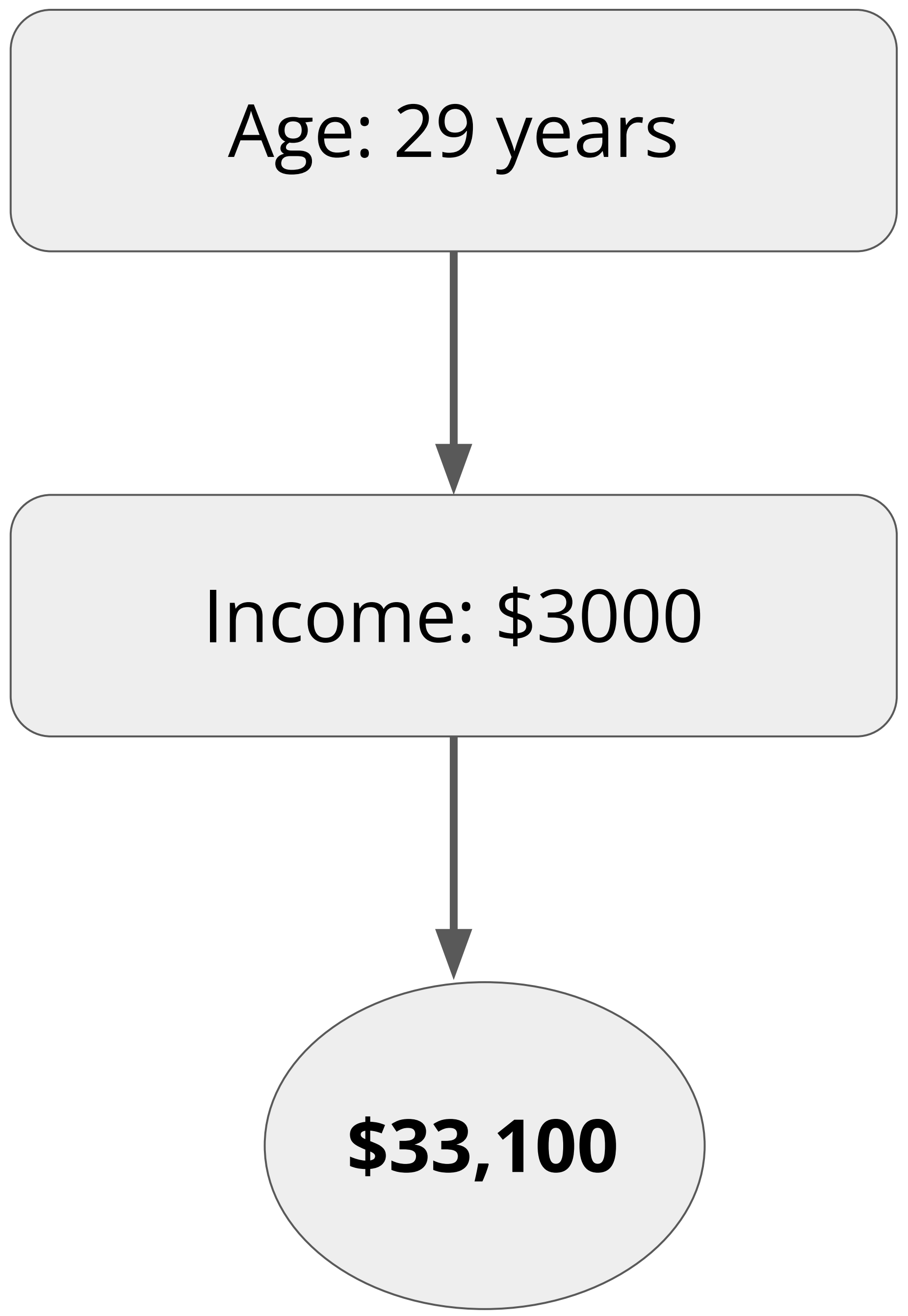
Interpretable Machine Learning

Interpretable models can be understood by a human without any other aids/techniques.

$$Y = 100*age + 10*income + 200$$

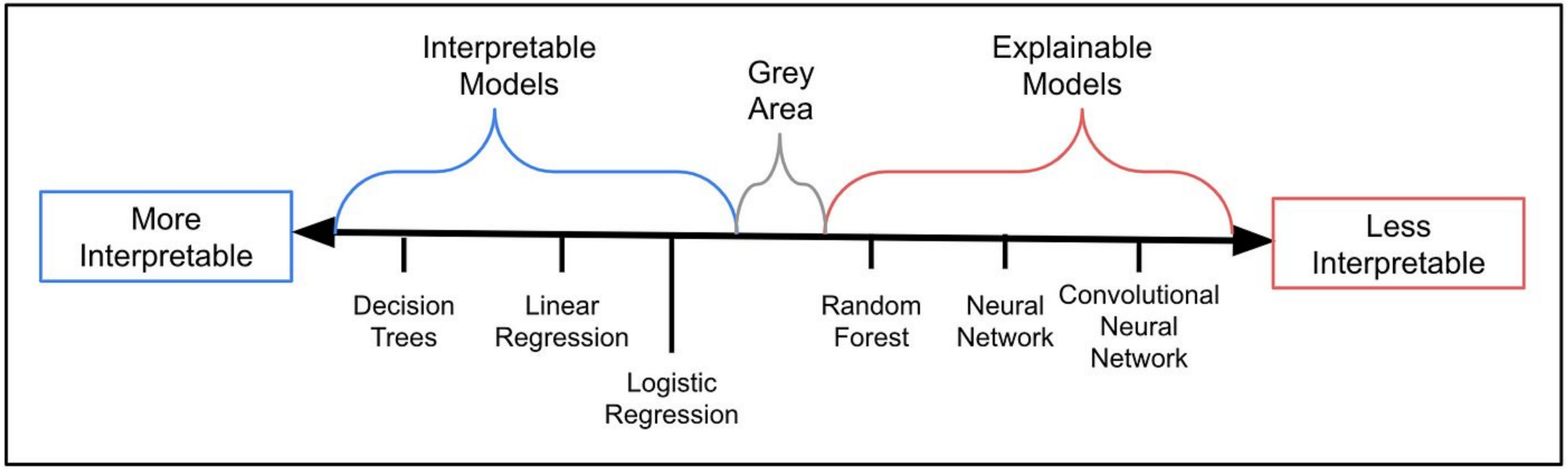
- $100 for every additional year of age
- $10 for every additional dollar of income

Predicting the maximum loan amount (Y)



Reference: https://towardsdatascience.com/interperable-vs-explainable-machine-learning-1fa525e12f48

Interpretability Dilemma

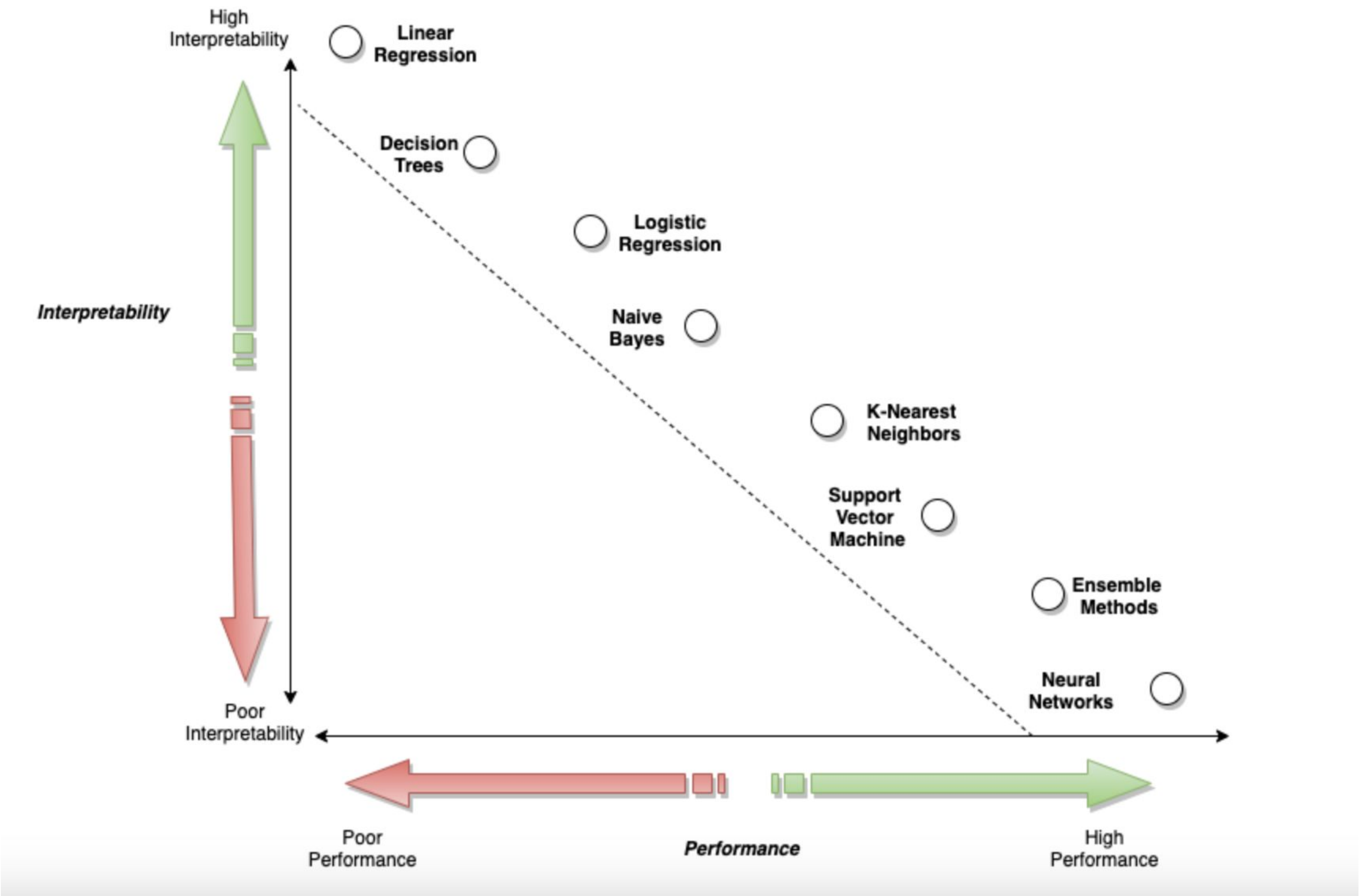


The Interpretability Spectrum

Example: A single Decision Tree or a Random forest with 2 trees is interpretable.
But, is a Random forest with 100 trees interpretable?

Reference: https://towardsdatascience.com/interperable-vs-explainable-machine-learning-1fa525e12f48

Interpretability versus performance trade-off



Interpretability versus performance trade-off given common ML algorithms

Is interpretability a hard business requirement?
- Regulations or business requirements for complete model transparency.

Can my dataset be used on a simpler model?
- If you can meet the objective using a simple AI/ML method with full transparency, select that approach.

Reference: https://docs.aws.amazon.com/whitepapers/latest/model-explainability-aws-ai-ml/interpretability-versus-explainability.html


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

Techniques to understand Explainability

- **SHAP**
  - SHapley Additive exPlanations
  - Explains why a particular example differs from the global expectation from a model
- **LIME**
  - Local Interpretable Model-Agnostic Explanations
  - Attempts to understand how local perturbations in a model's inputs affect the end-prediction of the model
- **PDPs and ICE Plots**
  - Visualise the relationship between model features and the target variable
  - Generally used when there are interactions between features

SHAP (SHapley Additive exPlanations)

SHAP values are based on Shapley values, a concept coming from game theory.

Imagine that we have a predictive model, then:

- the “game” is reproducing the outcome of the model.
- the “players” are the features included in the model.

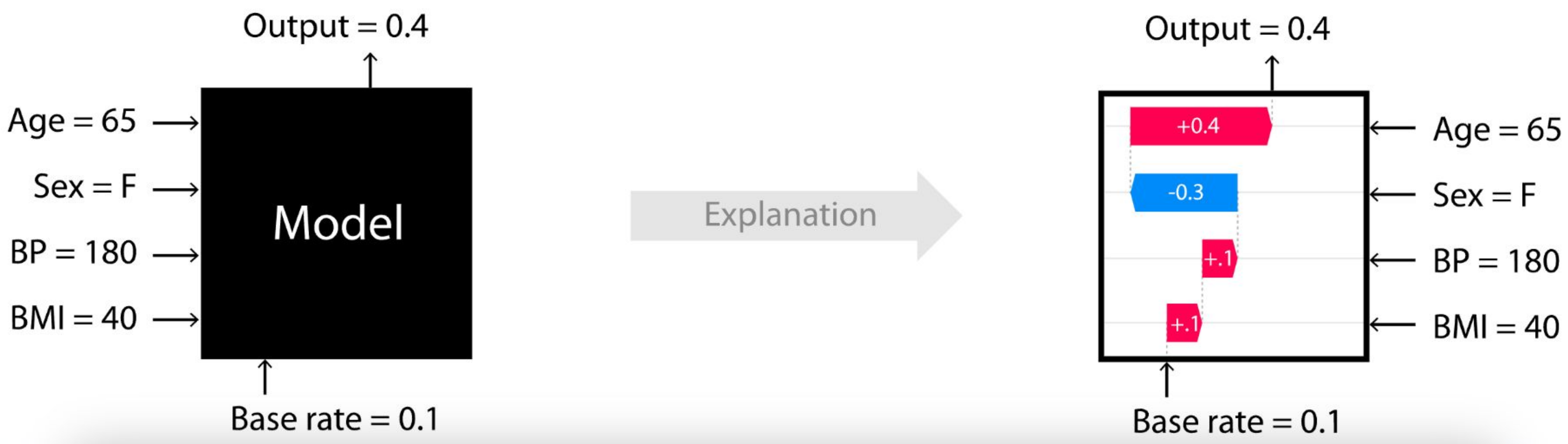
What Shapley does is quantifying the contribution that each player brings to the game

https://towardsdatascience.com/shap-explained-the-way-i-wish-someone-explained-it-to-me-ab81cc69ef30

SHAP (SHapley Additive exPlanations)

- Powerful **Python package** for understanding and debugging your models
- Tells us how each model feature contributes to an individual prediction
- By aggregating SHAP values, we can also understand trends across multiple predictions
- With a few lines of code, we are able to **identify and visualise important relationships** in our model
- SHAP Plots:
    - Waterfall plot
    - Force plots
    - Mean SHAP plot
    - Beeswarm plot
    - Dependence plots

SHAP


Output = 0.4
Age = 65
Sex = F
BP = 180
BMI = 40
Model
Base rate = 0.1
Explanation
Output = 0.4
+0.4
-0.3
+.1
+.1
Age = 65
Sex = F
BP = 180
BMI = 40
Base rate = 0.1

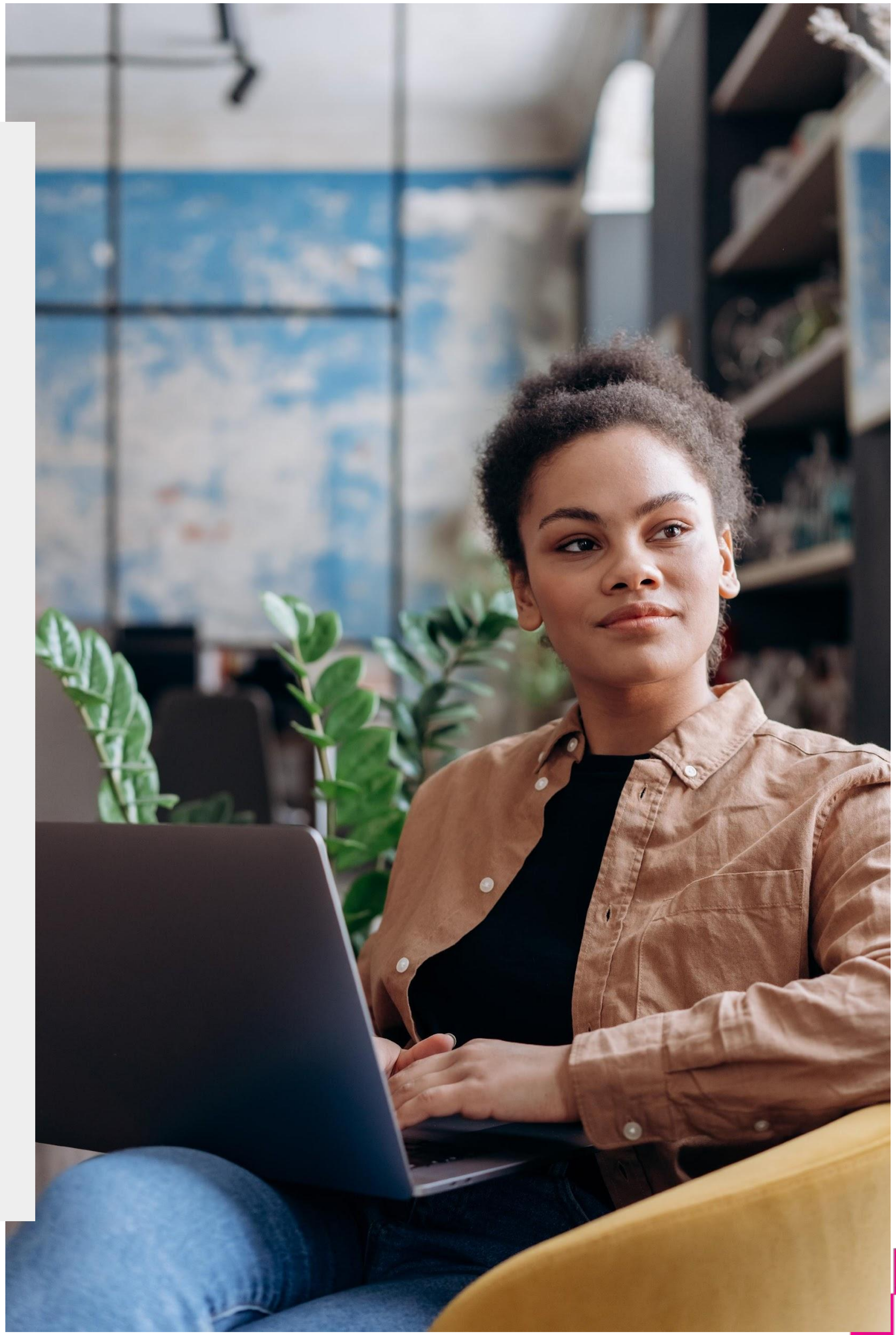
Ethical Considerations

- Facial Recognition Leads To False Arrest Of Black Man In Detroit : NPR
- Amazon scraps secret AI recruiting tool that showed bias against women
- Microsoft Chat Bot Goes On Racist, Genocidal Twitter Rampage
- Apple Card algorithm sparks gender bias inquiry

Organizations must adopt standards, processes, and controls to make AI systems ***compliant*** *and promote a culture of* ***responsible, ethical, and trustworthy AI.***

Resources

- https://adataodyssey.com/courses/xai-with-python/
- https://cloud.google.com/explainable-ai
- https://cloud.google.com/vertex-ai/docs/explainable-ai/overview
- https://github.com/marcotcr/lime
- https://github.com/slundberg/shap
- https://medium.com/60-leaders/the-ethical-concerns-associated-with-the-general-adoption-of-ai-ab893e9b5196



SHAP DEMO

Thank you!

Reach me at:
https://www.linkedin.com/in/ankitajamdade/

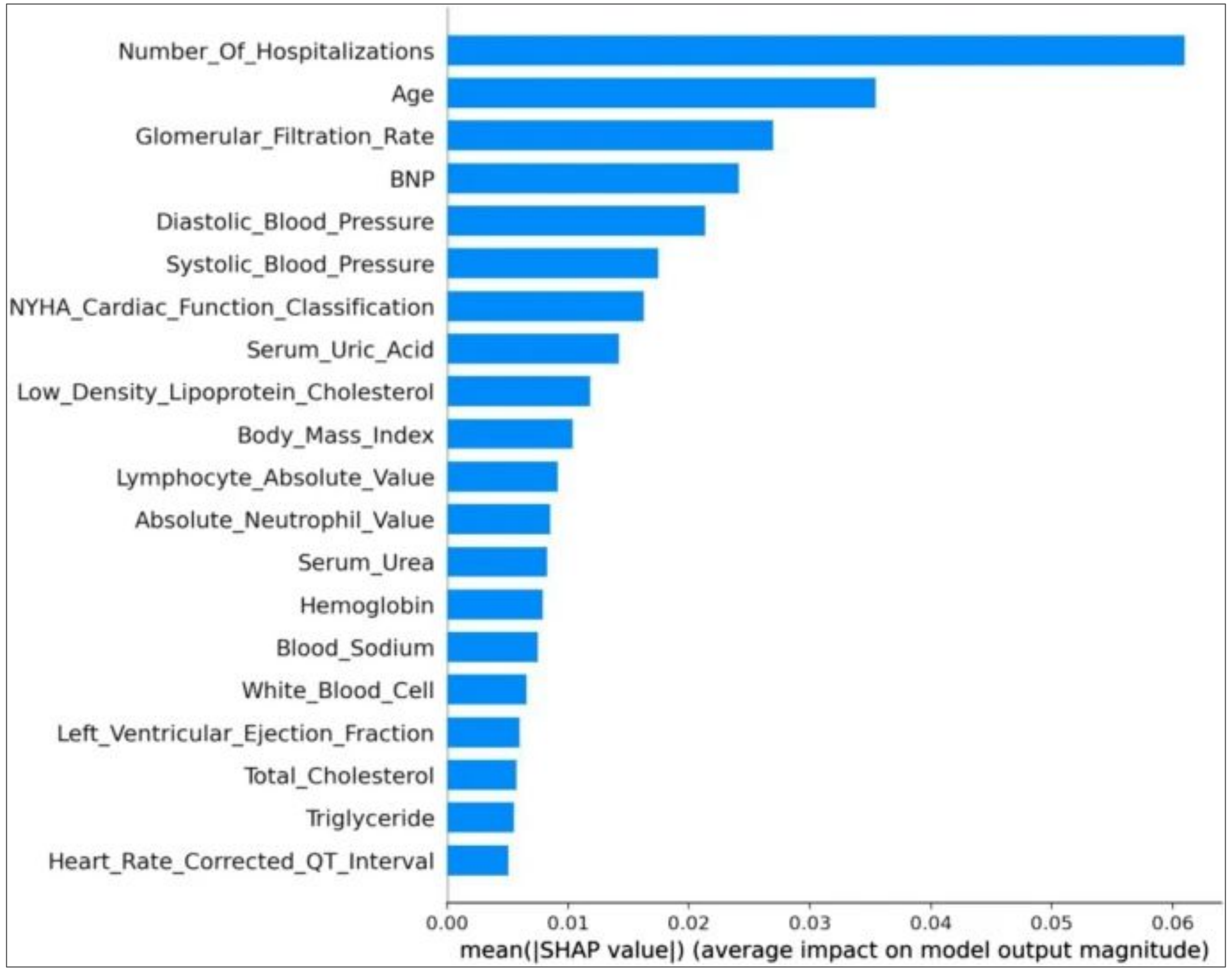
Appendix

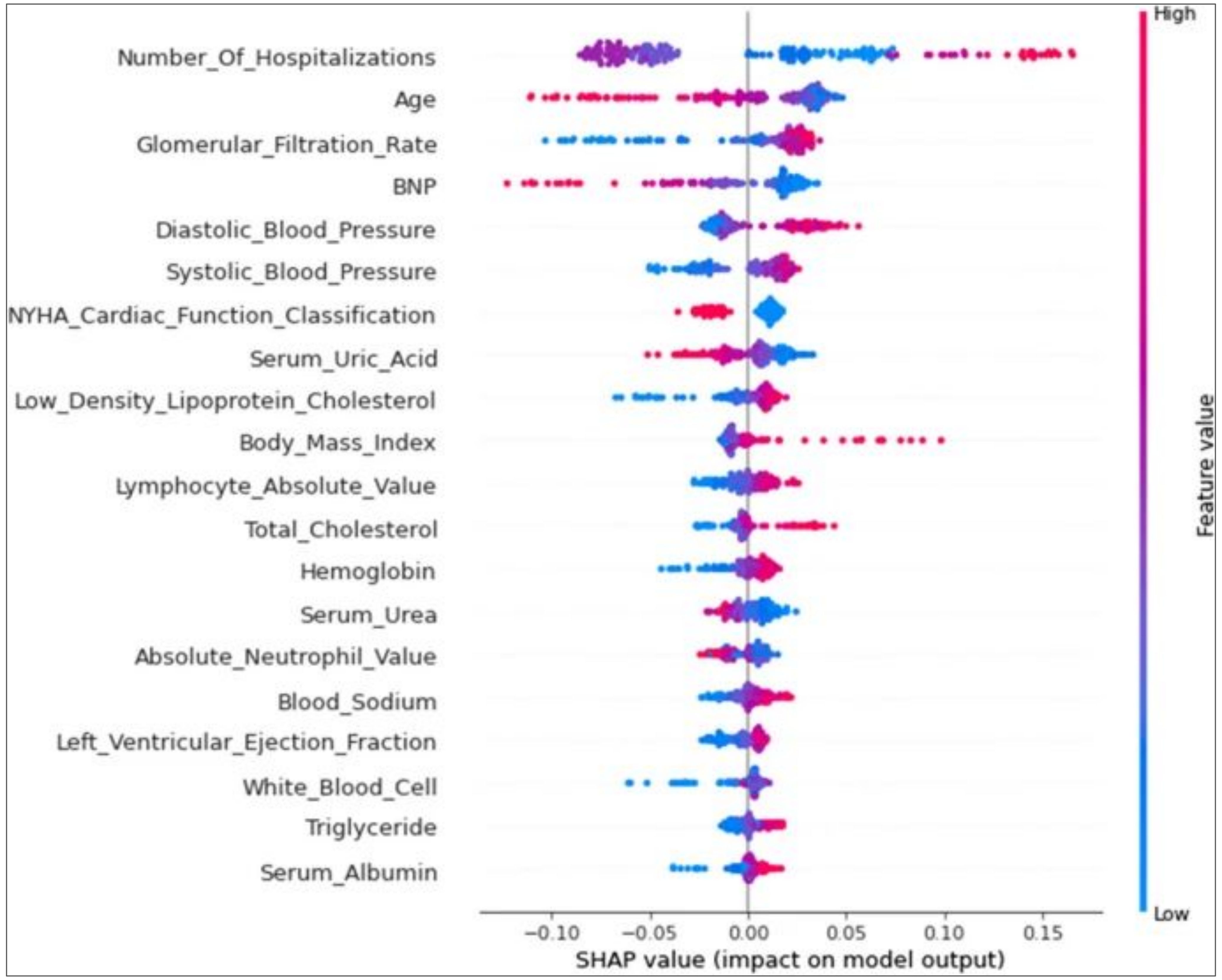
LIME vs SHAP

LIME	SHAP
Model-agnostic	Model-specific
Local explanation	Global explanation
Kernel-based	Game-theoretic approach
Accuracy trade-off	Simplicity trade-off

Healthcare use-case 1

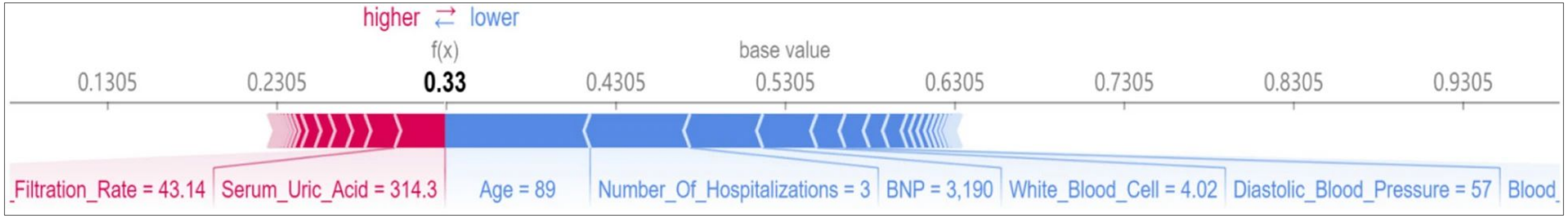
Interpretable prediction of 3-year all-cause mortality in patients with chronic heart failure based on Machine Learning
(https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-023-02371-5)





Healthcare use-case 1 (Continued)

Interpretable prediction of 3-year all-cause mortality in patients with chronic heart failure based on Machine Learning
(https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-023-02371-5)



Healthcare use-case 2

Data analysis with Shapley values for automatic subject selection in Alzheimer's disease
(https://alzres.biomedcentral.com/articles/10.1186/s13195-021-00879-4)