A Day with the Machine Learning Team

4 min

Next, Catherine is headed over to the Machine Learning and Algorithms Team. They use data to make predictions and create new products using data (like recommendation systems). Today they are trying to find trends among millions of learners according to their behavior on the site. They will use a cluster analysis.

Catherine is playing around with some sample data about penguins to learn about cluster analyses.

The dataset she's looking at is a collection of flipper and bill measurements for three different species of penguins collected by Dr. Kristen Gorman and the Palmer Station in Antarctica.

When you opened this exercise, some code running behind the scenes loaded the data and created a visualization of the flipper and bill measurements for three penguin species.

Take a look at the visualization in the learning environment. You might notice that there isn't much overlap between species on the plot. For example, Chinstrap penguins seem to usually have longer bills than Adelie penguins, and so Chinstrap penguins appear above Adelie penguins in the visualization. Regions like these that mostly feature one species over another are called **clusters**.

With learner data, you might find clusters centered around Data Science or Web Development.

We can use this to make predictions. For example, if we find a new penguin that has 180mm long flippers and a 35mm long bill, we might conclude (based on these clusters) that our penguin is more likely to be an Adelie penguin. In our code running behind the scenes, we've built a computer model to do this kind of prediction automatically.
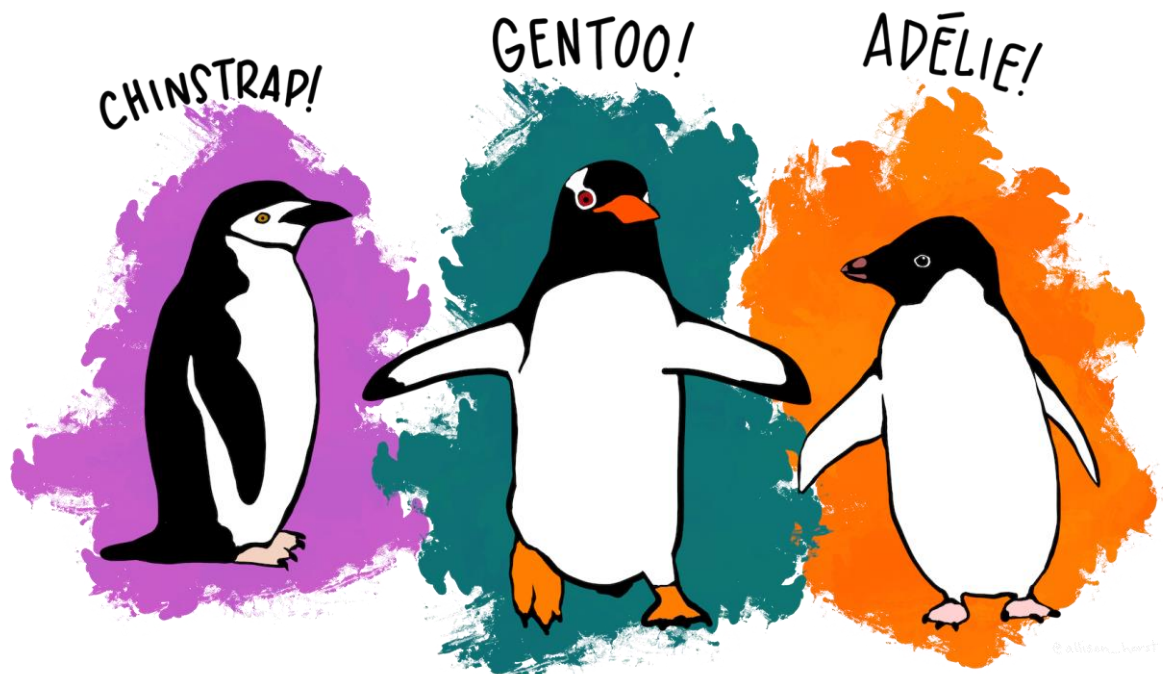
Instructions

1. Checkpoint 1 Passed

1.

Change the flipper and bill measurements and run the code. An algorithm we've loaded for you will attempt to predict which species of penguin matches the measurements you chose. How well do you think it does? What happens if you enter values that are between two regions, or values outside of the existing regions altogether?

In future Codecademy courses, you will learn to do all of this yourself! If you are curious, you can explore the code by clicking on the file icon in the center pane.

Did you change the measurements in lines 3 and 4?

script.py

Change these numbers and run to predict species!

mystery_penguin_flipper = 180

mystery_penguin_bill = 35

The model will use those values and the dataset to predict a species

from code import predict

predict(mystery_penguin_flipper,mystery_penguin_bill)

Prediction: this penguin belongs to the Adelie species!

Penguin Flippers and Bills by Species