GO DATA DRIVEN

Unpacking the Black Box

How to Interpret your Machine Learning Model

About us

Marysia Winkels

Data Science Educator @ GoDataDriven

James Hayward

Data Science Educator @ GoDataDriven

GO DATA DRIVEN

GoDataDriven

GO DATA DRIVEN

Organize

Data and AI Strategy Execution

Build

Data Platforms and AI Applications

Train

Data and AI Learning Journeys

Manage

Managed Data Platform and AI Services

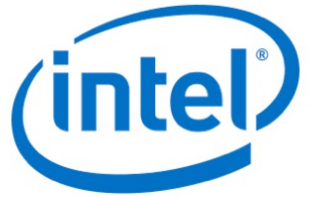
Democratize

Data Democratization

Solve

Turn-Key Cloud & Data Solutions

Trusted by global industry leaders



GO DATA DRIVEN

GoDataDriven Academy

A variety of courses

- Data Science with Python (introductory & advanced)
- Deep Learning
- Data Visualisation & Storytelling
- Analytics Translation
- Time Series
- Machine Learning Explainability

GO DATA DRIVEN

About you!

Join us at PollEverywhere

gdd.li/poll-amld

GO DATA DRIVEN

Workshop outline

Topics are covered in the following way:

1. Experiment *first*
2. Explain *later*

Put it in practice:

- Hackathon
- Lightning talks!

GO DATA DRIVEN

Workshop content

What we will cover:

- Intrinsicly explainable machine learning models
- Methods for model-agnostic interpretability
- Advantages and disadvantages of each method

What this workshop is **not** about:

- The theory behind every machine learning model (*algorithmic transparency)*
- Shapley values
- Neural networks

GO DATA DRIVEN

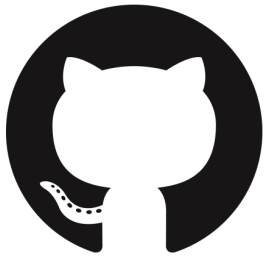
Material

Two ways to access the material:

1. **Binder**

   Does not require any installation

2. Download or clone the material from our **GitHub** repository



gdd.li/github

gdd.li/binder

GO DATA DRIVEN

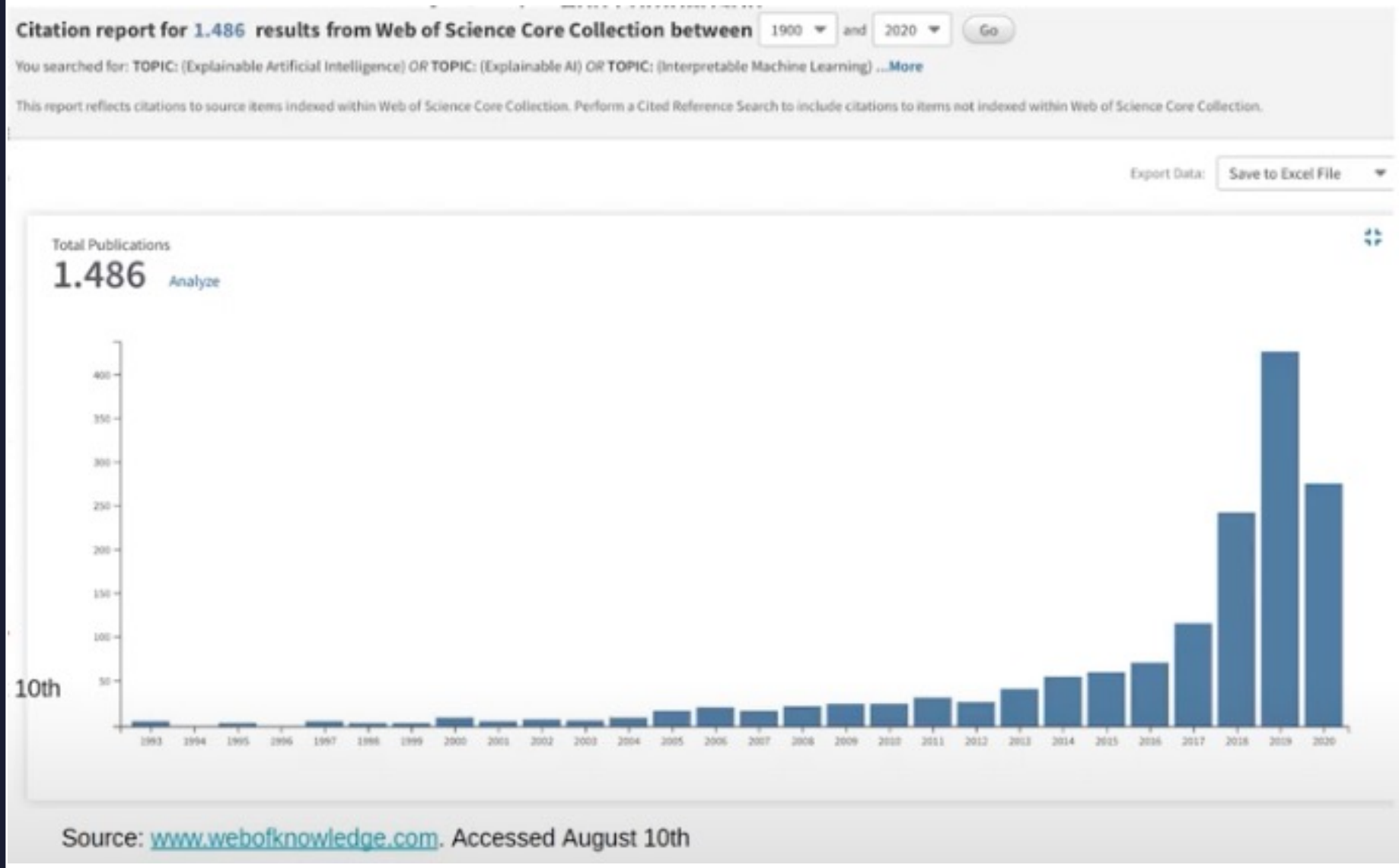
Unpacking the black box

GO DATA DRIVEN

History

- 1800s – Linear Regression
- 1960s – Decision Tree
- 2001 – Random Forests
- 2012 – Deep Learning
- 2016 – *Interpretable Machine Learning*

GO DATA DRIVEN

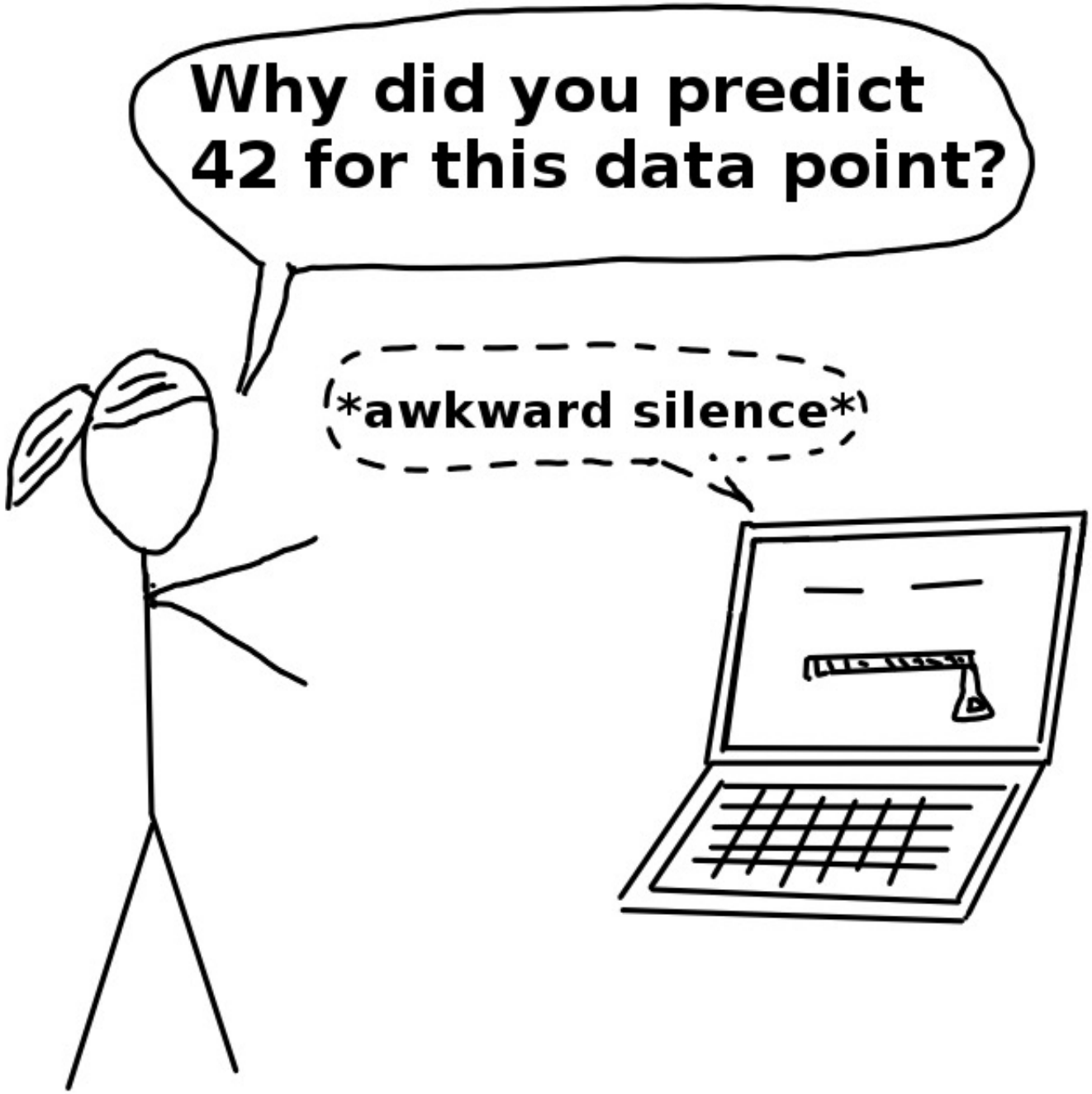


GO DATA DRIVEN

"Interpretability is the degree to which a human can understand the cause of a decision"

Source: Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences.

Why of Interpretability



GO DATA DRIVEN

Why of Interpretability

- Engender trust

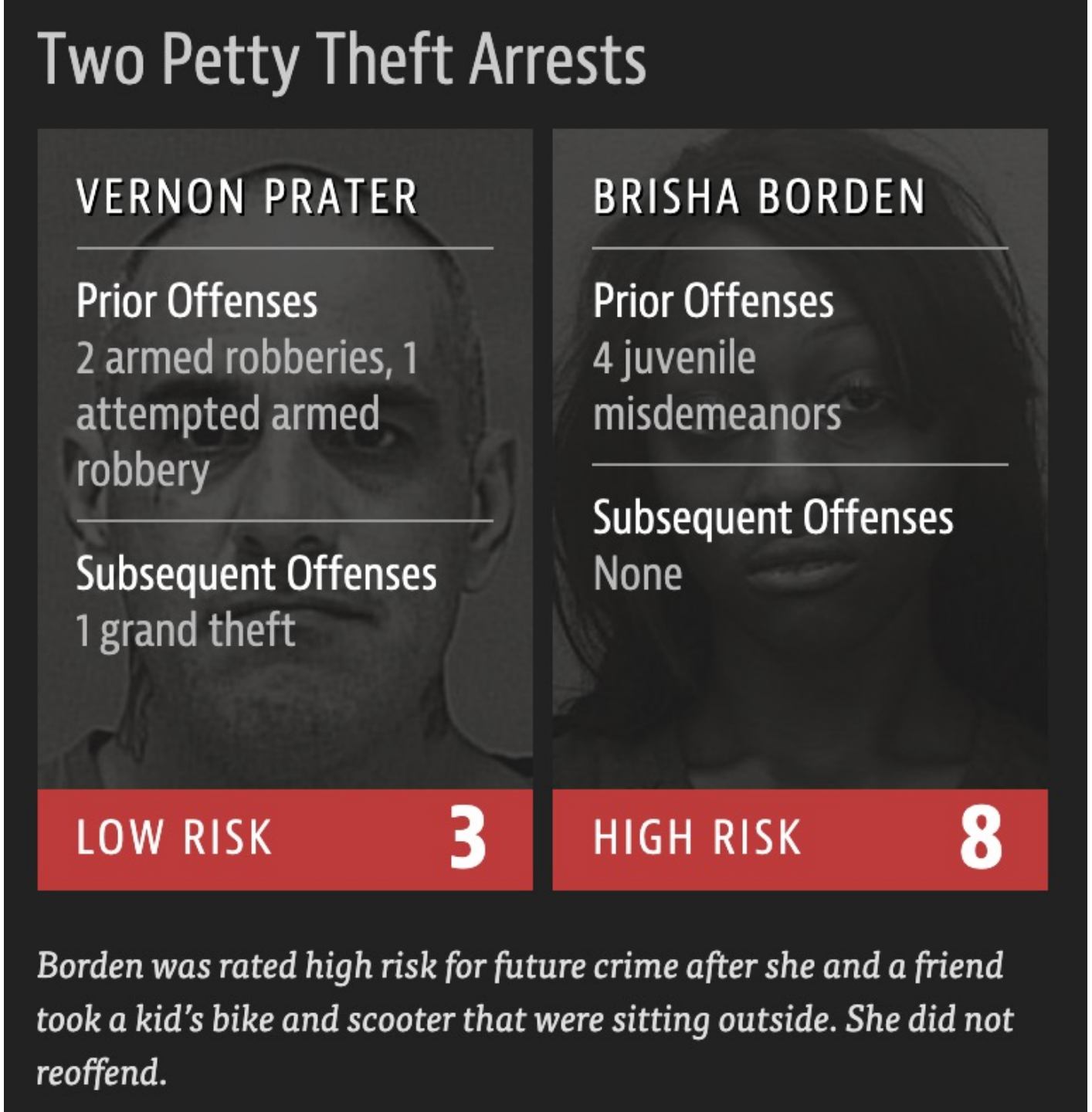
GO DATA DRIVEN

Why of Interpretability

- Engender trust
- Right to explanation

Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

GO DATA DRIVEN



Why of Interpretability

- Engender trust
- Right to explanation
- **Debug models**

(a) Husky classified as wolf

(b) Explanation

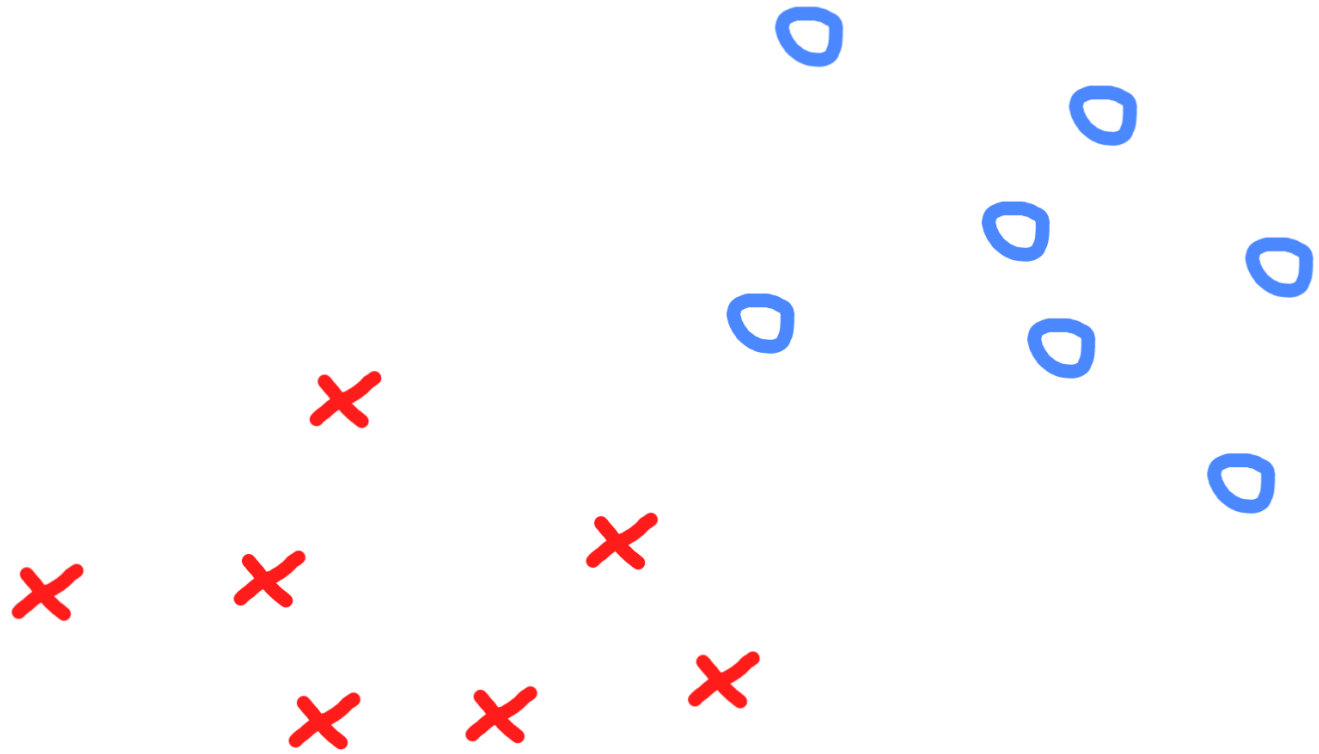
GO DATA DRIVEN

GO DATA DRIVEN

"Thing B is similar to thing A,
and thing A caused Y,
so I predict that B will cause Y!"

Example-based explanations

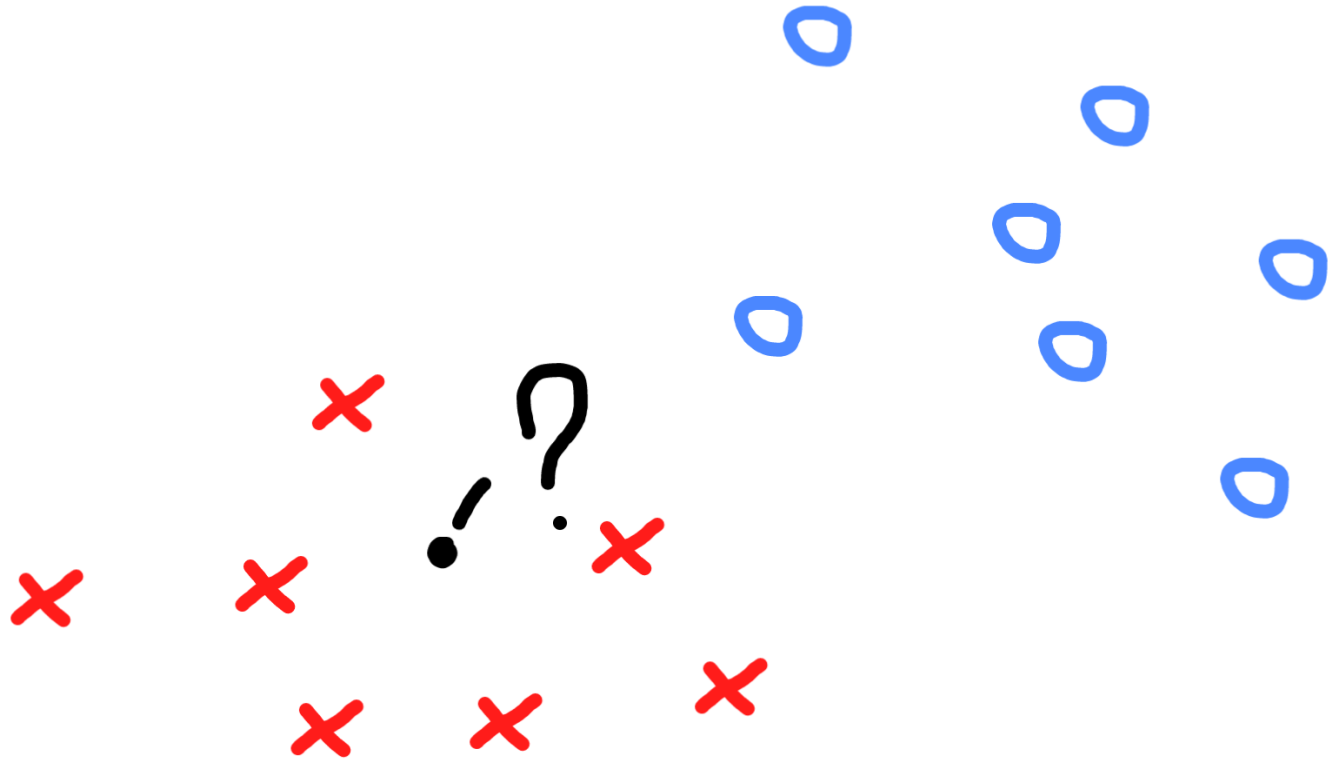
1. Some dataset with two classes



GO DATA DRIVEN

Example-based explanations

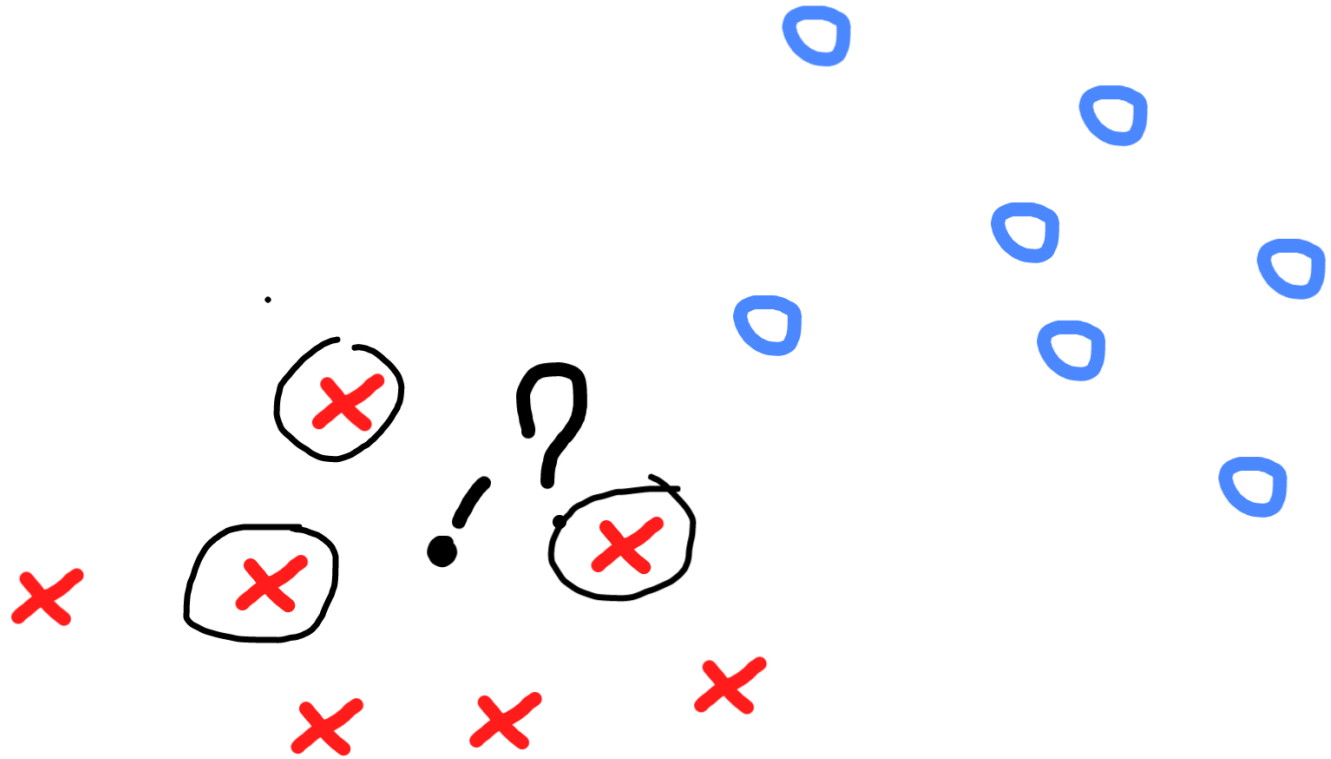
1. Some dataset with two classes
2. A new data point to classify



GO DATA DRIVEN

Example-based explanations

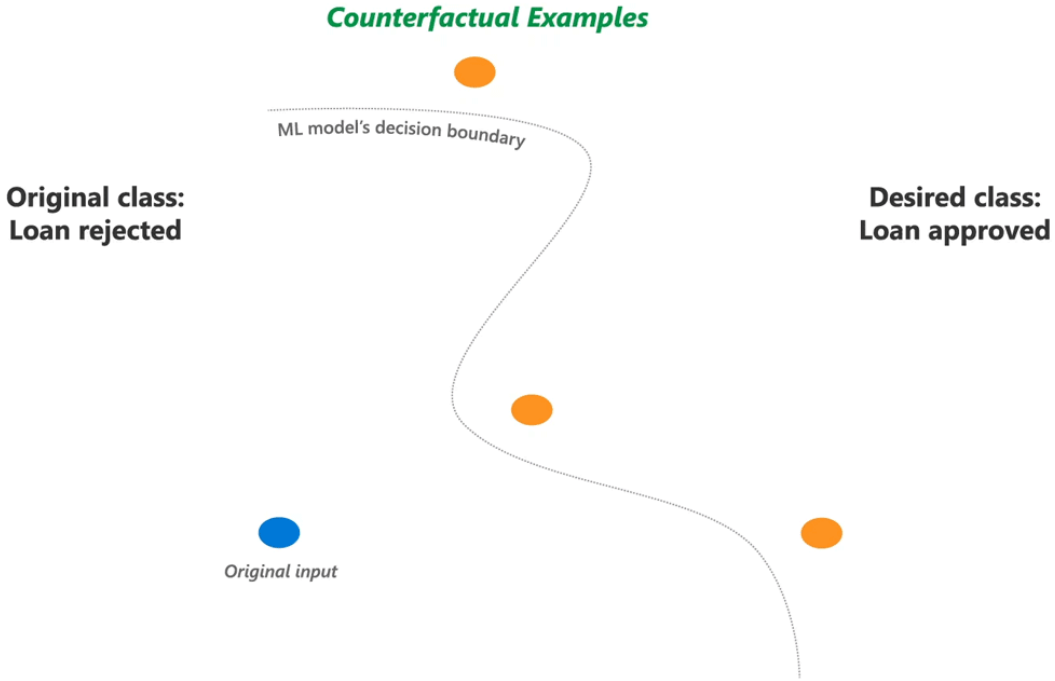
1. Some dataset with two classes
2. A new data point to classify
3. **An explanation based on nearby data points**



GO DATA DRIVEN

Example-based explanations

- Some models
- Counterfactual explanations
- Adversarial examples
- Prototypes & Criticisms
- Influential instances



GO DATA DRIVEN

Jupyter Notebook 1

01_Introduction.ipynb

GO DATA DRIVEN

The inherent interpretability of ML models

GO DATA DRIVEN

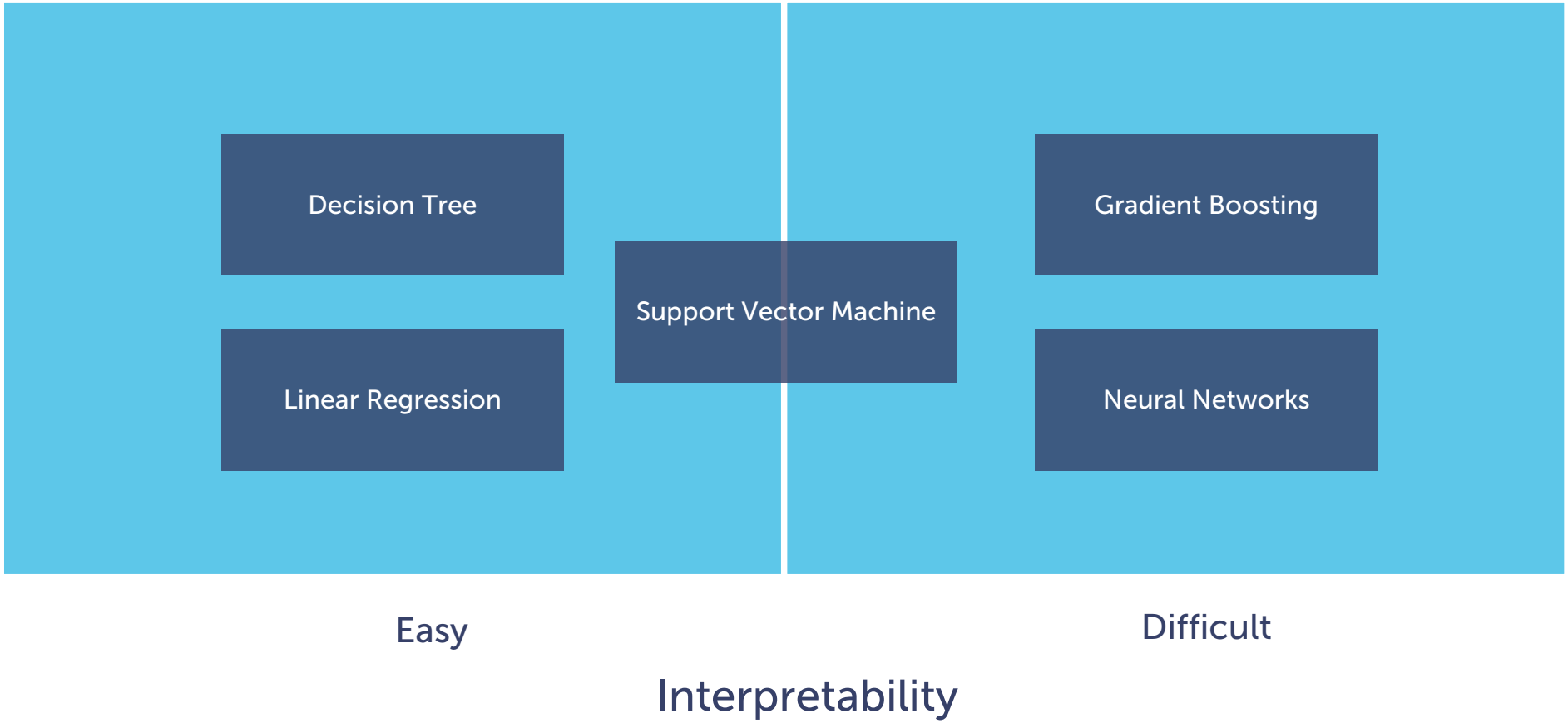
Penguin 218

- Species: Chinstrap
- Flipper length: 210mm
- Body mass: 4100g

MISCLASSIFED AS GENTOO

GO DATA DRIVEN

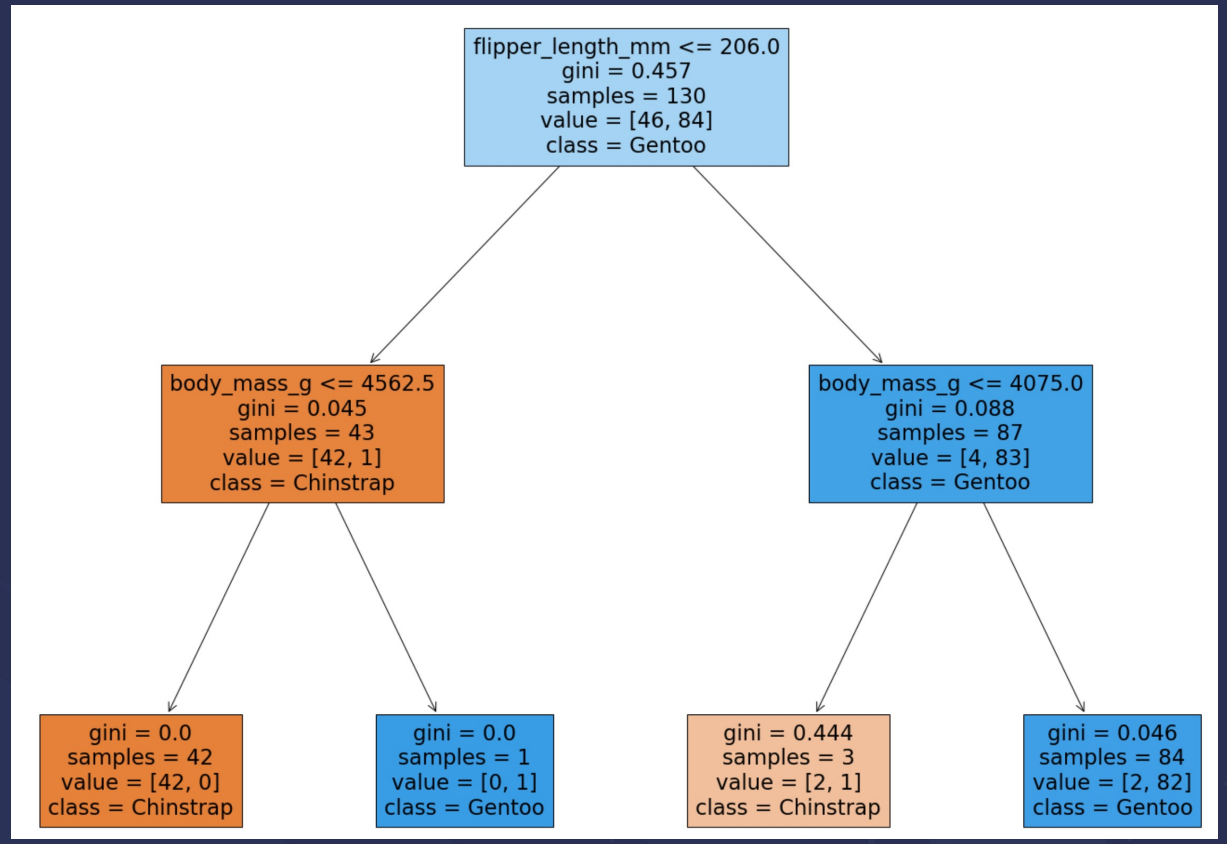
Inherent model interpretability



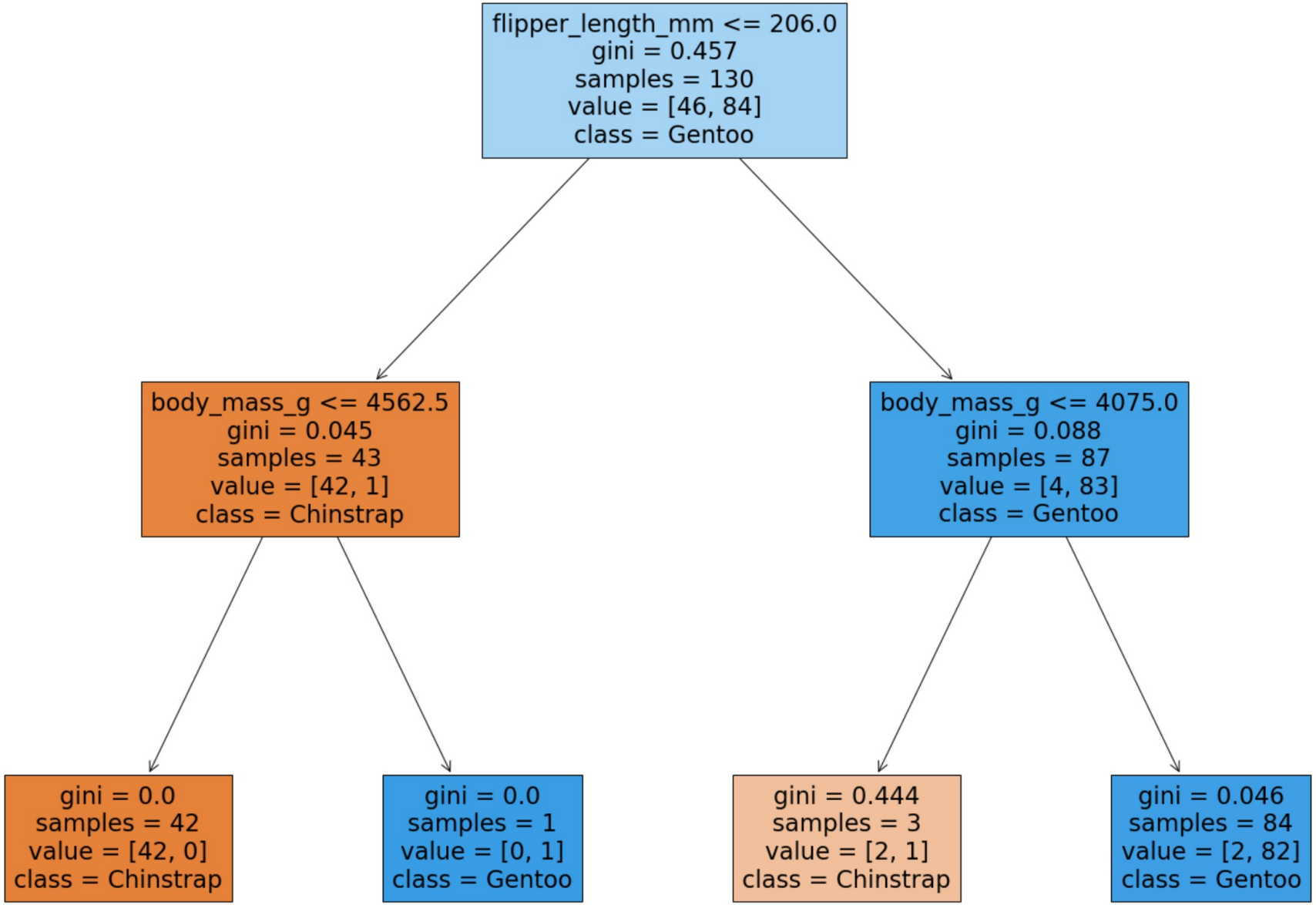
Decision Tree Classifer

GO DATA DRIVEN

The inherent interpretability of ML models



Decision Tree Classifer

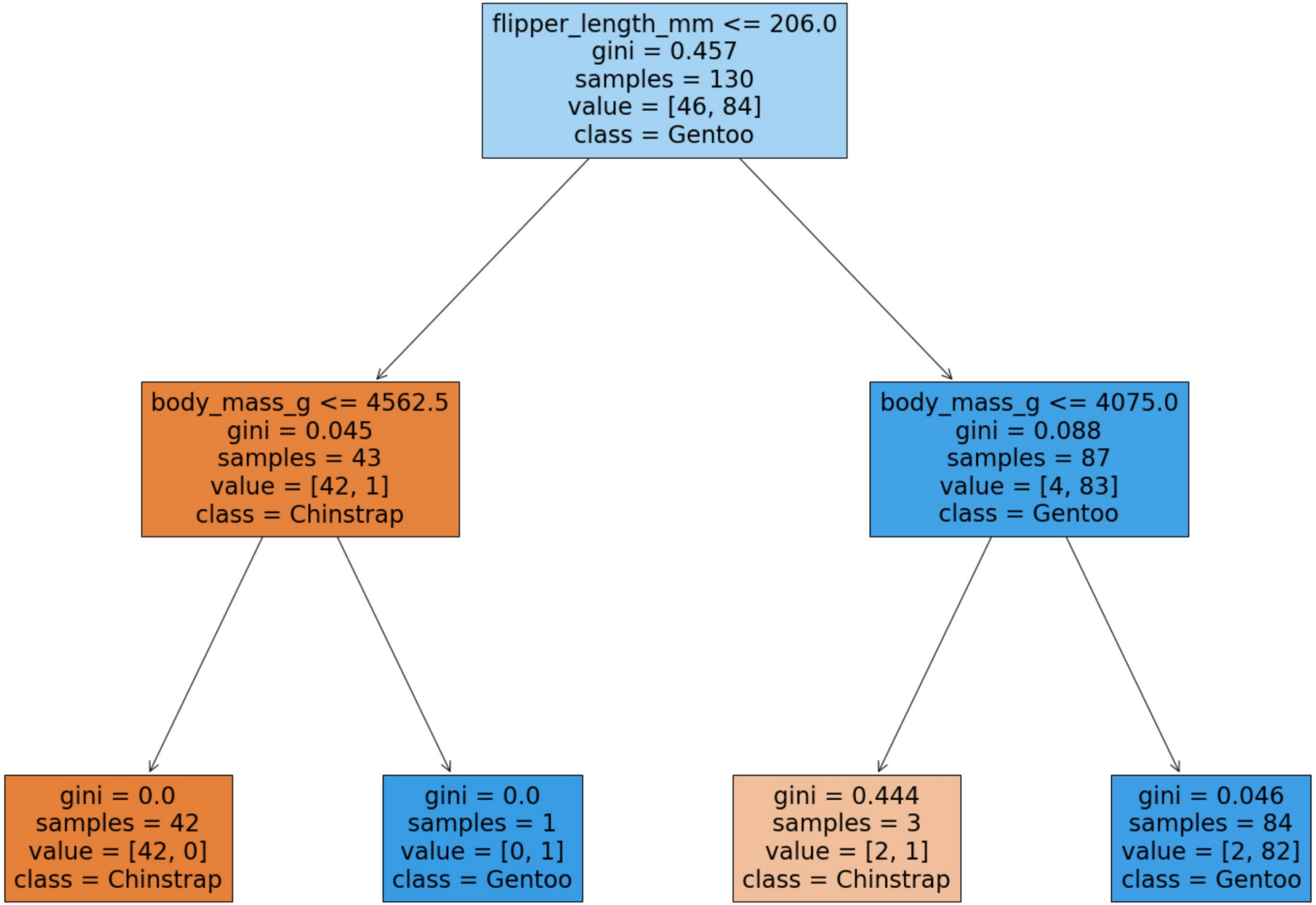


GO DATA DRIVEN

Decision Tree Classifer

GO DATA DRIVEN

Decision Tree Classifer



GO DATA DRIVEN

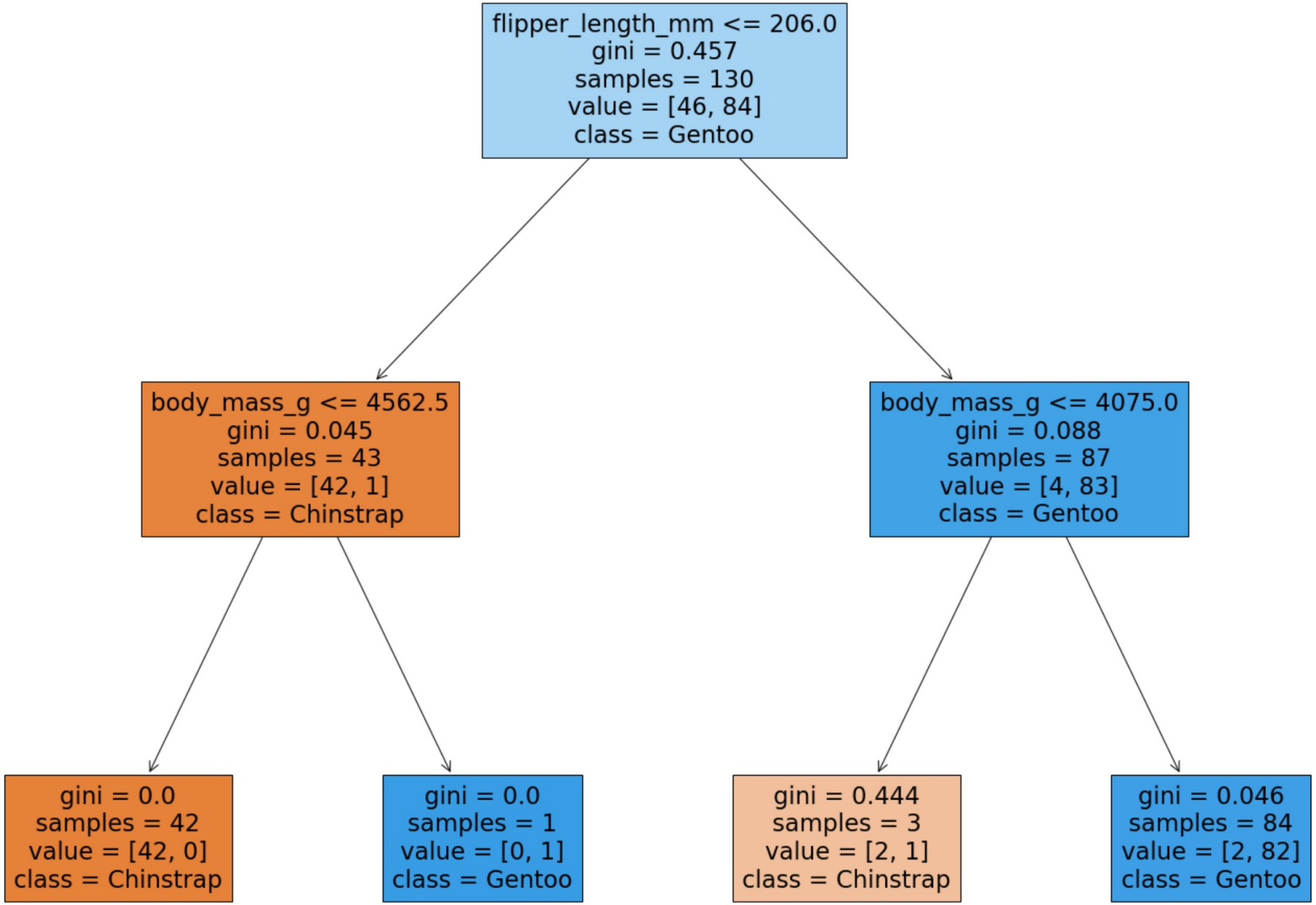
Decision Tree Classifer

gini = 0.457
samples = 130
value = [46, 84]
class = Gentoo

$$\text{gini} = 1 - \frac{46}{130}^2 - \frac{84}{130}^2 = 0.457\ldots$$

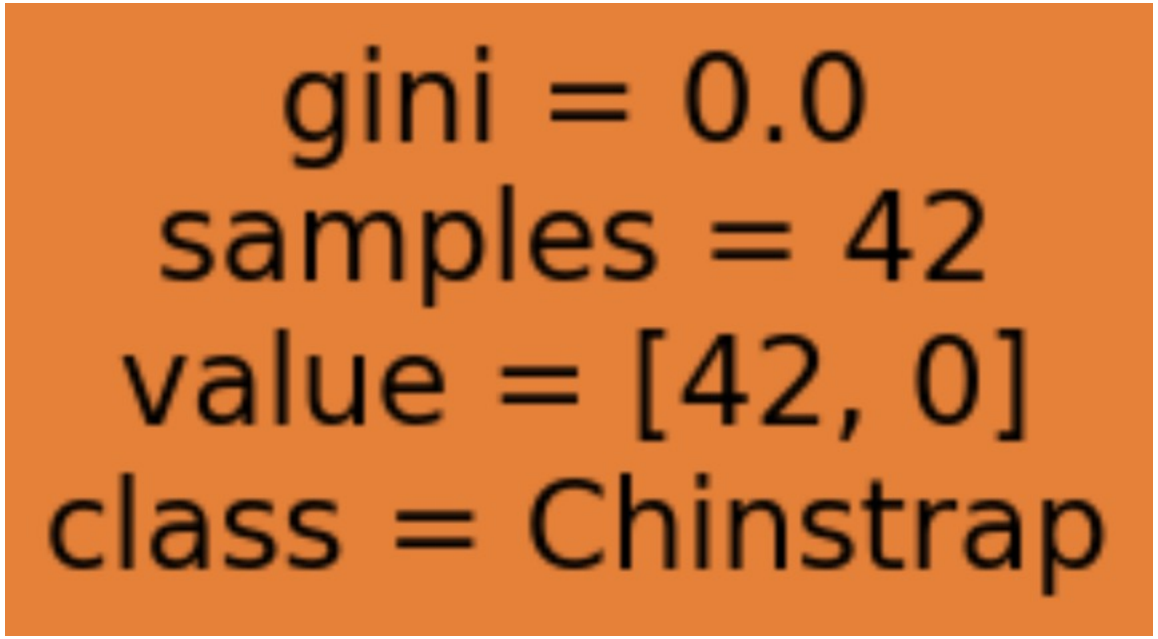
GO DATA DRIVEN

Decision Tree Classifer



GO DATA DRIVEN

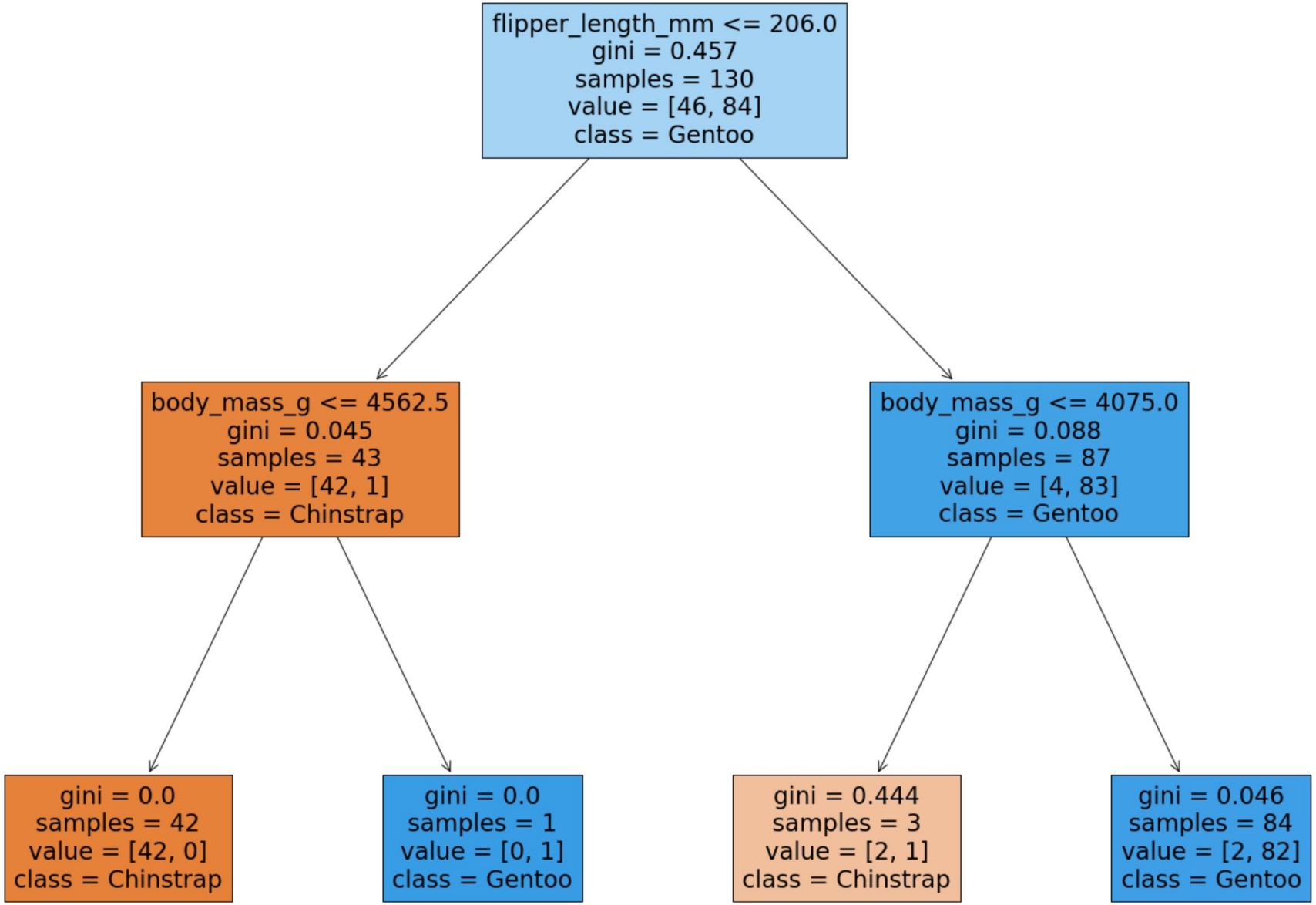
Decision Tree Classifer



$$\text{gini} = 1 - \frac{42}{42}^2 - \frac{0}{42}^2 = 1 - 1 - 0 = 1$$

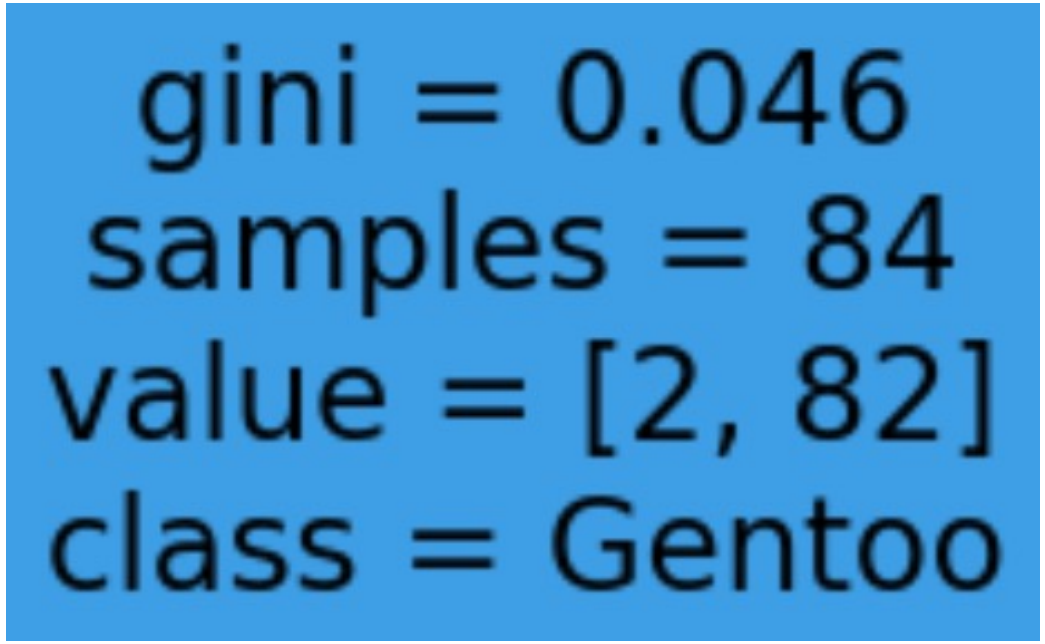
GO DATA DRIVEN

Decision Tree Classifer



GO DATA DRIVEN

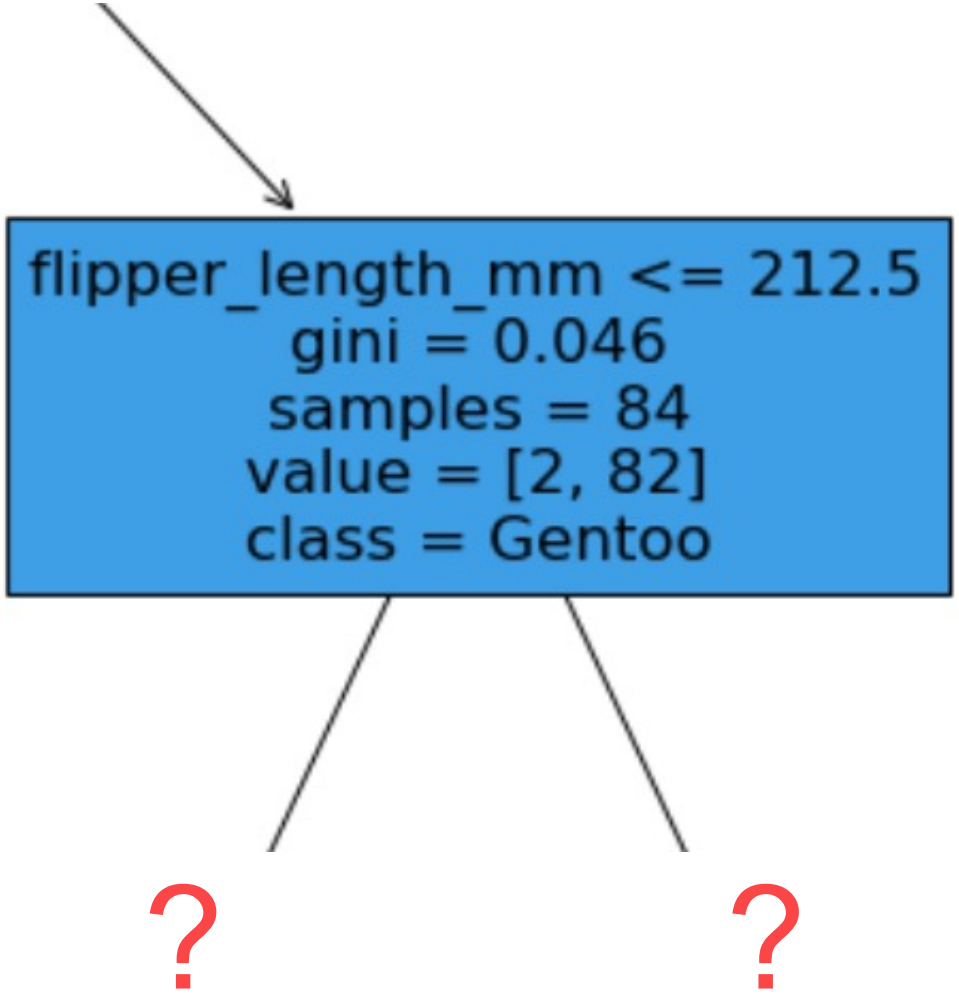
Decision Tree Classifer



$$\text{gini} = 1 - \frac{2}{84}^2 - \frac{82}{84}^2 = 0.046$$

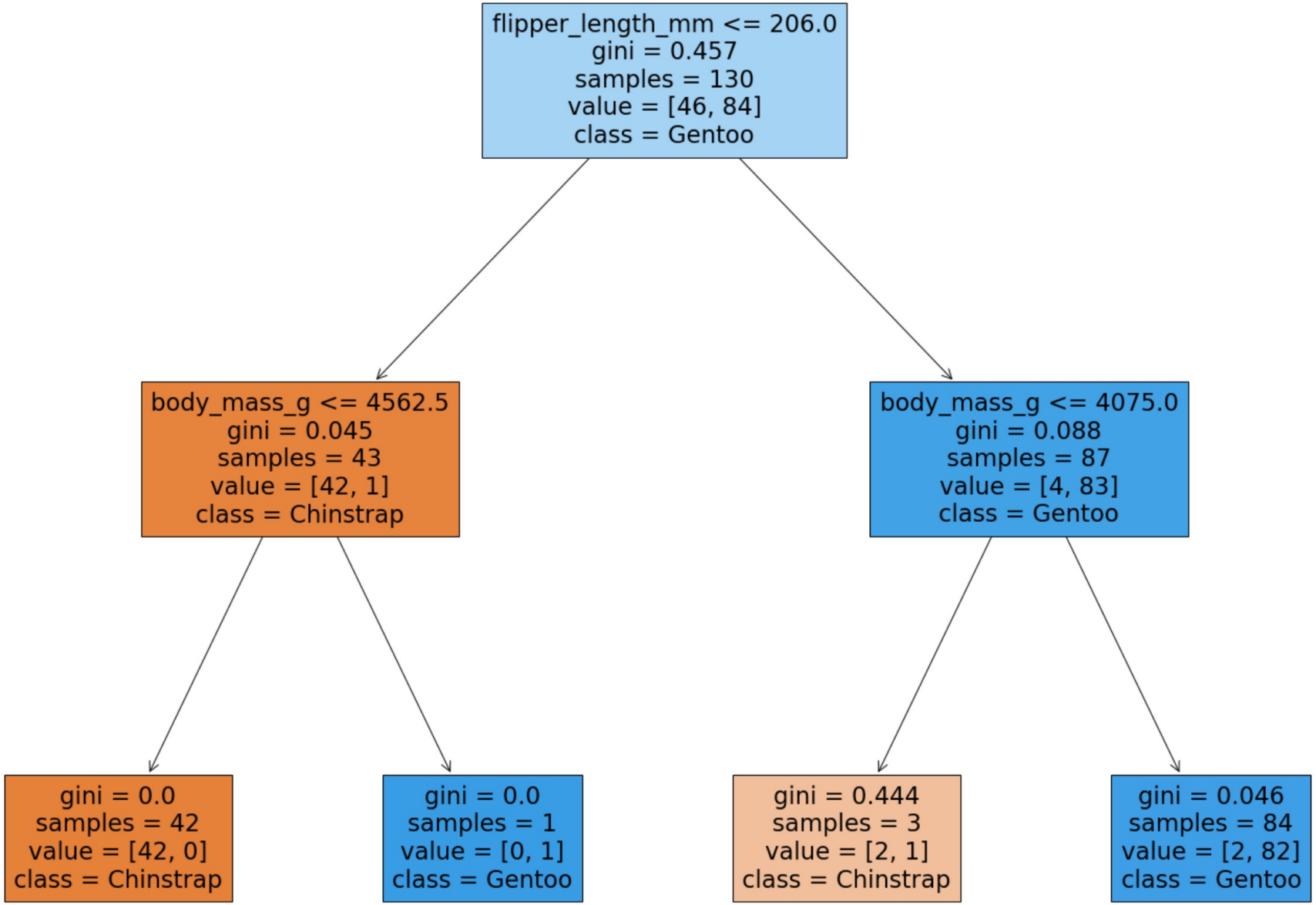
GO DATA DRIVEN

Decision Tree Classifer



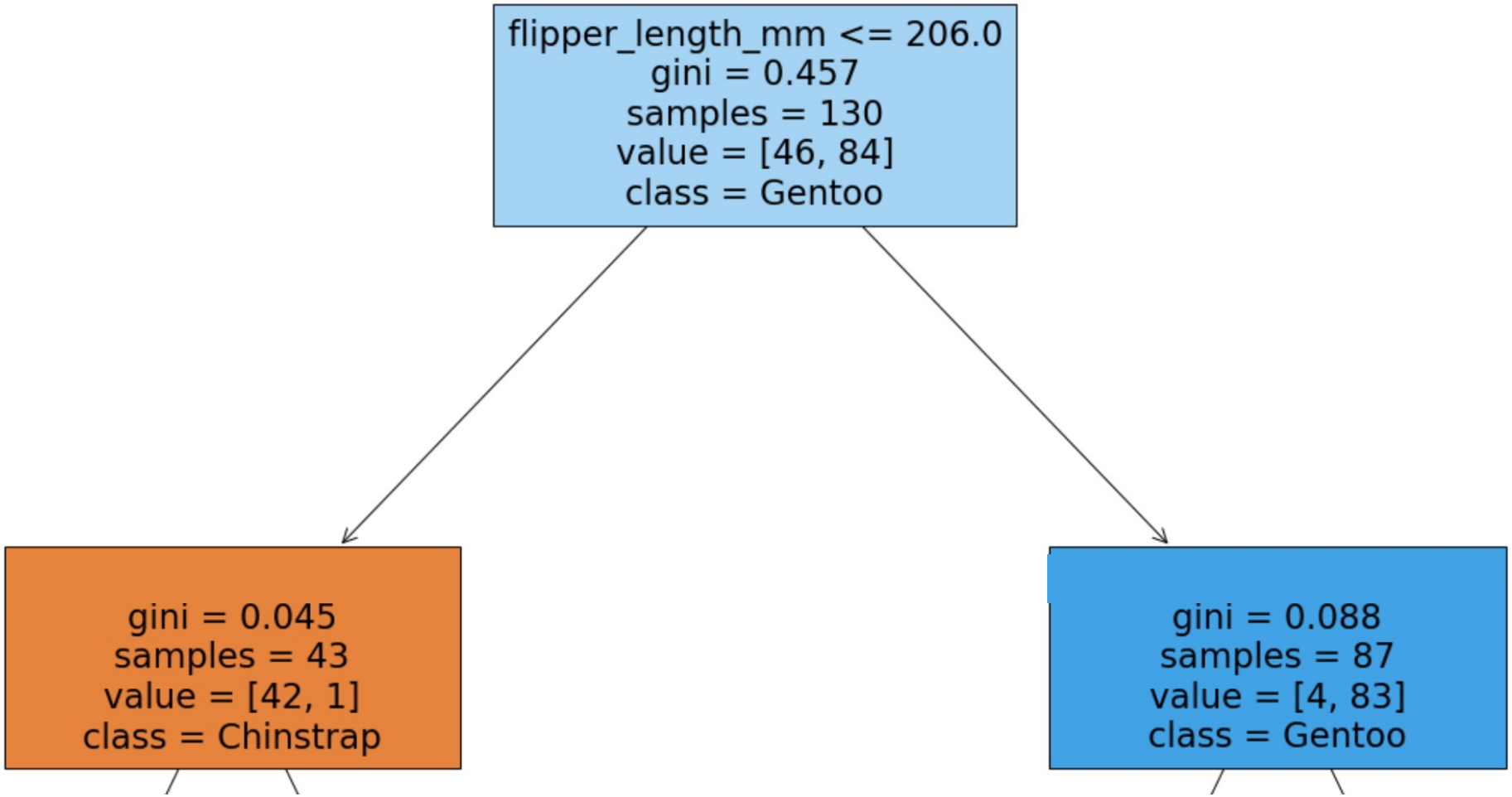
GO DATA DRIVEN

Decision Tree Classifer



GO DATA DRIVEN

Decision Tree Classifer



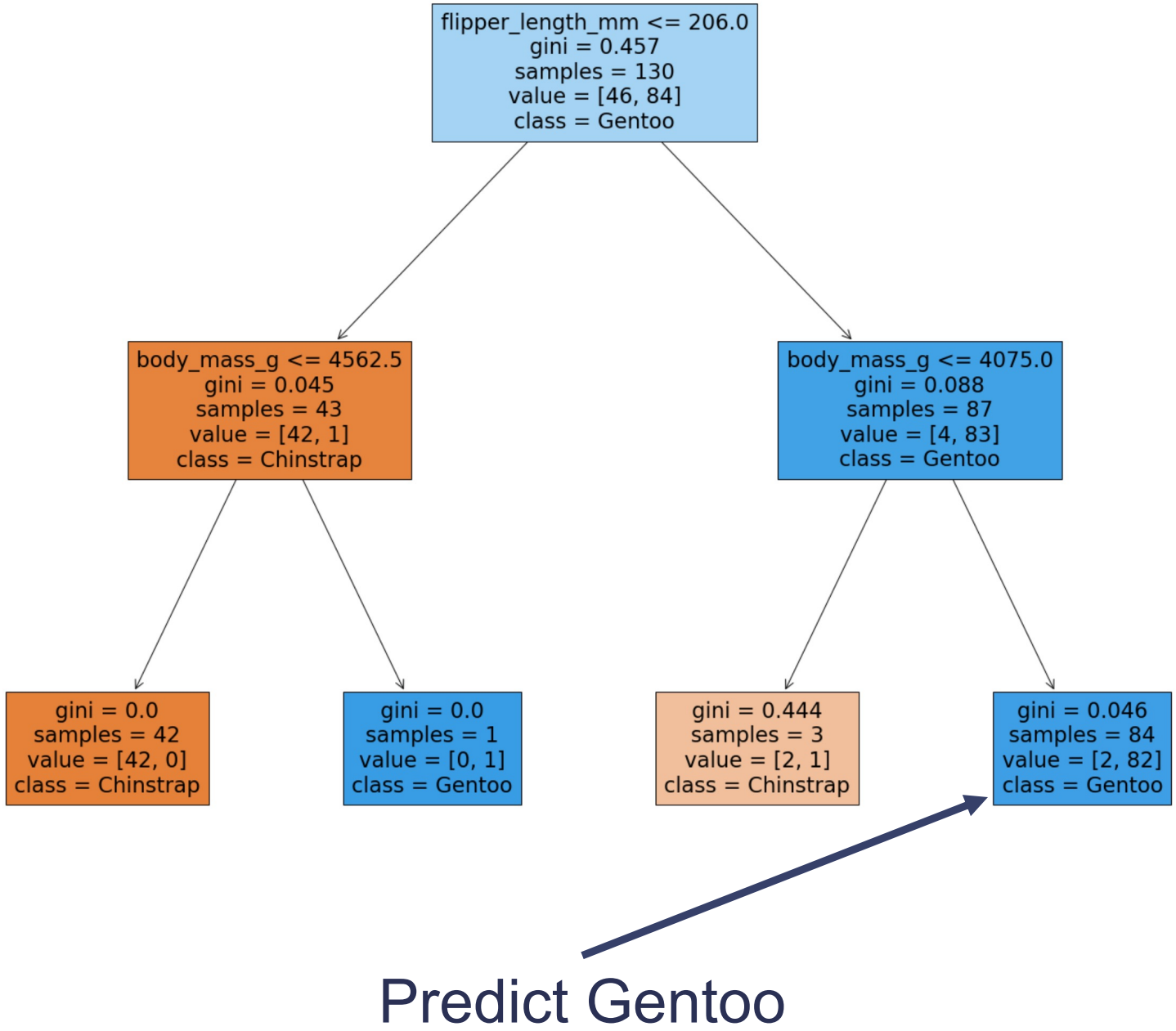
GO DATA DRIVEN

$$0.045 \times \frac{43}{130} + 0.088 \times \frac{87}{130} = 0.074$$

Penguin 218

- Flipper length: 210mm
- Body mass: 4100g

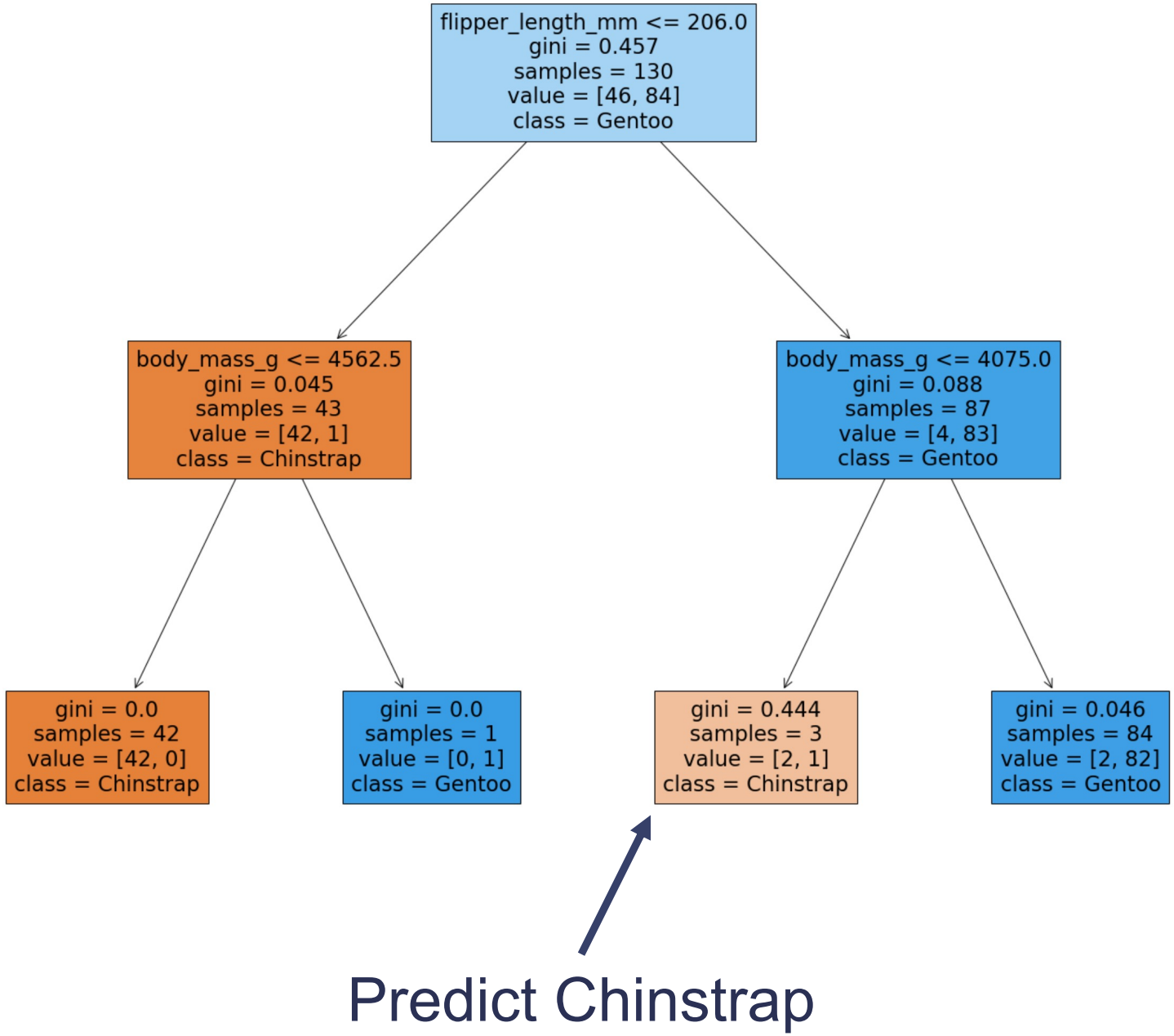
GO DATA DRIVEN



Penguin 218

- Flipper length: 210mm
- Body mass: 4070g

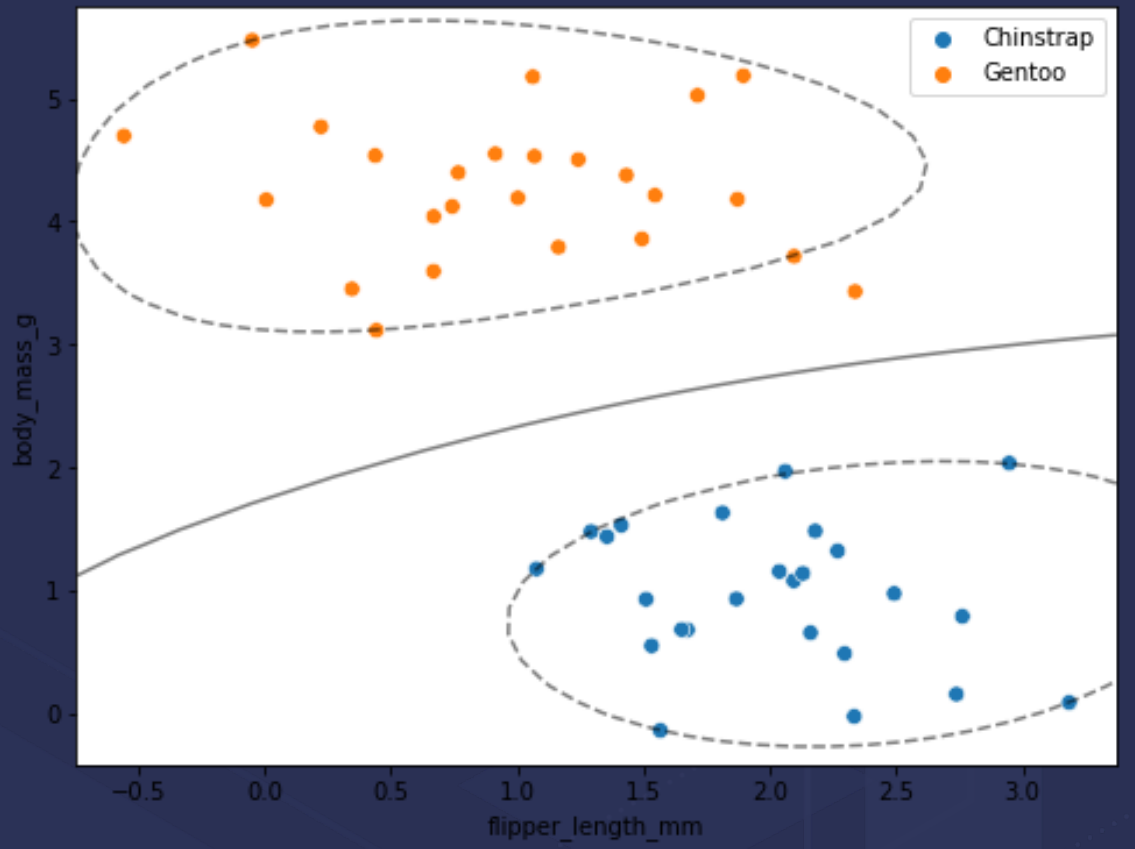
GO DATA DRIVEN



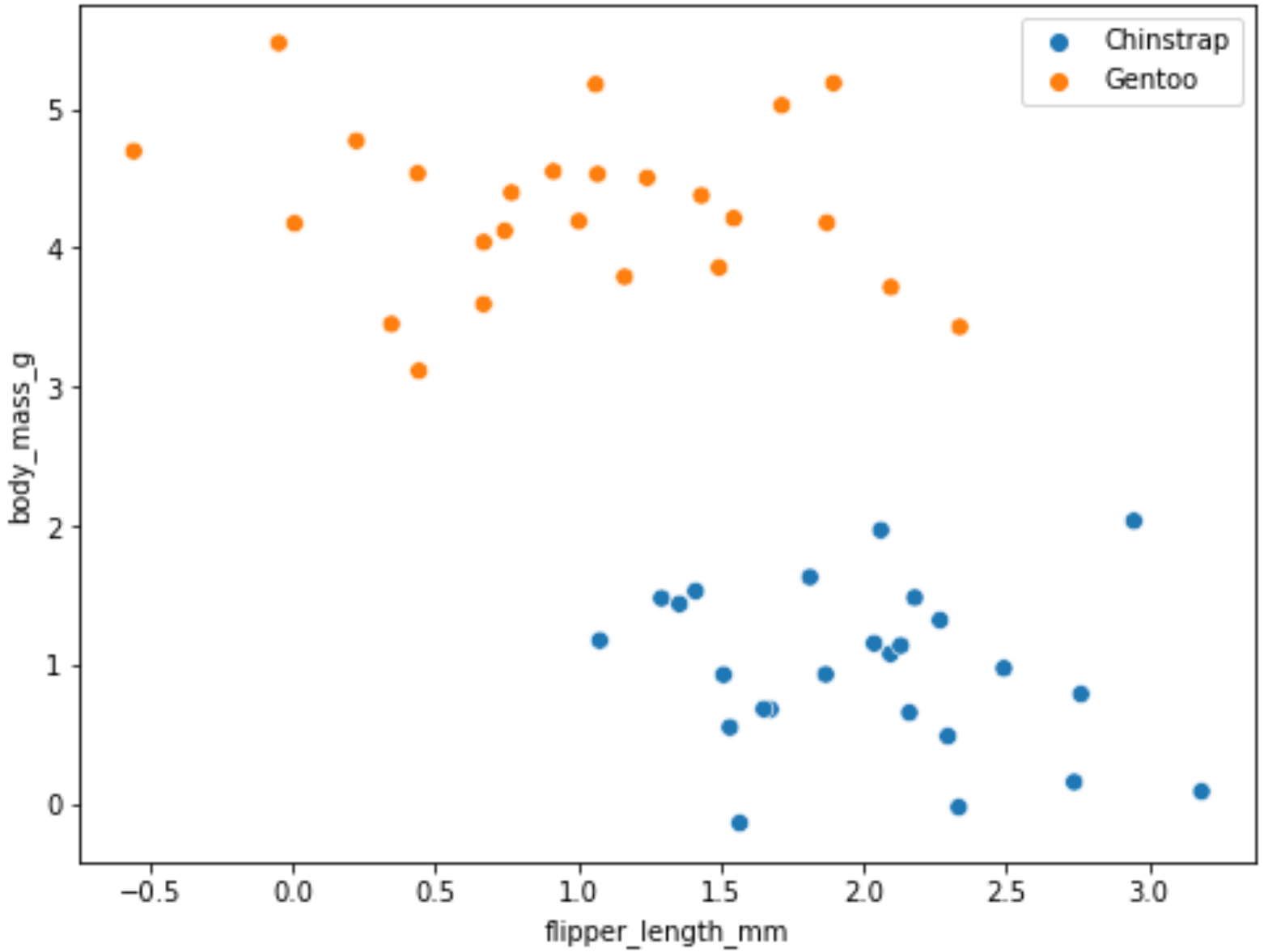
Support Vector Machine Classifer

GO DATA DRIVEN

The inherent interpretability of ML models

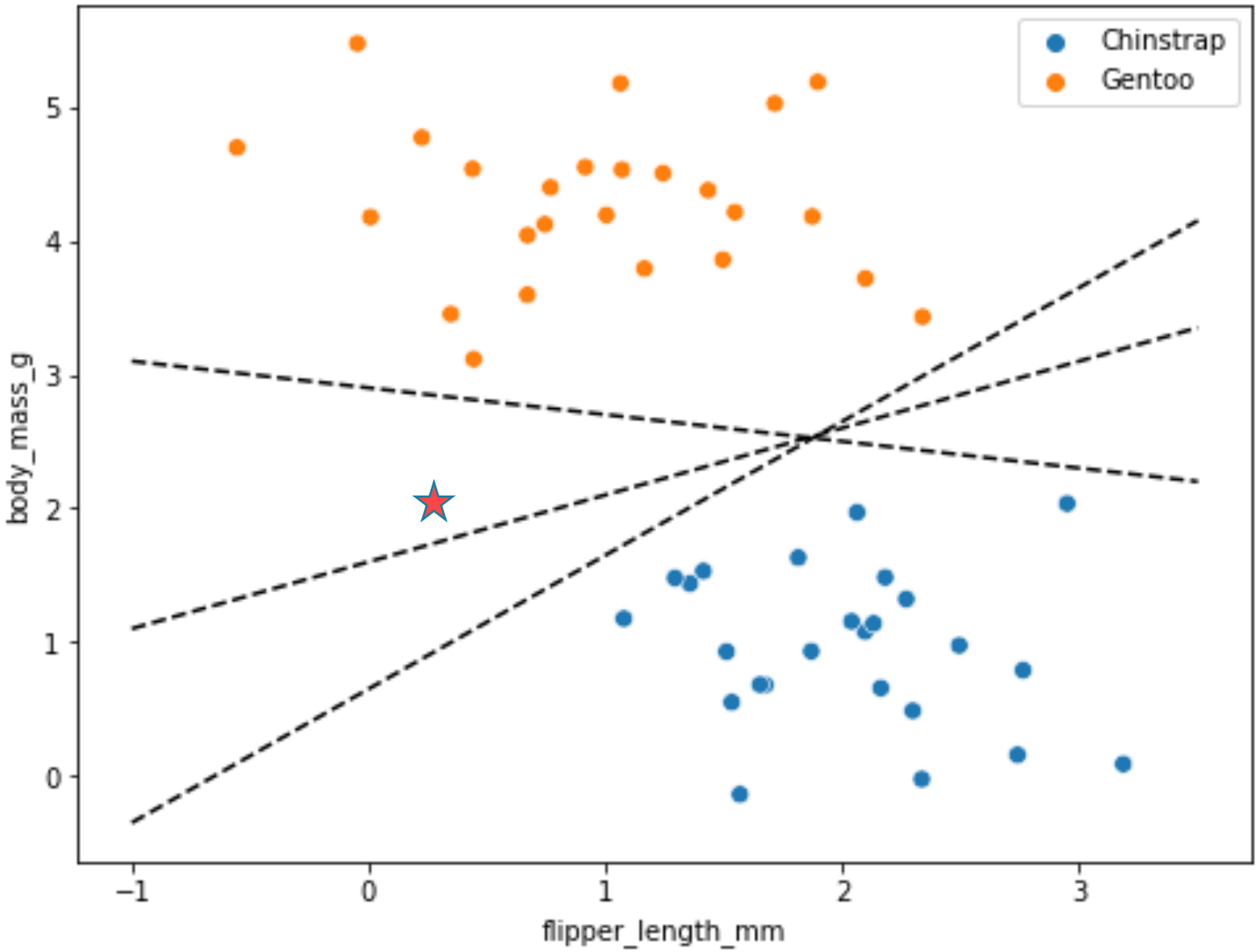


Support Vector Machine Classifier



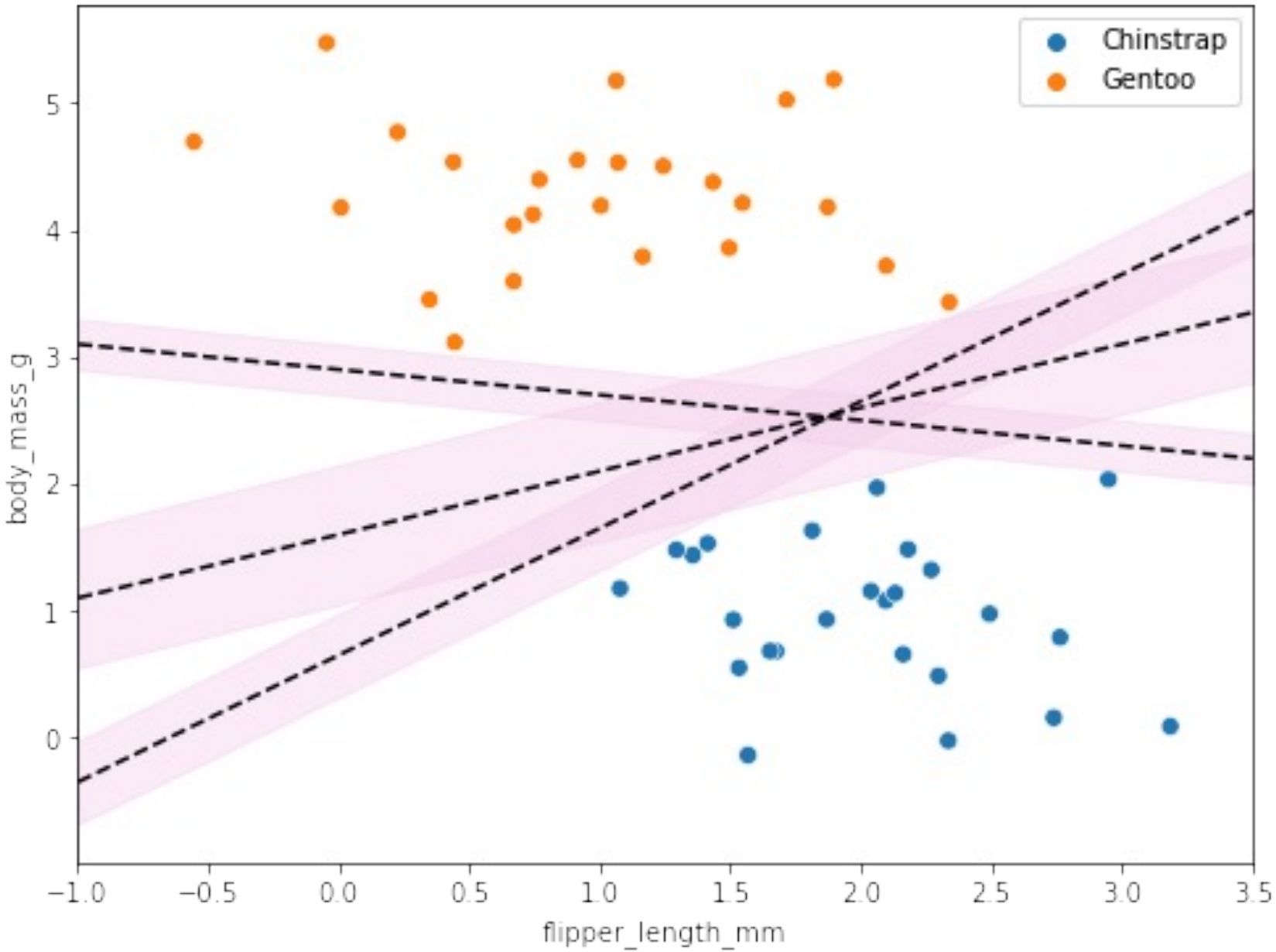
GO DATA DRIVEN

Support Vector Machine Classifier



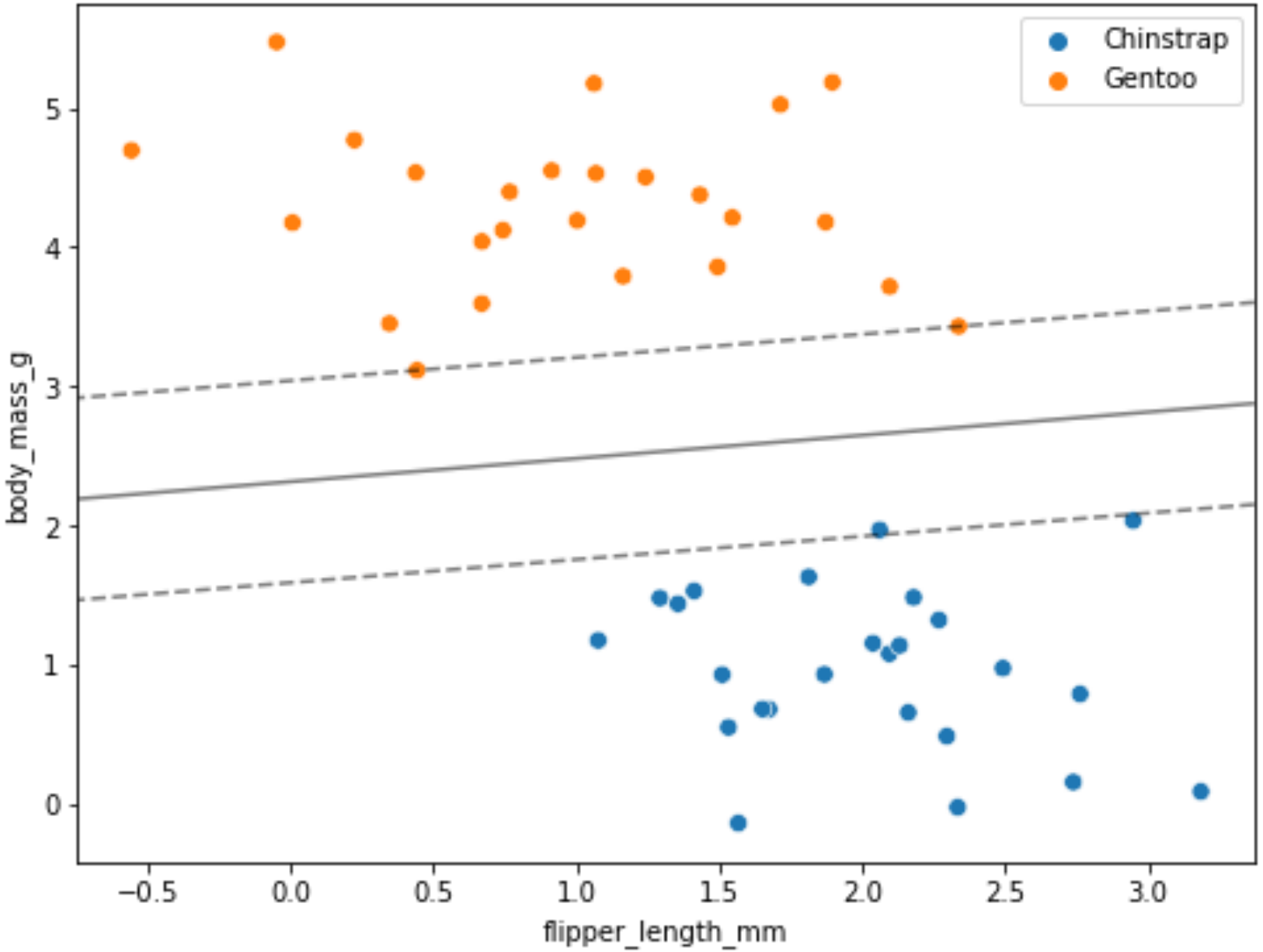
GO DATA DRIVEN

Support Vector Machine Classifier



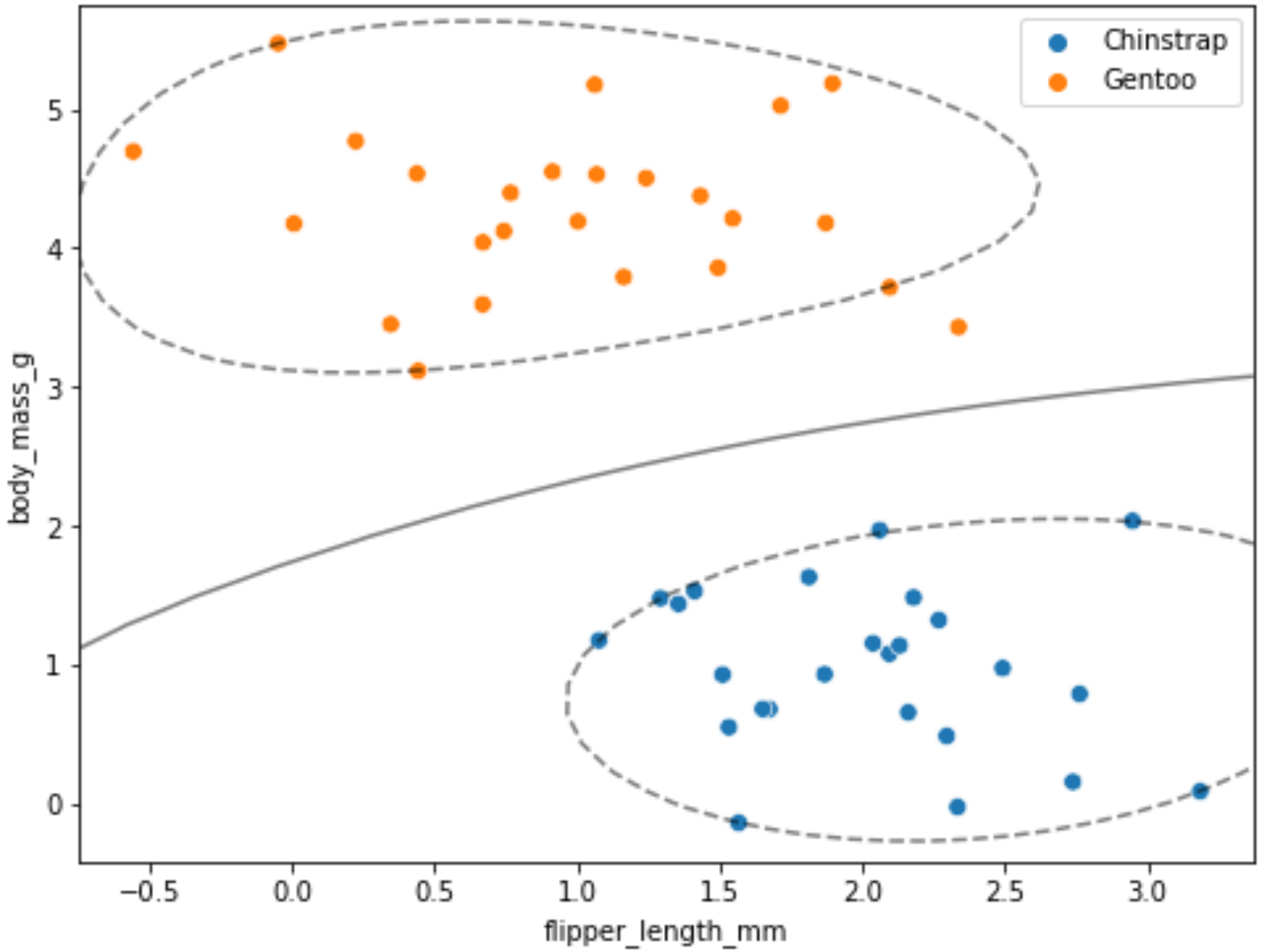
GO DATA DRIVEN

Support Vector Machine Classifier



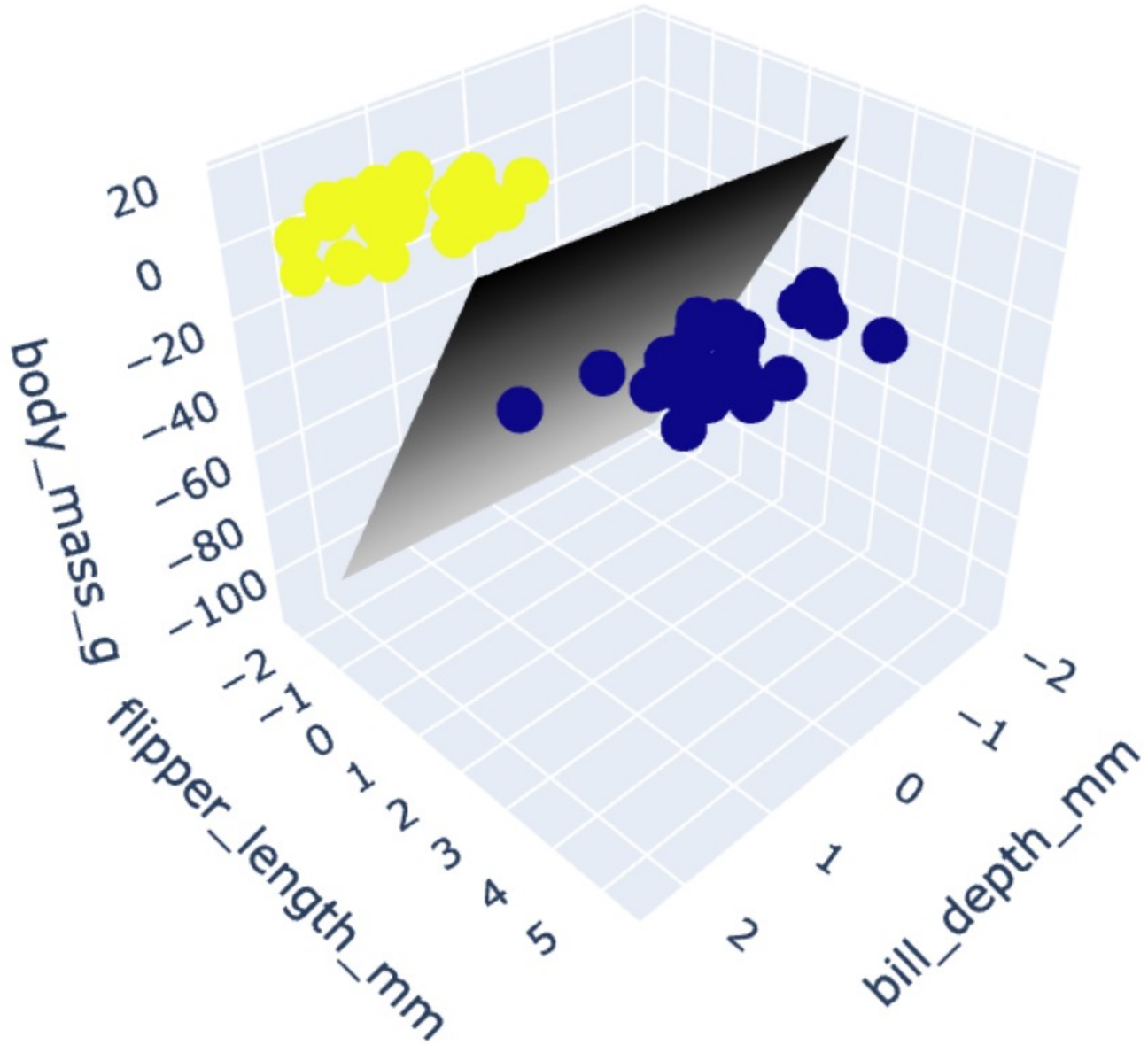
GO DATA DRIVEN

Support Vector Machine Classifier



GO DATA DRIVEN

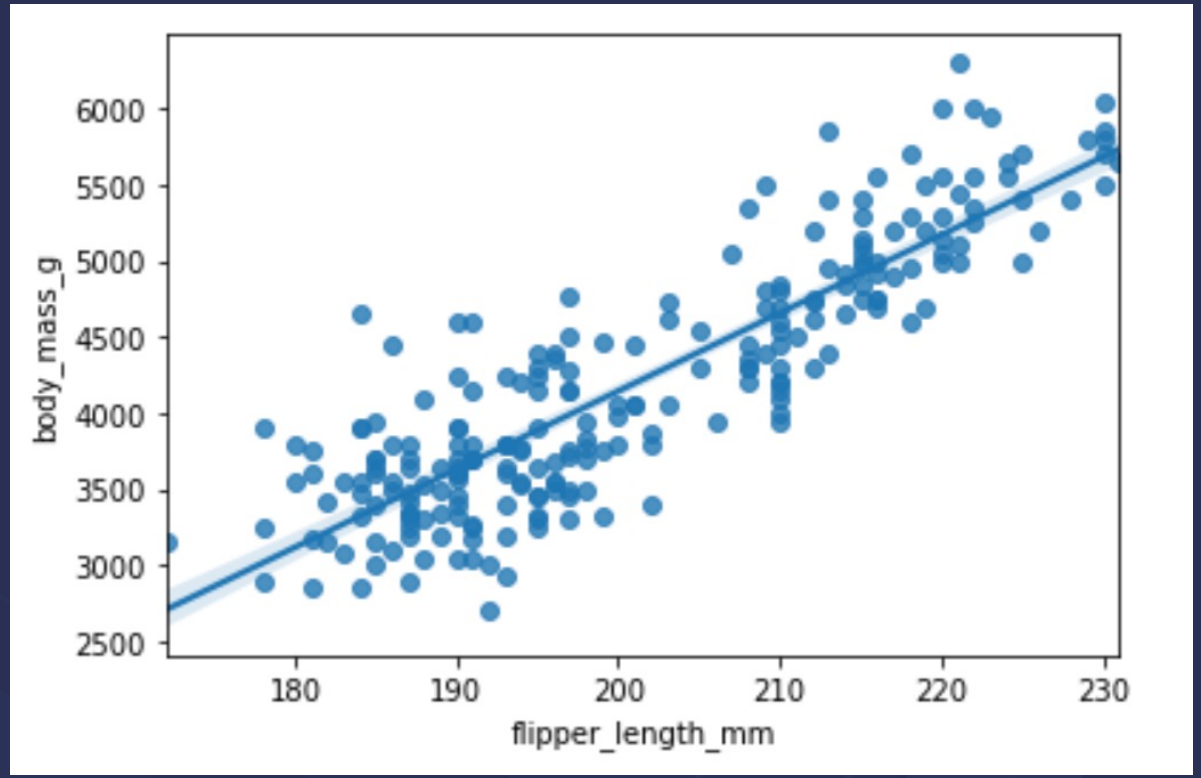
Support Vector Machine Classifier



GO DATA DRIVEN

Linear Regression

The inherent interpretability of ML models



Classification

bill_depth_mm	flipper_length_mm	species
18.7	181.0	Adelie
17.4	186.0	Adelie
18.0	195.0	Adelie
19.3	193.0	Adelie
20.6	190.0	Adelie

GO DATA DRIVEN

Regression

bill_depth_mm	flipper_length_mm	body_mass_g
18.7	181.0	3750.0
17.4	186.0	3800.0
18.0	195.0	3250.0
19.3	193.0	3450.0
20.6	190.0	3650.0

GO DATA DRIVEN

Linear Regression

*m*bill depth + n*flipper length + c = body mass*

bill_depth_mm	flipper_length_mm	body_mass_g
18.7	181.0	3750.0
17.4	186.0	3800.0
18.0	195.0	3250.0
19.3	193.0	3450.0
20.6	190.0	3650.0

GO DATA DRIVEN

Linear Regression

*m*bill depth + n*flipper length + c = body mass*

*14*bill depth + 53*flipper length - 6600 = body mass*

bill_depth_mm	flipper_length_mm	body_mass_g
18.7	181.0	3750.0

*14*18.7+ 53*181.0 - 6600 = 3171*

GO DATA DRIVEN

Model-agnostic interpretability methods

GO DATA DRIVEN

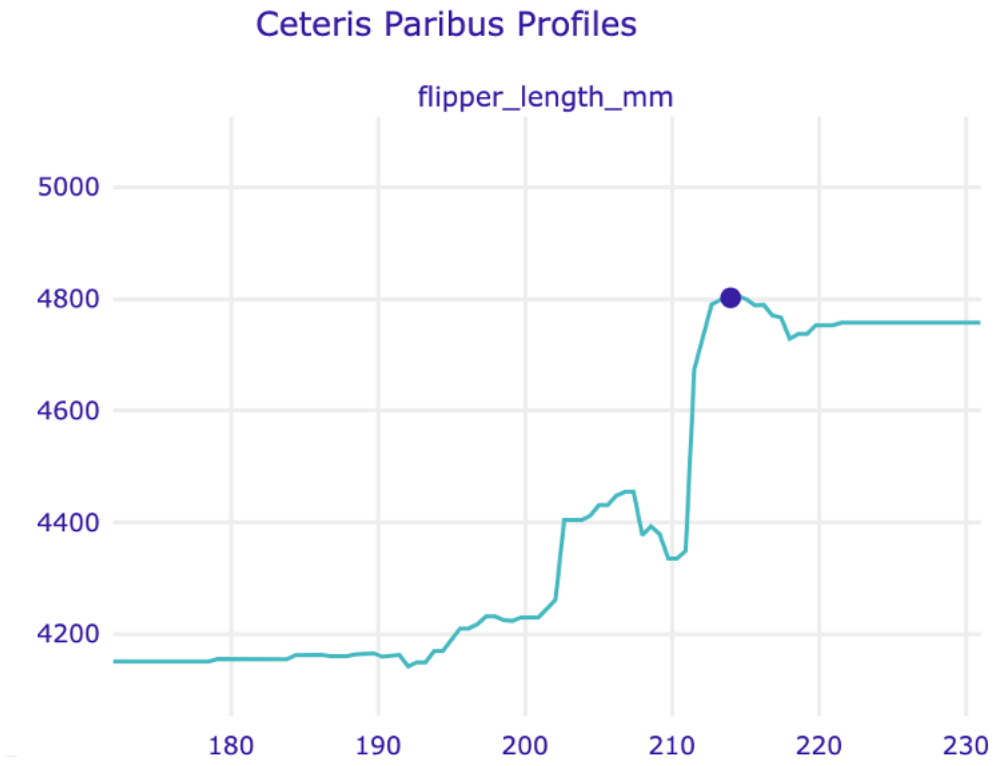
Jupyter Notebook 2

02_Ceteris_Paribus_Plots.ipynb

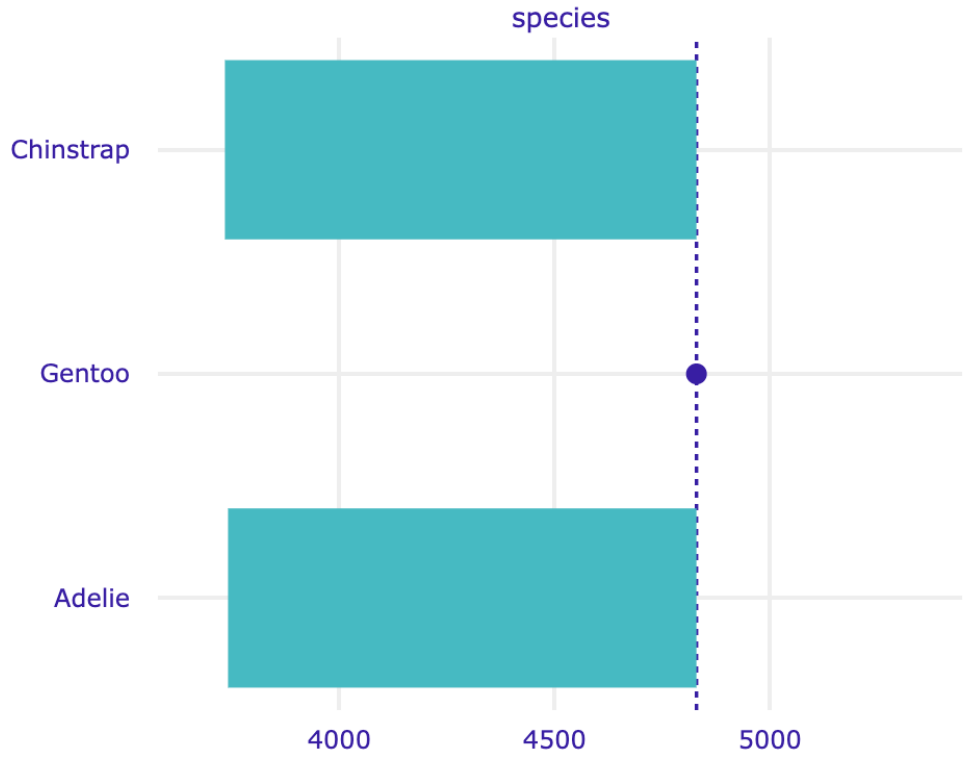
GO DATA DRIVEN

Ceteris Paribus profile

Numerical and categorical features



Numerical features

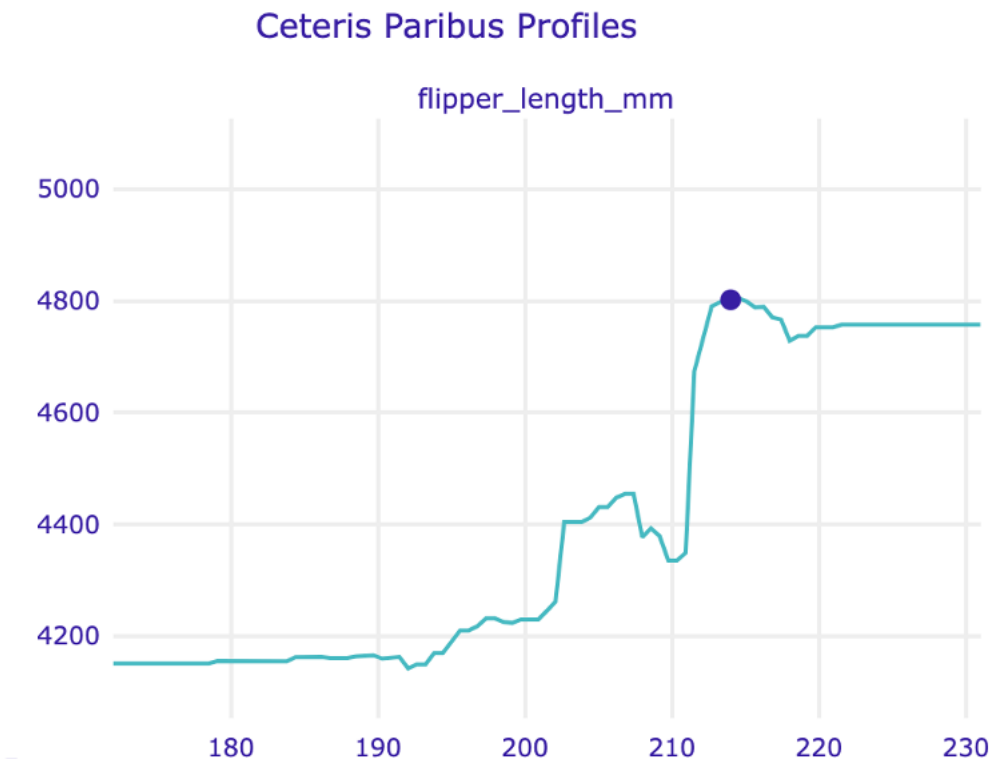


Categorical features

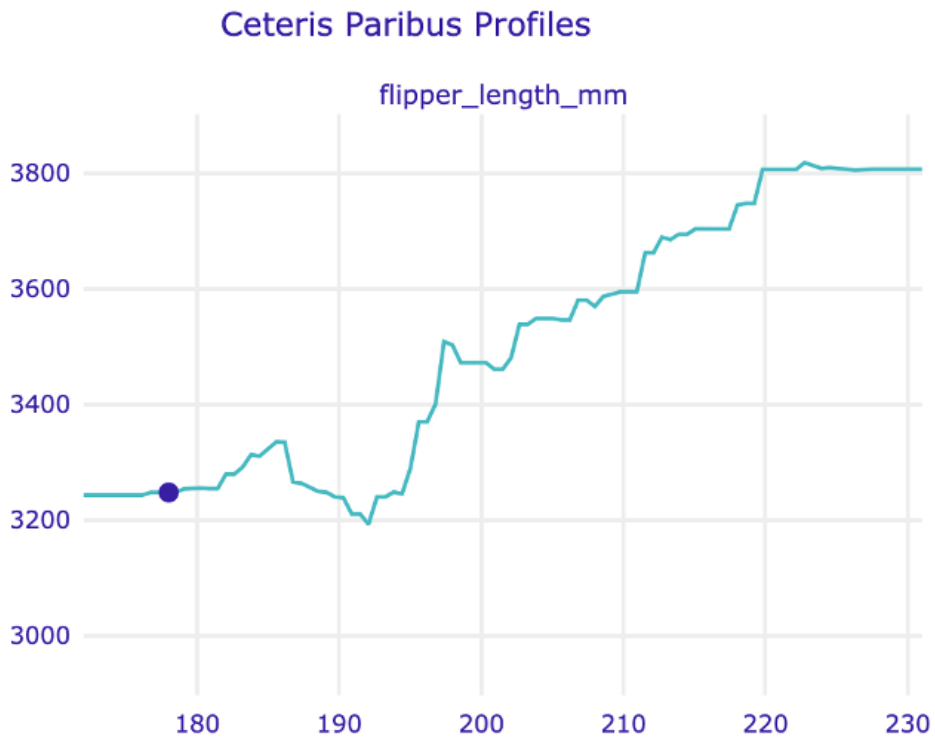
GO DATA DRIVEN

Ceteris Paribus profile

Different plots for different *data points*



Row 231: A Gentoo penguin

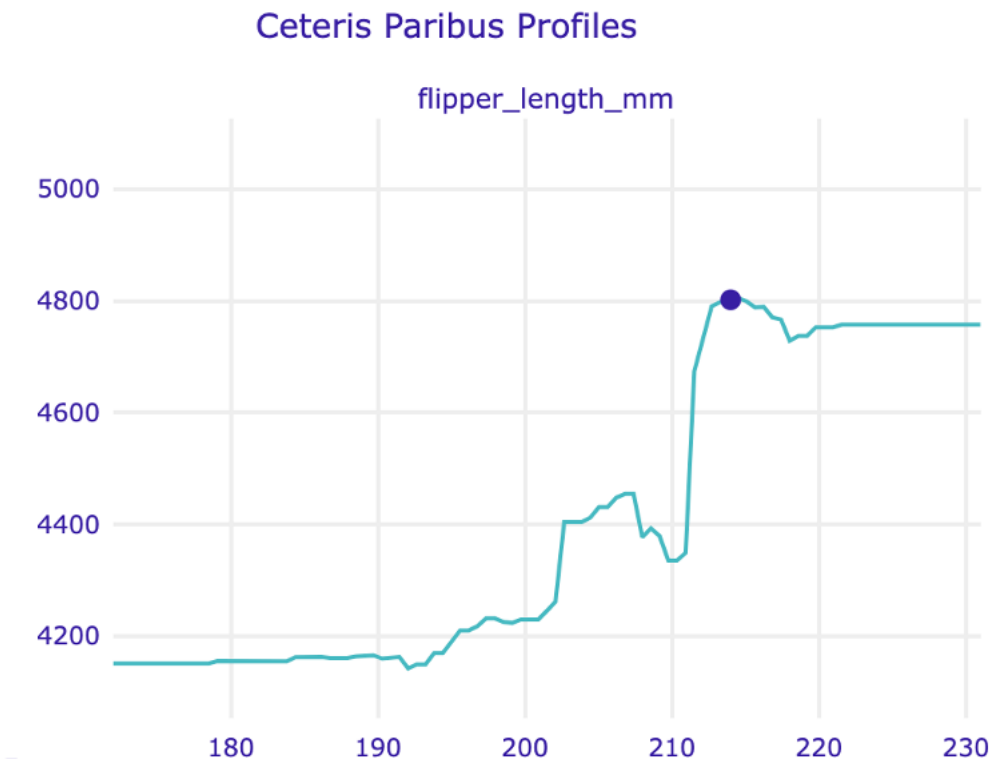


Row 30: An Adélie penguin

GO DATA DRIVEN

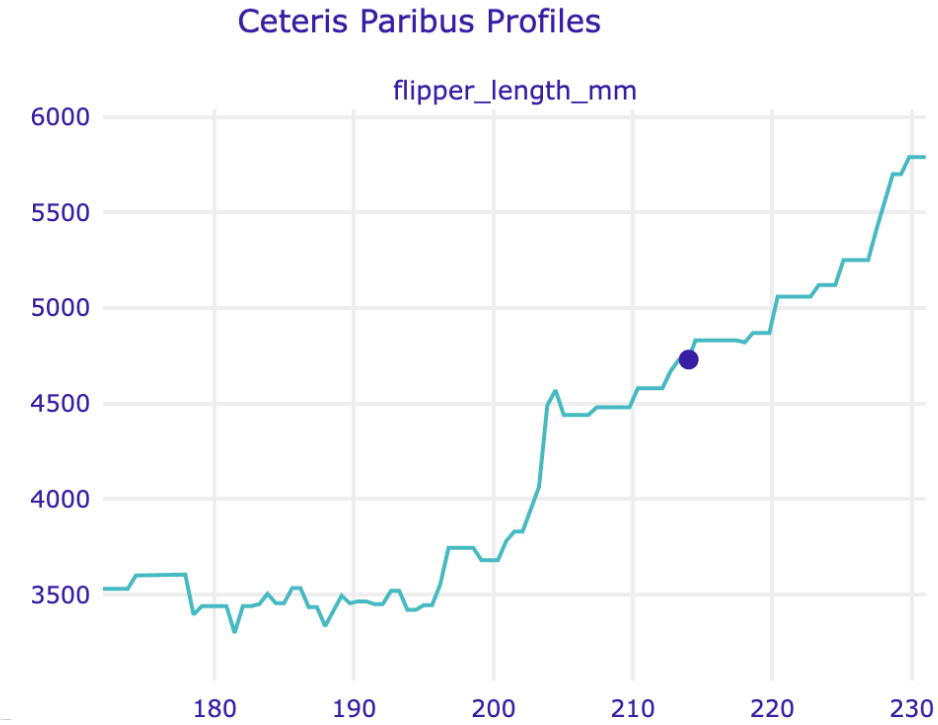
Ceteris Paribus profile

Different plots for different *models*



Row 231: A Gentoo penguin

Random Forest Regressor



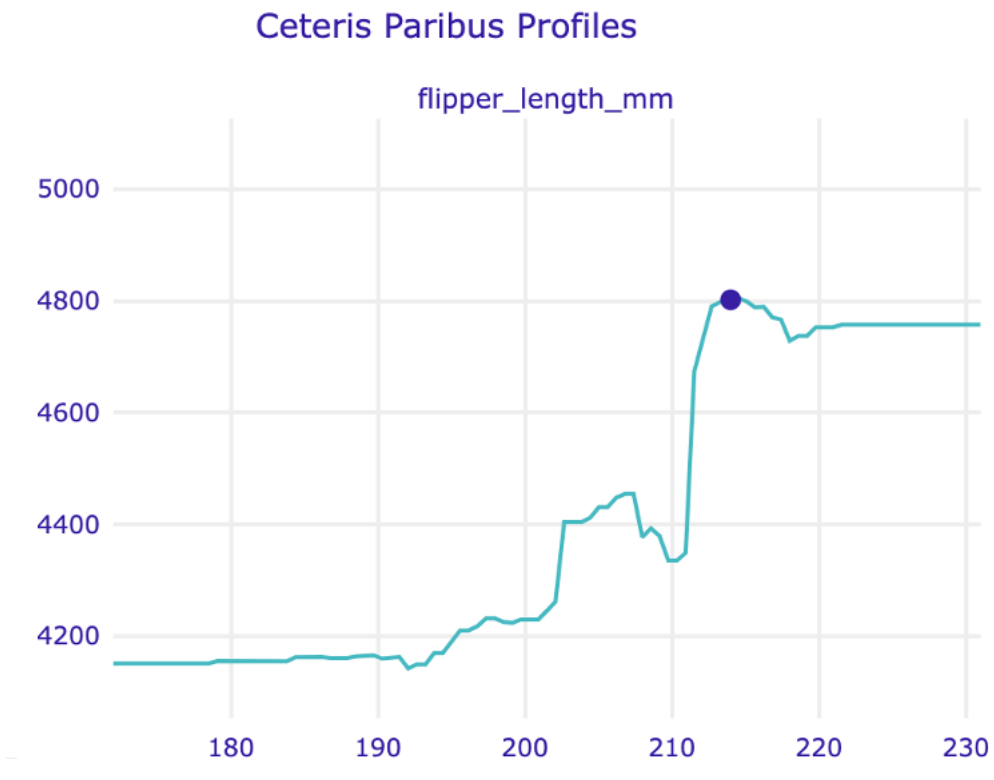
Row 231: A Gentoo penguin

K-Nearest Neighbors Regressor

GO DATA DRIVEN

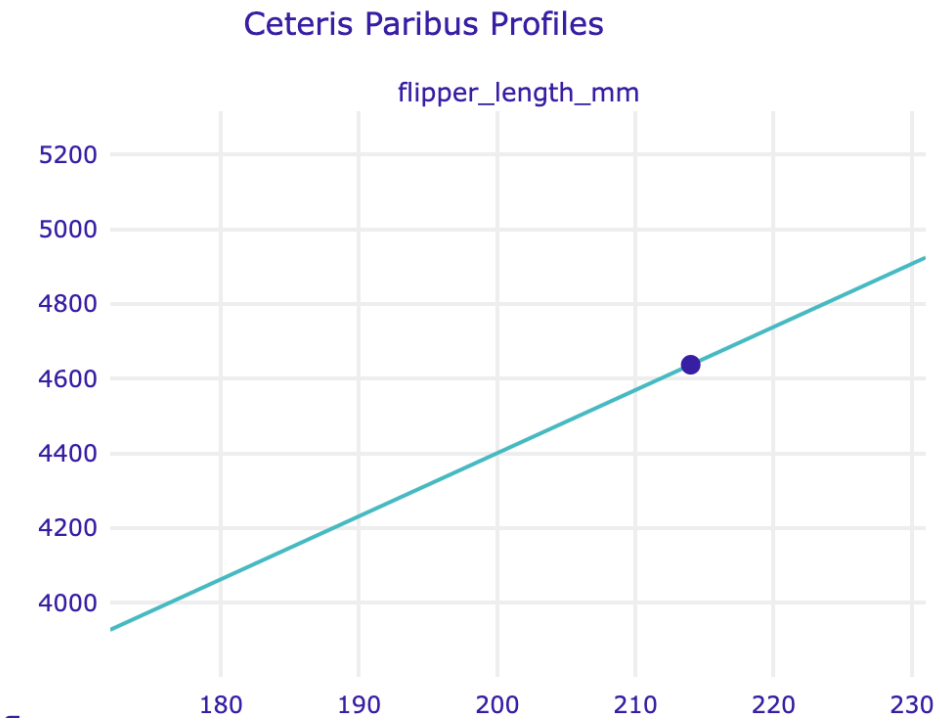
Ceteris Paribus profile

Different plots for different *models*



Row 231: A Gentoo penguin

Random Forest Regressor



Row 231: A Gentoo penguin

Linear Regression

GO DATA DRIVEN

Ceteris Paribus profile

Advantages

- Works on a **local** level for an individual prediction
- Displays model's **sensitivity** to a feature
- **Intuitive** to understand
- (**Easy** to implement)

Disadvantages

- Can only display **one feature** meaningfully at the time
- Misleading for **correlated** features
- Only applicable to **one** data point at the time – does not tell you anything about overall trend
- Does not tell you **how much a feature contributes** to individual prediction

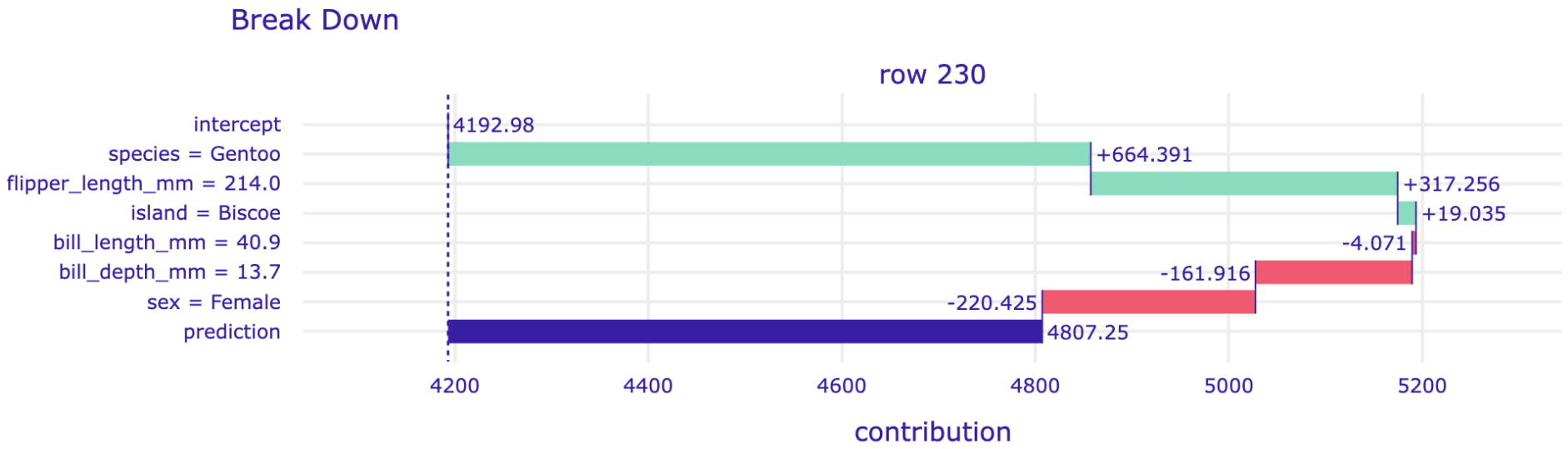
GO DATA DRIVEN

Jupyter Notebook 3

03_Prediction_Break_Down.ipynb

GO DATA DRIVEN

Break-down Plots for Additive Attributions

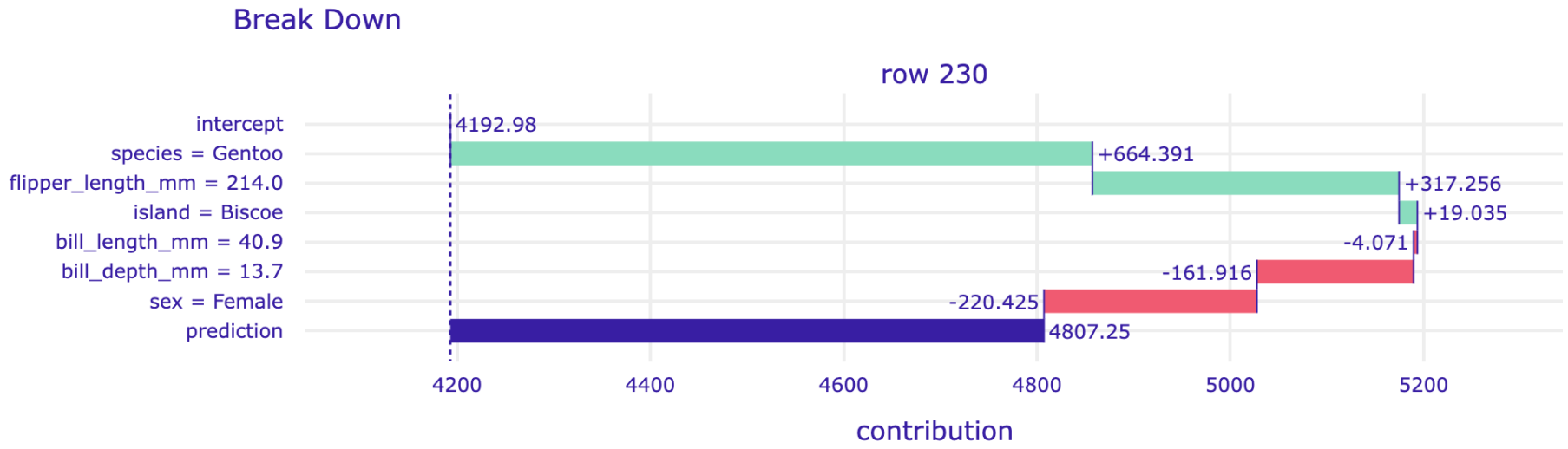


How to calculate the intercept:

1. Generate predictions with the model on the **train set** `y_hat = model.predict(X_train)`
2. Calculate the **average prediction** on the train set `intercept = y_hat.mean()`

GO DATA DRIVEN

Break-down Plots for Additive Attributions



How to calculate the value for *species = Gentoo*:

1. Assume that **all** datapoints in your dataset have the value Gentoo

```
X_train['species'] = 'Gentoo'
```

2. Calculate the **average prediction** on the altered (!) train set

```
value = model.predict(X_train).mean()
```

3. Subtract the intercept

```
value - intercept
```

GO DATA DRIVEN

Ceteris Paribus profile

Advantages

- Works on a **local** level for an individual prediction
- Displays a feature's **importance**
- **Fast** to calculate
- **Visual**

Disadvantages

- The **order** of the features matters
- Misleading for **interactions**
- **Messy** for many variables
- Does not tell you how **sensitive** a model is to a feature
- Only applicable to **one** data point at the time – does not tell you anything about feature importance

GO DATA DRIVEN

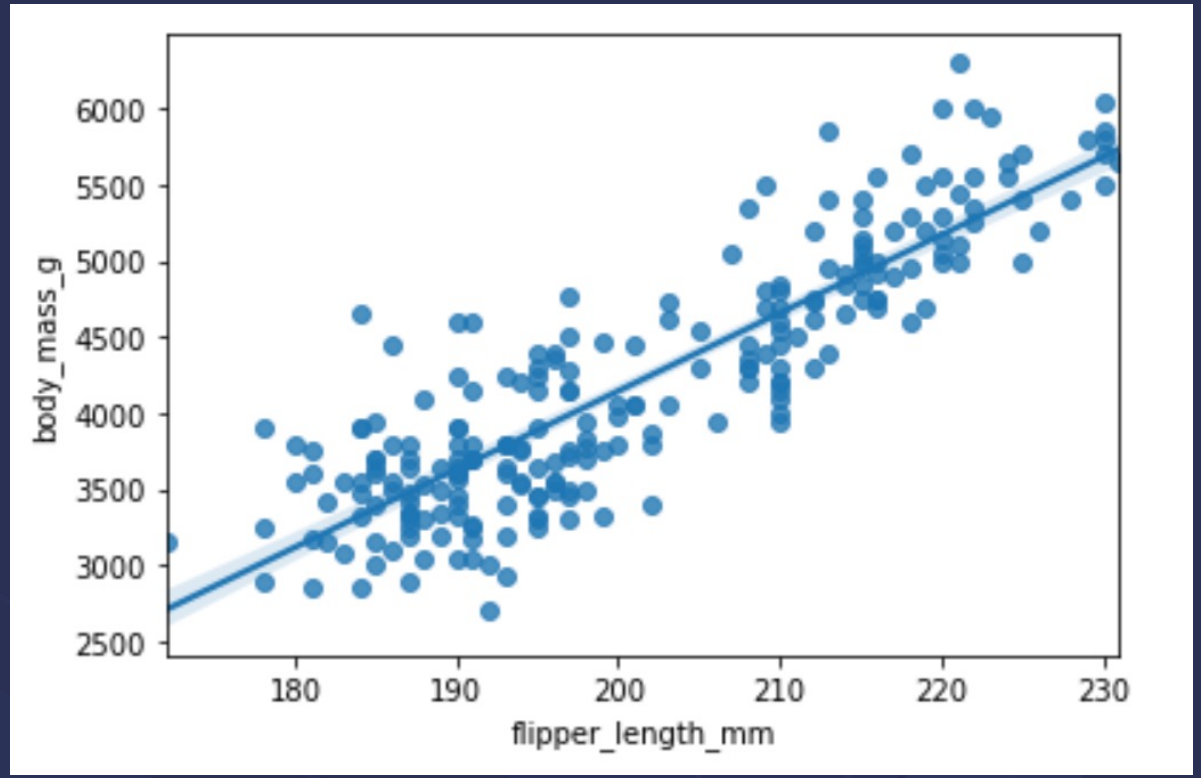
Global Feature Importance

GO DATA DRIVEN

Linear Regression

GO DATA DRIVEN

Global feature impotances



Linear Regression

*m*bill depth + n*flipper length + c = body mass*

*14*bill depth + 53*flipper length - 6600 = body mass*

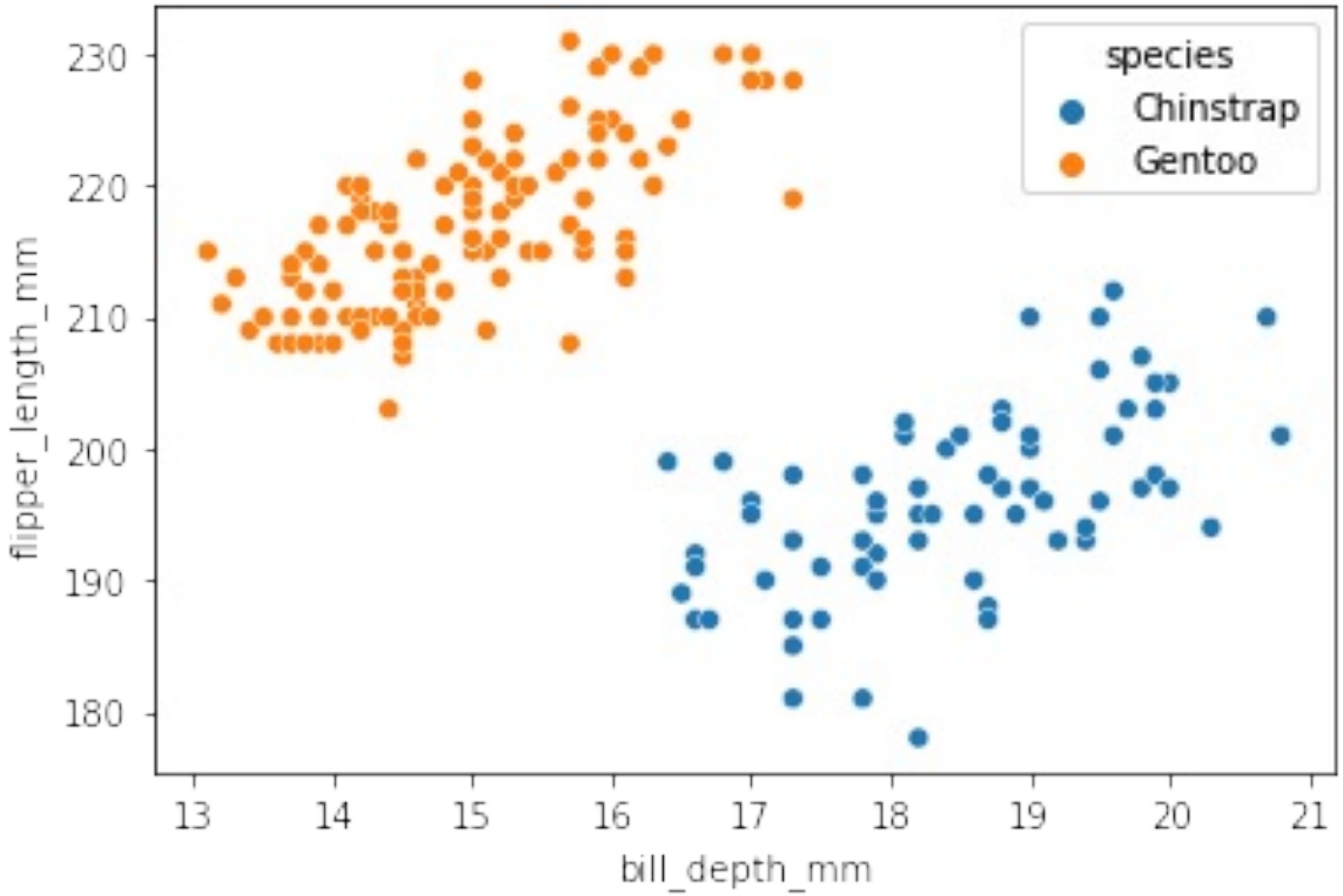
bill_depth_mm	flipper_length_mm	body_mass_g
18.7	181.0	3750.0

*14*18.7+ 53*181.0 - 6600 = 3171*

GO DATA DRIVEN

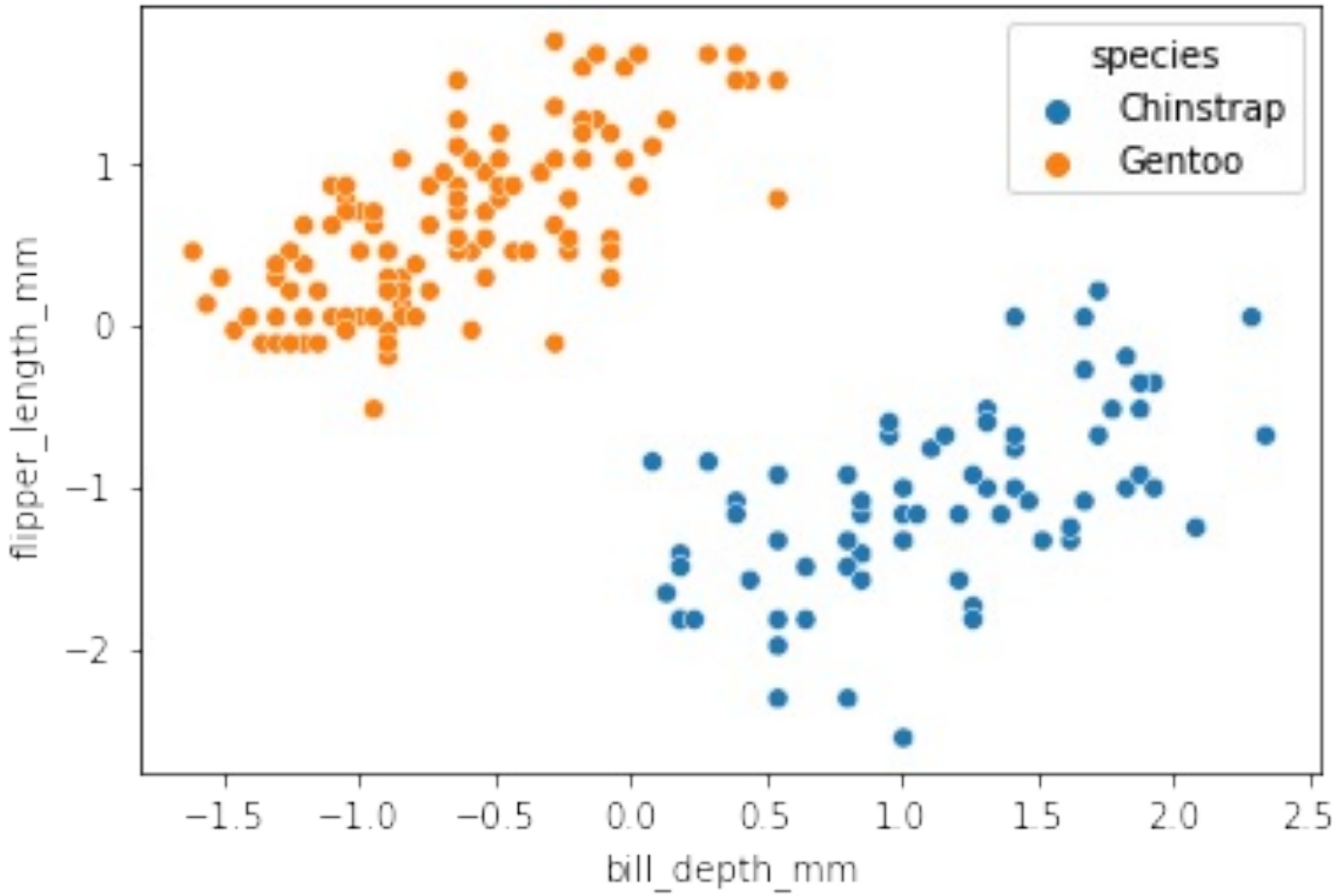
Linear Regression

*14*bill depth + 53*flipper length - 6600 = body mass*



GO DATA DRIVEN

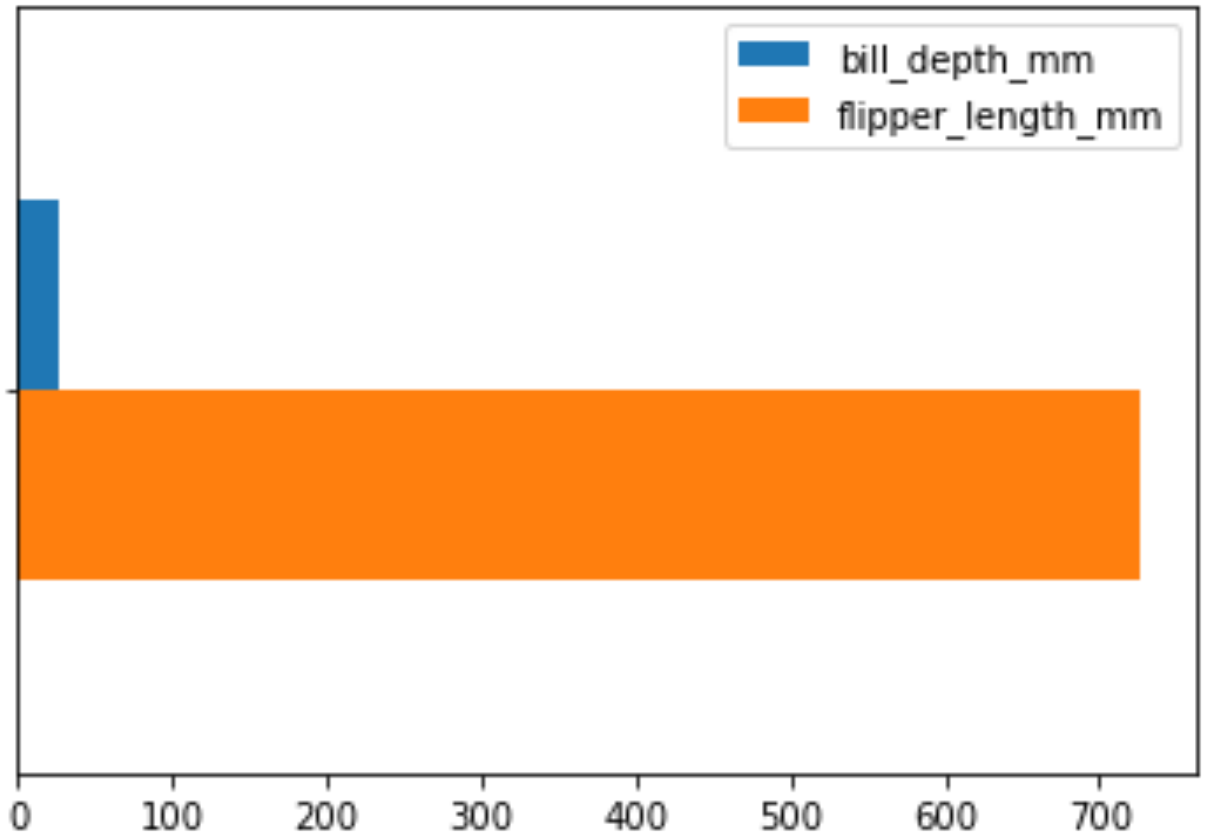
Linear Regression



GO DATA DRIVEN

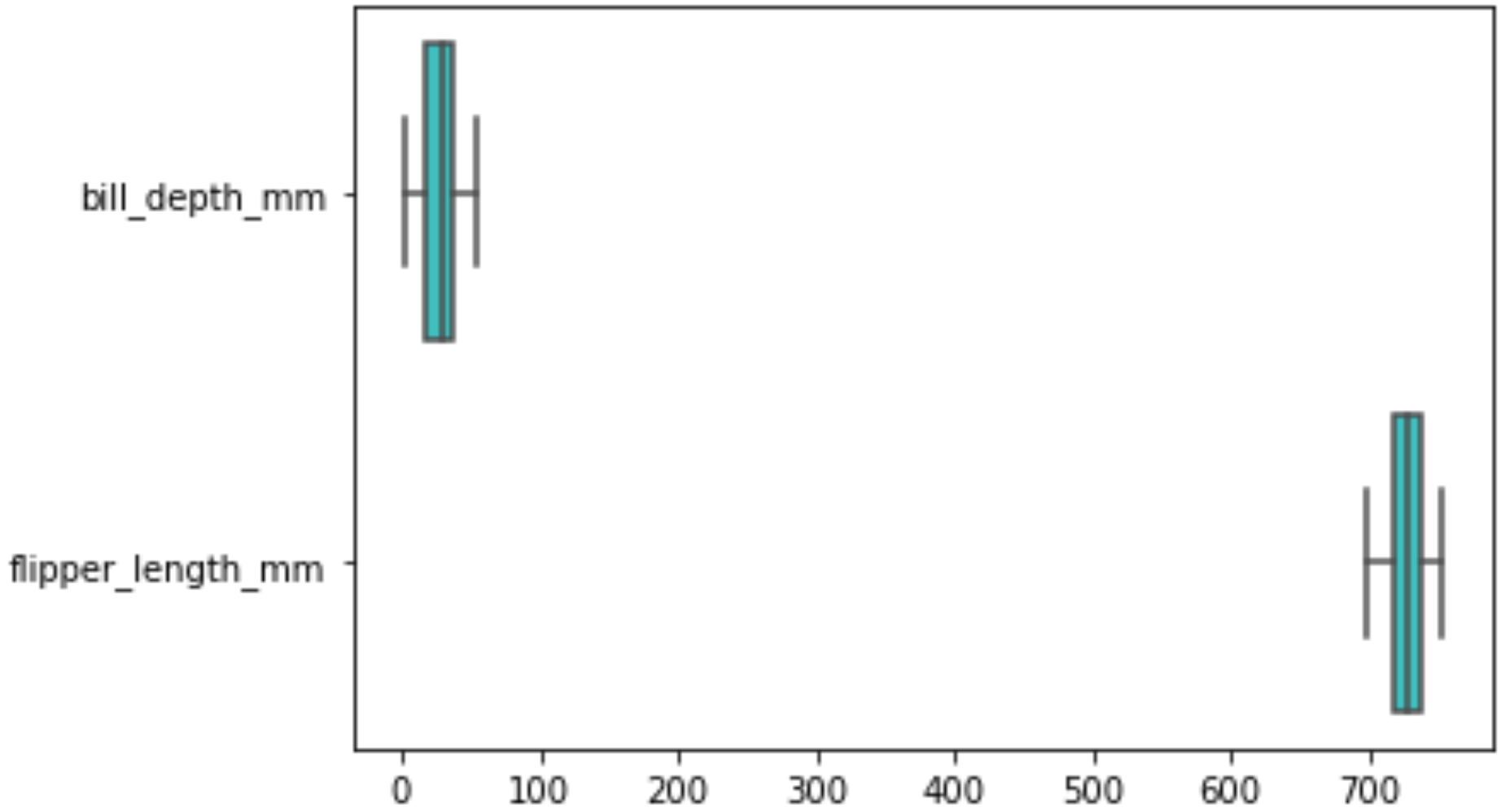
Linear Regression

*28*bill depth + 727*flipper length + 4199 = body mass*



GO DATA DRIVEN

Linear Regression

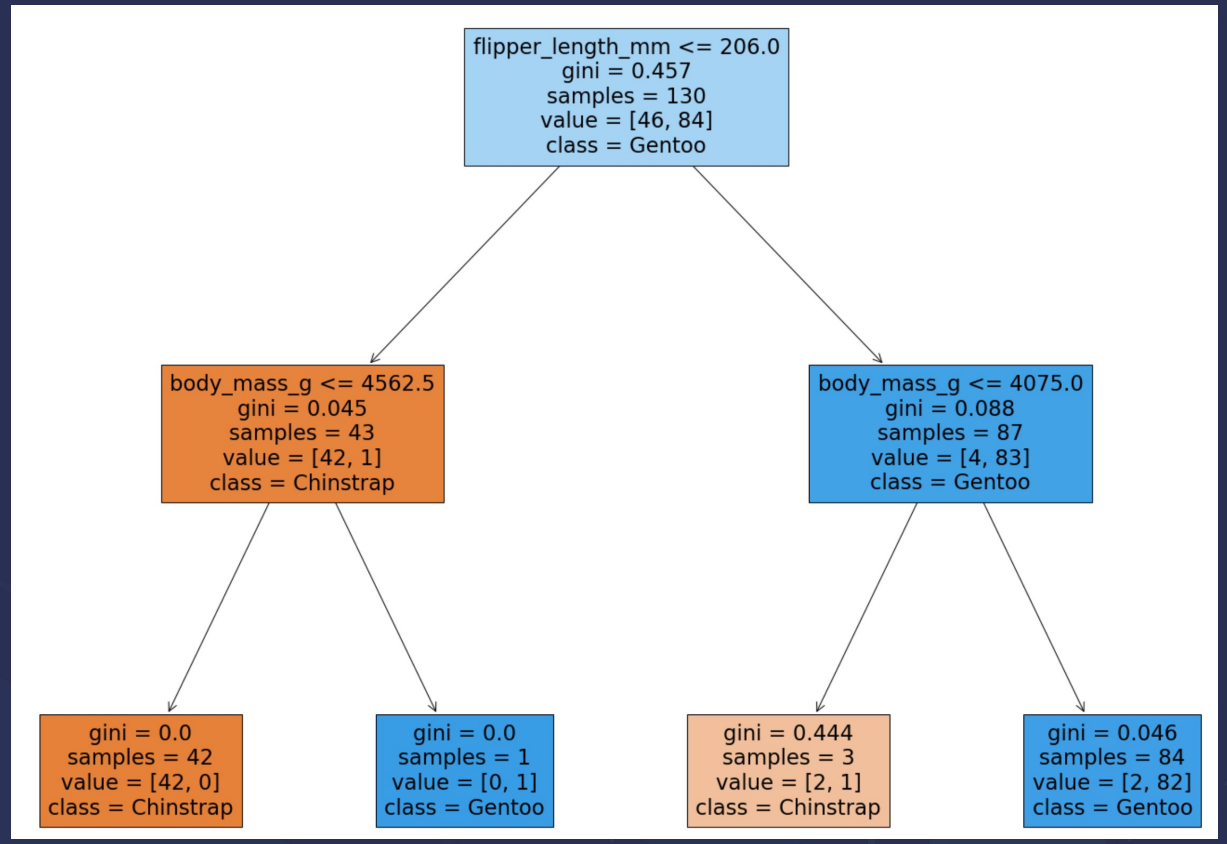


GO DATA DRIVEN

Decision Tree Classifer

GO DATA DRIVEN

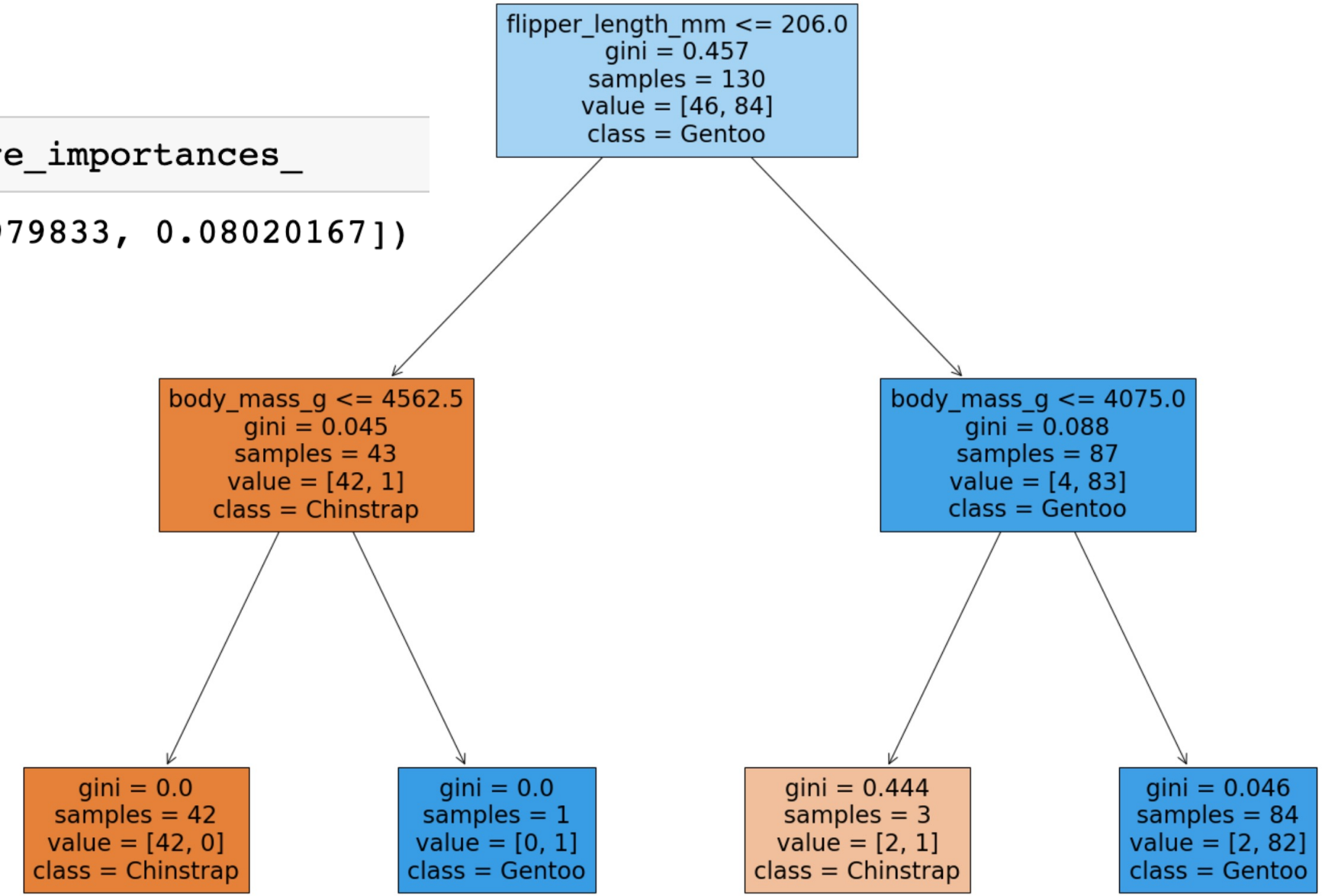
Global feature impotances



Decision Tree Classifier

```
model.feature_importances_
```

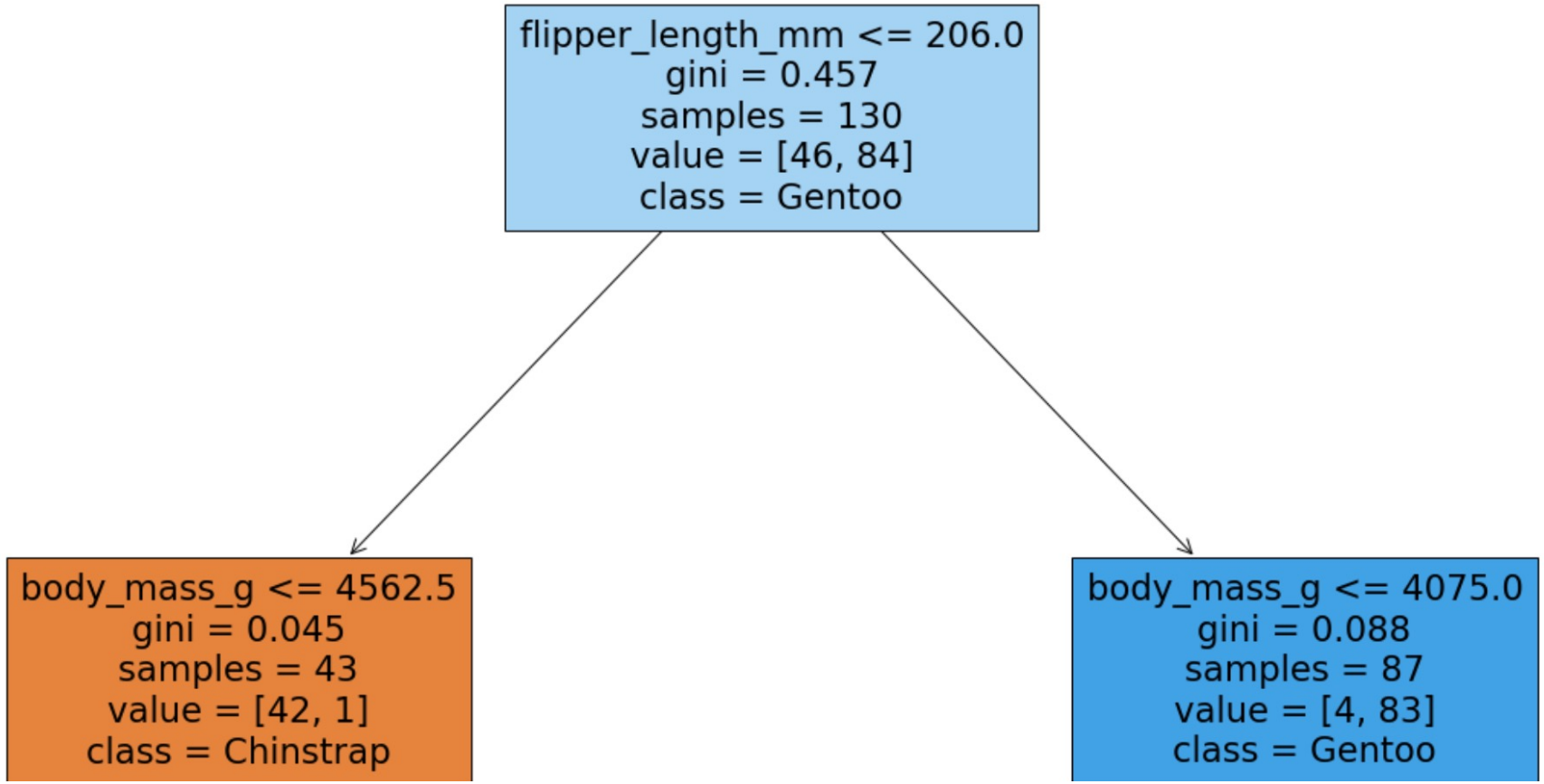
```
array([0.91979833, 0.08020167])
```



GO DATA DRIVEN

Decision Tree Classifier

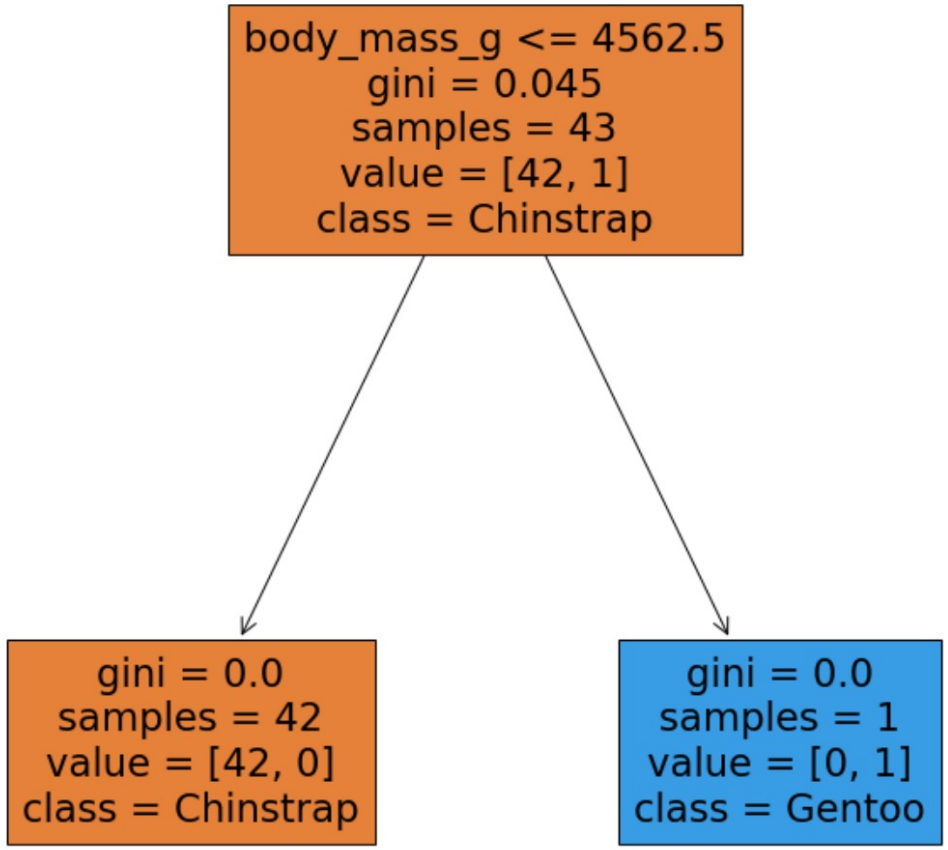
	feature importances
flipper_length_mm	0.919798
body_mass_g	0.080202

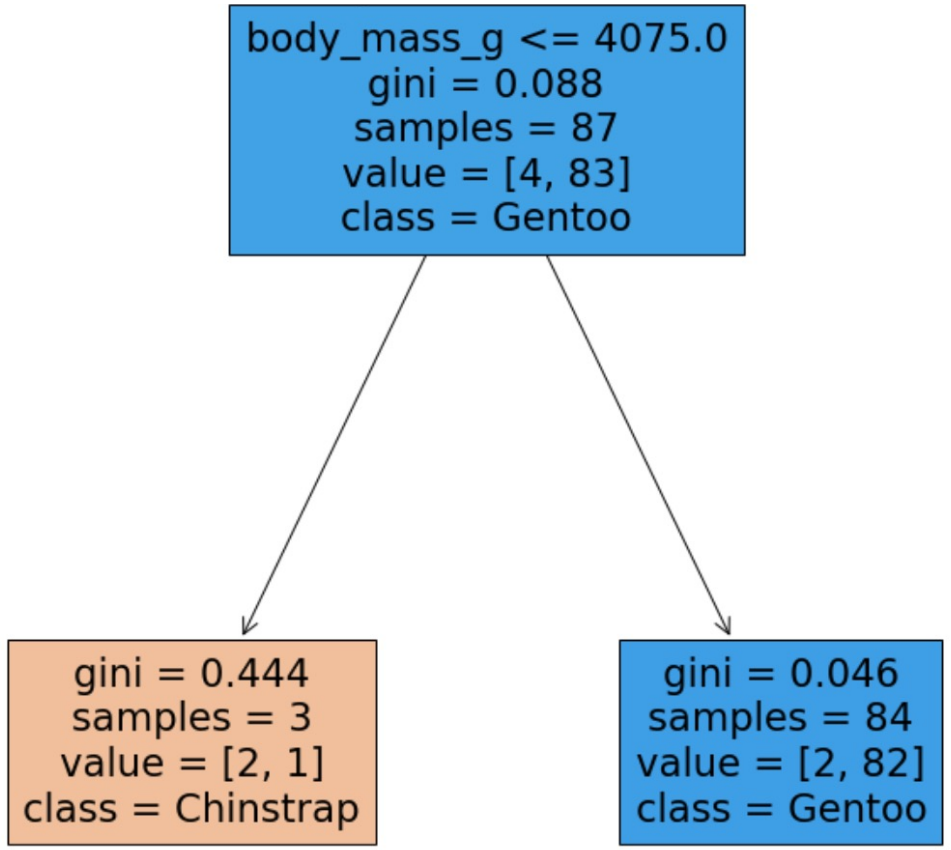


GO DATA DRIVEN

Decision Tree Classifier

feature importances	
flipper_length_mm	0.919798
body_mass_g	0.080202



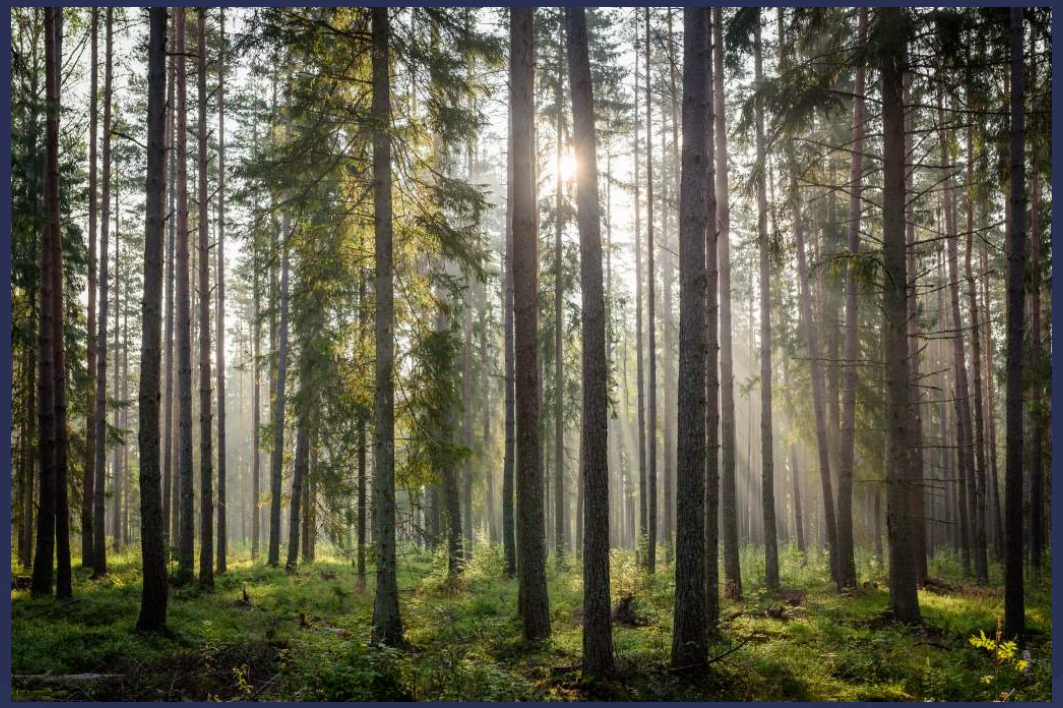


GO DATA DRIVEN

Random Forest Classifer

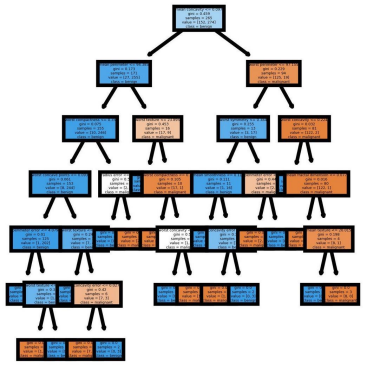
GO DATA DRIVEN

Global feature impotances

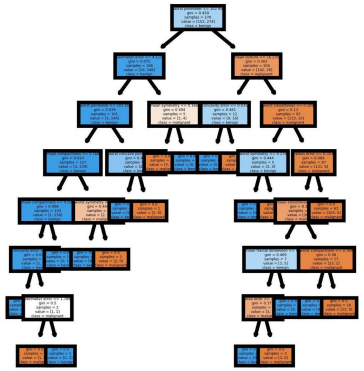


Random Forest Classifier

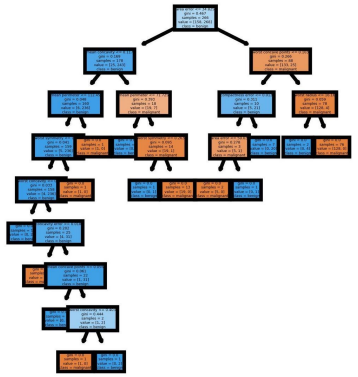
Estimator: 0



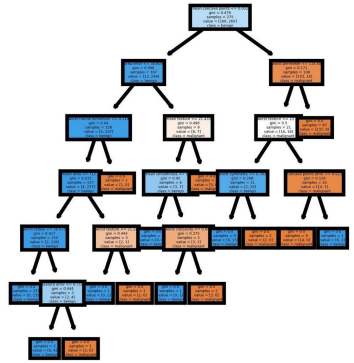
Estimator: 1



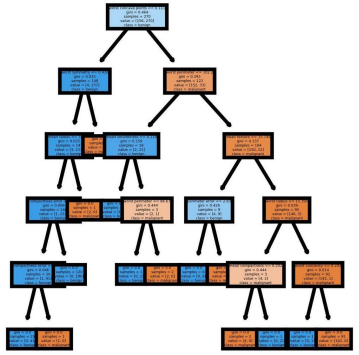
Estimator: 2



Estimator: 3



Estimator: 4

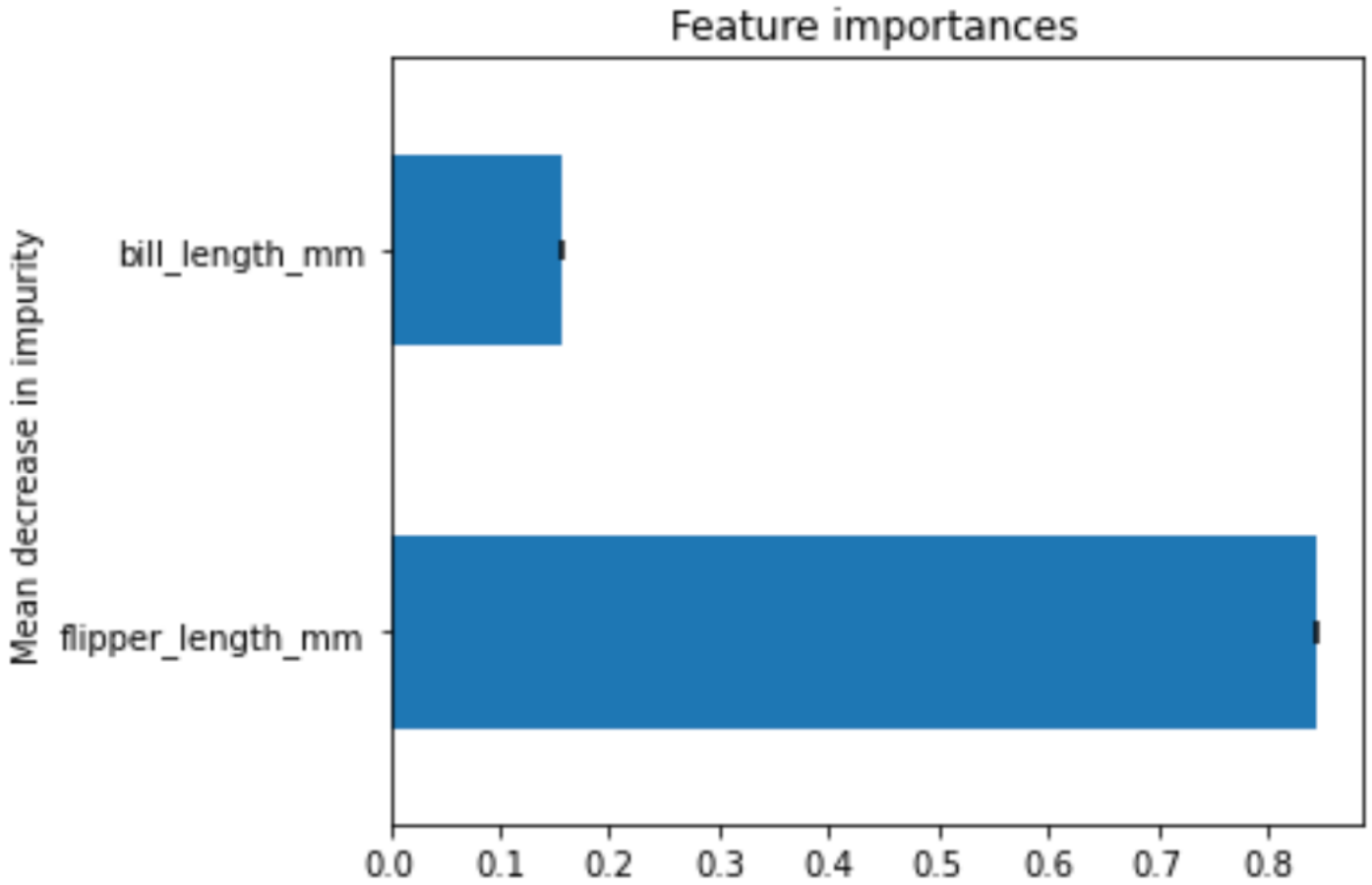


```
model.feature_importances_
```

```
array([0.84263378, 0.15736622])
```

GO DATA DRIVEN

Random Forest Classifier



GO DATA DRIVEN

Jupyter Notebook 4

04_Permutation_Feature_Importance.ipynb

GO DATA DRIVEN

Model-Agnostic Feature Importance

	flipper_length_mm	bill_length_mm	bill_depth_mm	sex	island	species
27	187.0	40.5	17.9	Female	Biscoe	Adelie
292	221.0	48.2	15.6	Male	Biscoe	Gentoo
302	212.0	47.4	14.6	Female	Biscoe	Gentoo
62	185.0	37.6	17.0	Female	Biscoe	Adelie
181	205.0	52.8	20.0	Male	Dream	Chinstrap
15	185.0	36.6	17.8	Female	Torgersen	Adelie
231	216.0	49.0	16.1	Male	Biscoe	Gentoo
233	213.0	48.4	14.6	Male	Biscoe	Gentoo
87	189.0	36.9	18.6	Female	Dream	Adelie
159	197.0	51.3	18.2	Male	Dream	Chinstrap

0.856

GO DATA DRIVEN

Drop Feature Importance

	flipper_length_mm	bill_length_mm	bill_depth_mm	sex	island	species
27	187.0	40.5	17.9	Female	Biscoe	Adelie
292	221.0	48.2	15.6	Male	Biscoe	Gentoo
302	212.0	47.4	14.6	Female	Biscoe	Gentoo
62	185.0	37.6	17.0	Female	Biscoe	Adelie
181	205.0	52.8	20.0	Male	Dream	Chinstrap
15	185.0	36.6	17.8	Female	Torgersen	Adelie
231	216.0	49.0	16.1	Male	Biscoe	Gentoo
233	213.0	48.4	14.6	Male	Biscoe	Gentoo
87	189.0	36.9	18.6	Female	Dream	Adelie
159	197.0	51.3	18.2	Male	Dream	Chinstrap

0.856

	flipper_length_mm	bill_length_mm	sex	island	species
27	187.0	40.5	Female	Biscoe	Adelie
292	221.0	48.2	Male	Biscoe	Gentoo
302	212.0	47.4	Female	Biscoe	Gentoo
62	185.0	37.6	Female	Biscoe	Adelie
181	205.0	52.8	Male	Dream	Chinstrap
15	185.0	36.6	Female	Torgersen	Adelie
231	216.0	49.0	Male	Biscoe	Gentoo
233	213.0	48.4	Male	Biscoe	Gentoo
87	189.0	36.9	Female	Dream	Adelie
159	197.0	51.3	Male	Dream	Chinstrap

???

GO DATA DRIVEN

Permutation Feature Importance

	flipper_length_mm	bill_length_mm	bill_depth_mm	sex	island	species
27	187.0	40.5	17.9	Female	Biscoe	Adelie
292	221.0	48.2	15.6	Male	Biscoe	Gentoo
302	212.0	47.4	14.6	Female	Biscoe	Gentoo
62	185.0	37.6	17.0	Female	Biscoe	Adelie
181	205.0	52.8	20.0	Male	Dream	Chinstrap
15	185.0	36.6	17.8	Female	Torgersen	Adelie
231	216.0	49.0	16.1	Male	Biscoe	Gentoo
233	213.0	48.4	14.6	Male	Biscoe	Gentoo
87	189.0	36.9	18.6	Female	Dream	Adelie
159	197.0	51.3	18.2	Male	Dream	Chinstrap

0.856

	flipper_length_mm	bill_length_mm	sex	island	species
27	187.0	40.5	Female	Biscoe	Adelie
292	221.0	48.2	Male	Biscoe	Gentoo
302	212.0	47.4	Female	Biscoe	Gentoo
62	185.0	37.6	Female	Biscoe	Adelie
181	205.0	52.8	Male	Dream	Chinstrap
15	185.0	36.6	Female	Torgersen	Adelie
231	216.0	49.0	Male	Biscoe	Gentoo
233	213.0	48.4	Male	Biscoe	Gentoo
87	189.0	36.9	Female	Dream	Adelie
159	197.0	51.3	Male	Dream	Chinstrap

???

GO DATA DRIVEN

Permutation Feature Importance

Advantages

- Works on a **global** level for a model
- Displays a feature's **importance**
- **Fast** to calculate
- **Additional** tricks

Disadvantages

- Dependent on **randomness**
- Misleading for **correlations**
- Can lead to **unrealistic data** points
- Need access to **true** outcome
- Does not tell you how **sensitive** a model is to a feature

GO DATA DRIVEN

Jupyter Notebook 5

05_Partial_Dependence_Plots.ipynb

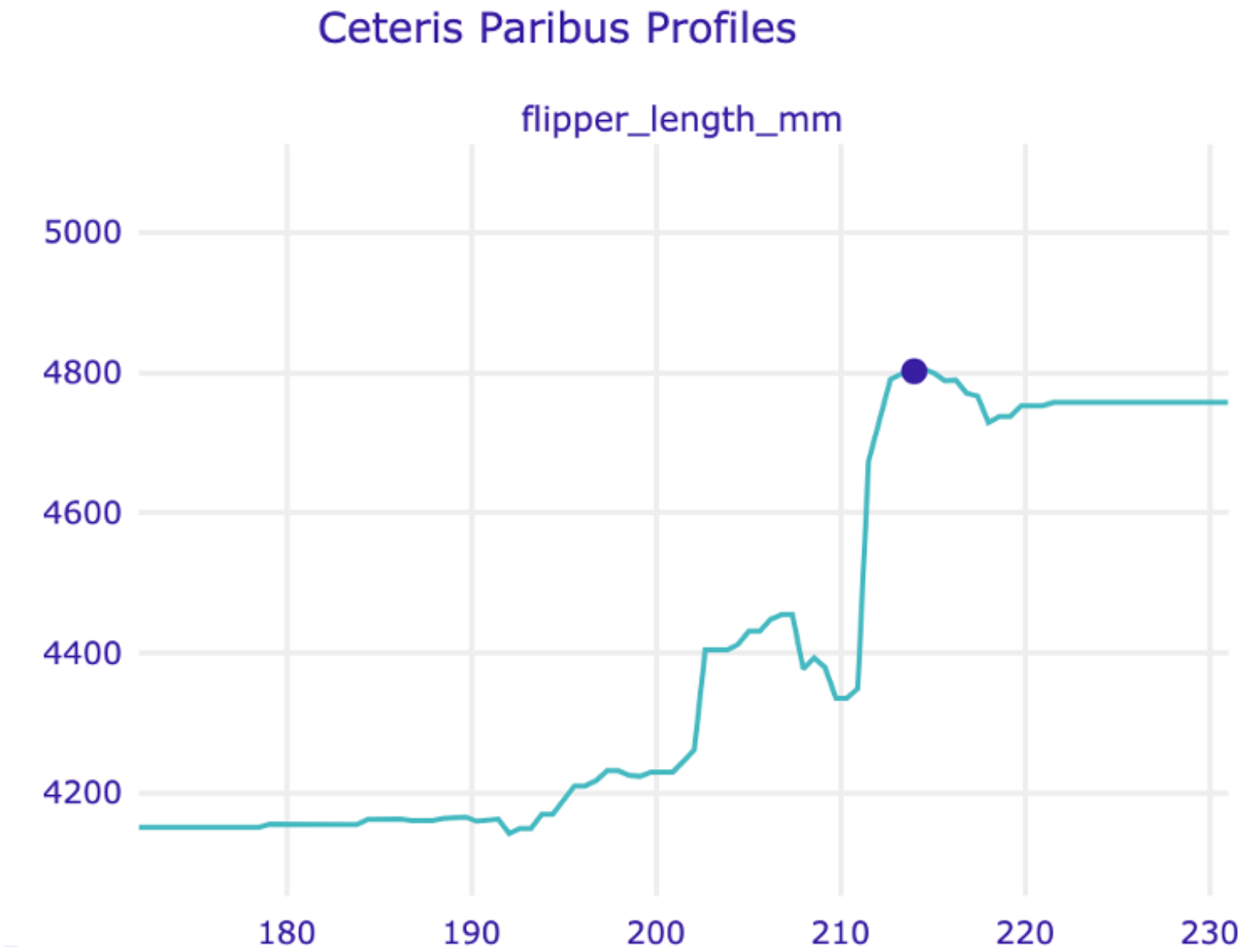
GO DATA DRIVEN

Global Sensitivity

GO DATA DRIVEN

Ceteris Paribus

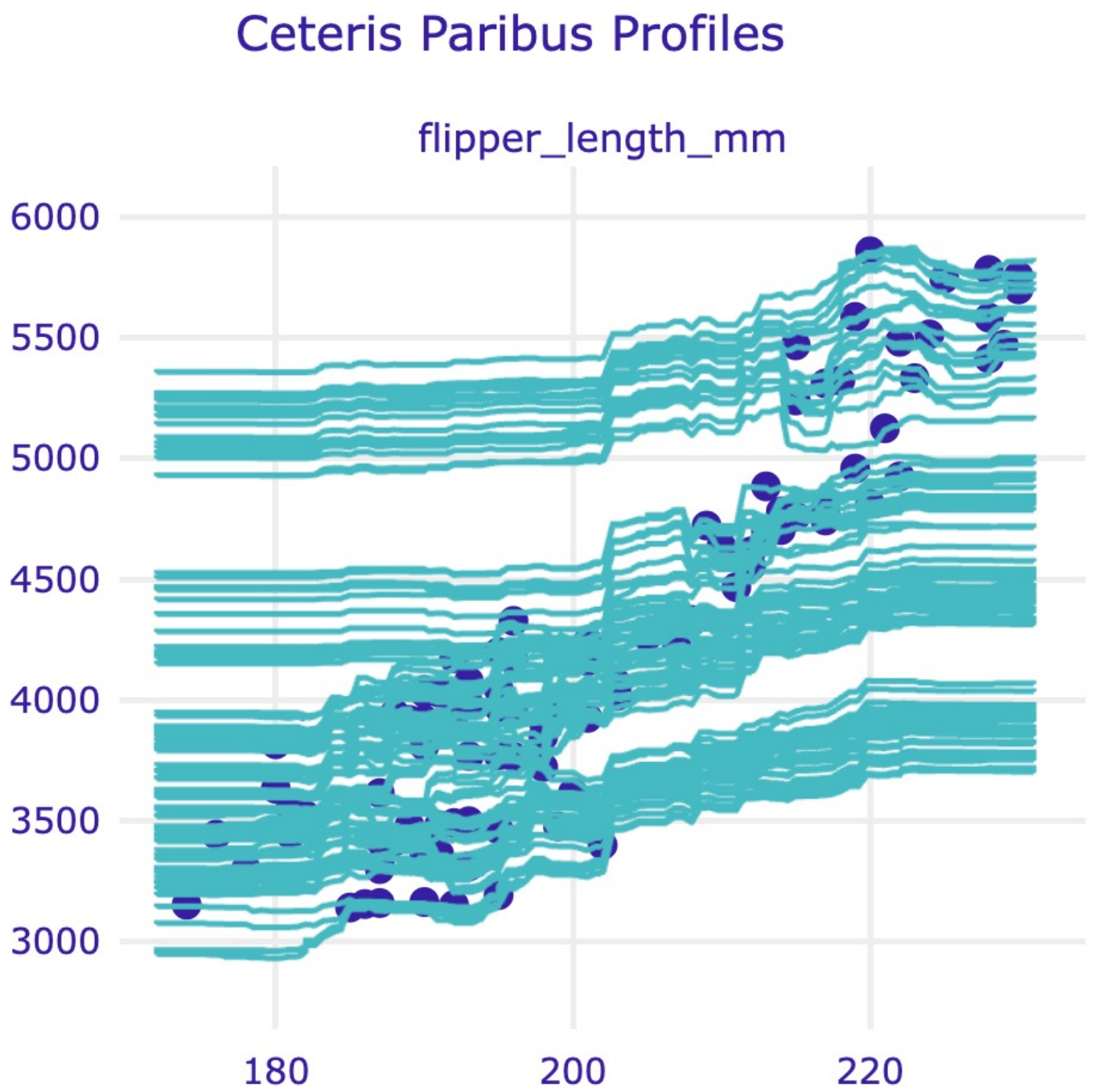
- **Sensivity** to feature
- On a **local** level



GO DATA DRIVEN

Individual Conditional Expectation

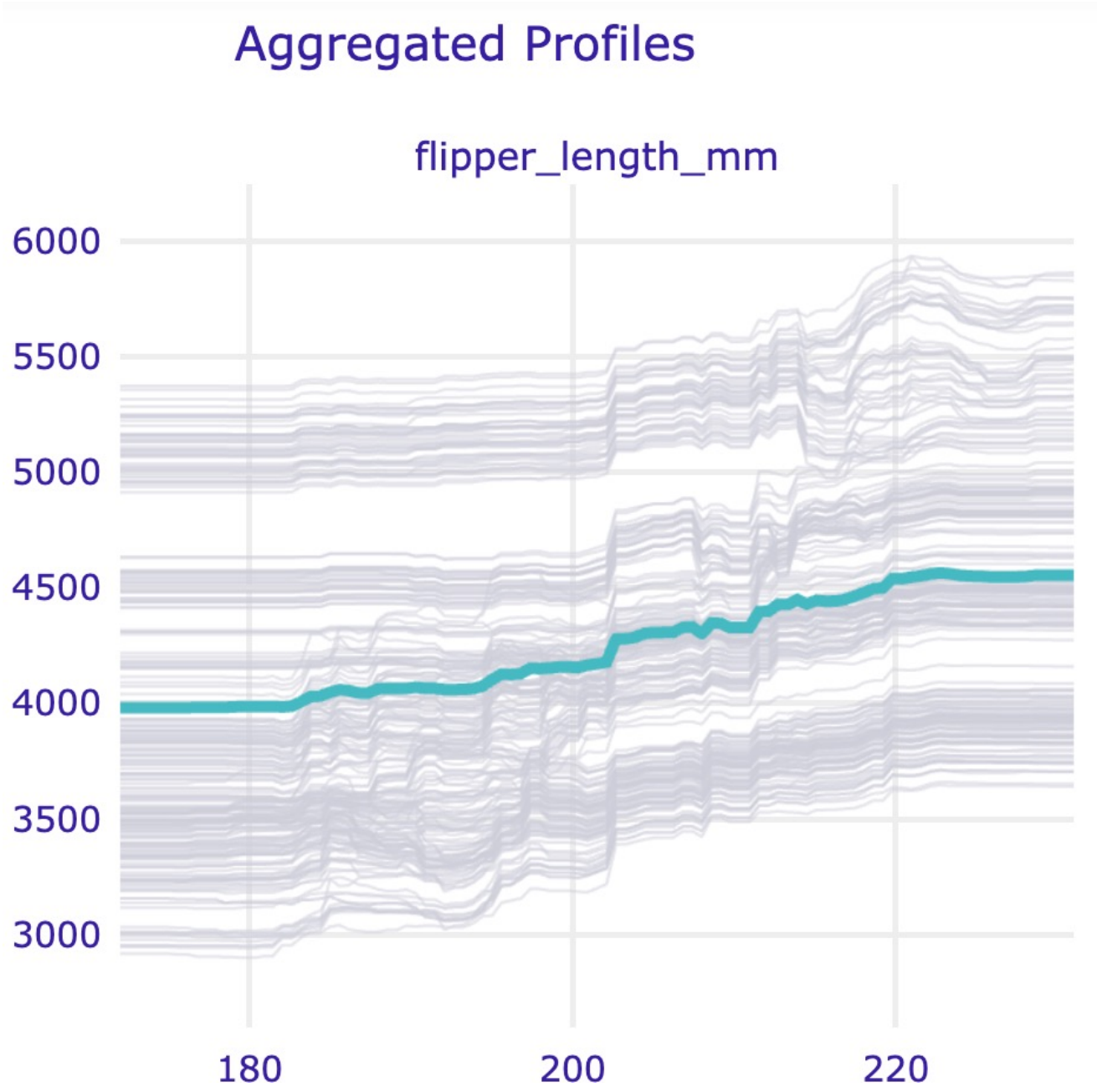
- **Sensivity** to feature
- One line per instance



GO DATA DRIVEN

Partial Dependence Plot

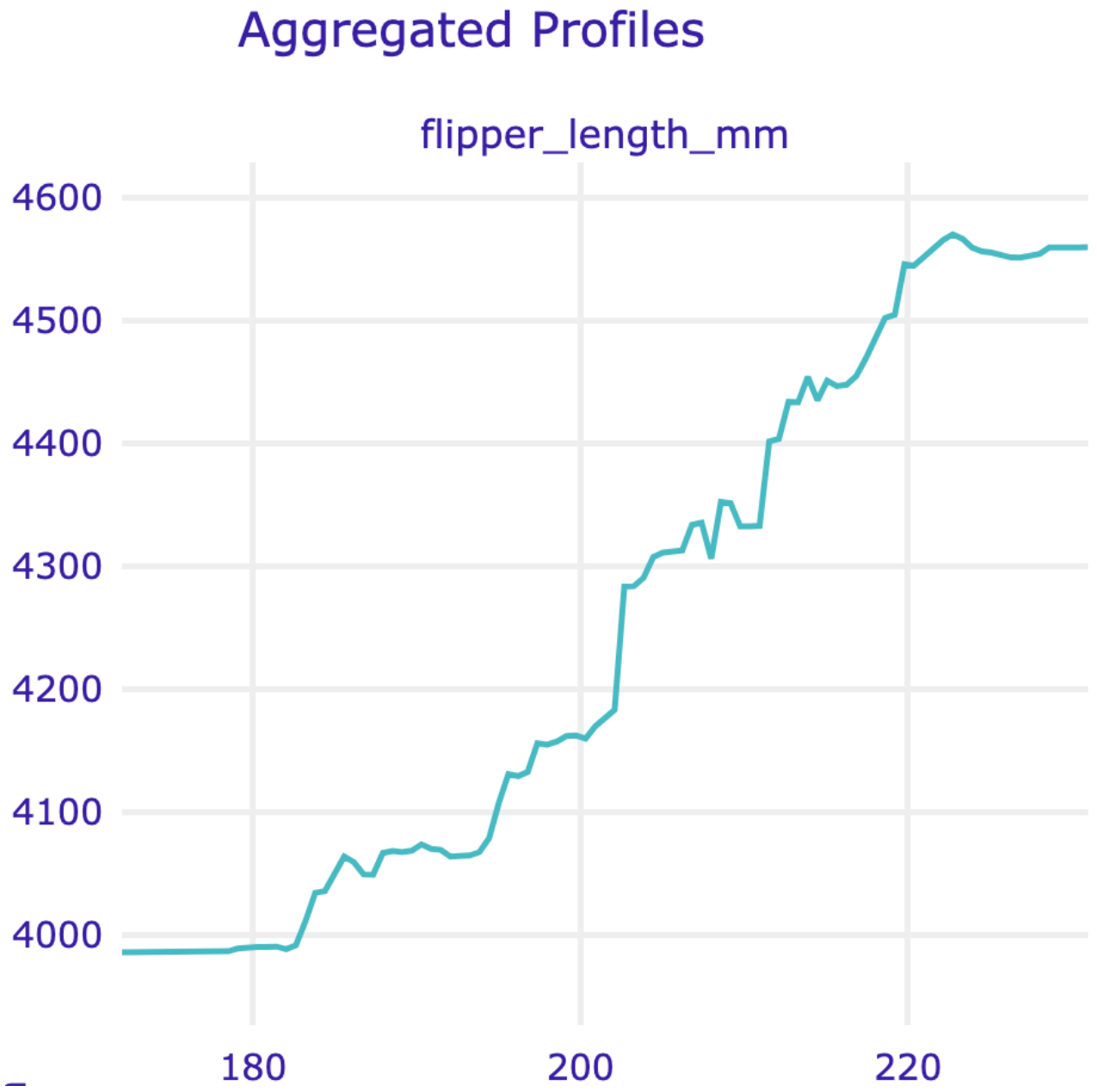
- **Sensivity** to feature
- Average of ICE plot
- **Global method**



GO DATA DRIVEN

Partial Dependence Plot

- **Sensivity** to feature
- Average of ICE plot
- **Global method**



GO DATA DRIVEN

Summary

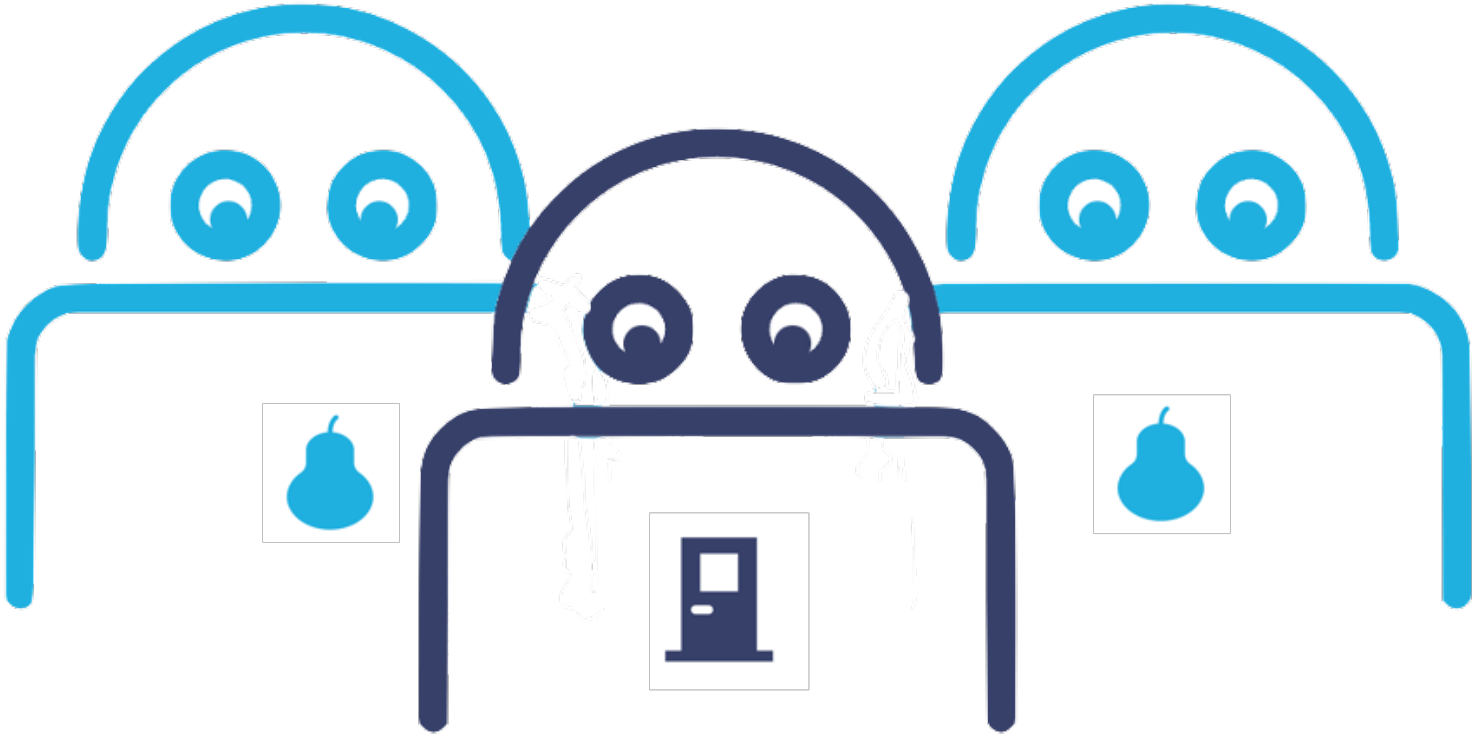
- Intrinsicly interpretable machine learning models
- Model-agnostic interpretability techniques

	Local	Global
Importance	Break-down plots	Permutation Feature Importance
Sensivitiy	Ceteris Paribus	Partial Dependence Plots

GO DATA DRIVEN

Hackathon!

- Credit Lending
- Laptop Prices
- Medical Diagnoses
- Discussion Hackathon (theory)



GO DATA DRIVEN

GO DATA DRIVEN

Unpacking the Black Box

How to Interpret your Machine Learning Model

GoDataDriven

@ AMLD Switzerland 2022

The Data & AI Maturity Journey

13:30-17:30, March 28 @ 1A

GO DATA DRIVEN

GoDataDriven Academy

- Public trainings
- In-company tailored trainings
- Virtual & in-person

Visit godatadriven.academy

GO DATA DRIVEN

GO DATA DRIVEN

Unpacking the Black Box

How to Interpret your Machine Learning Model