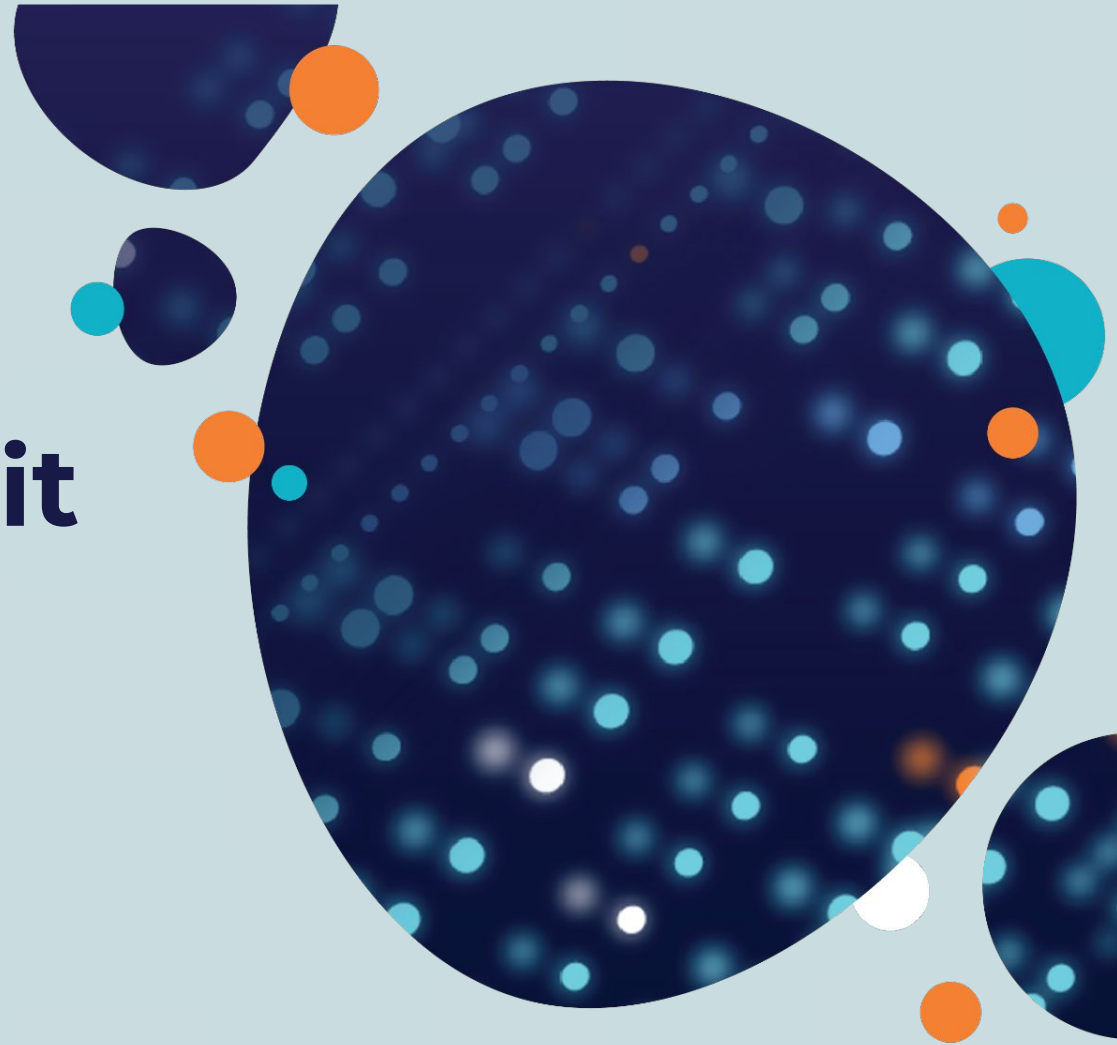
data science
student society

Intro to SciKit Learn

January 15, 2025



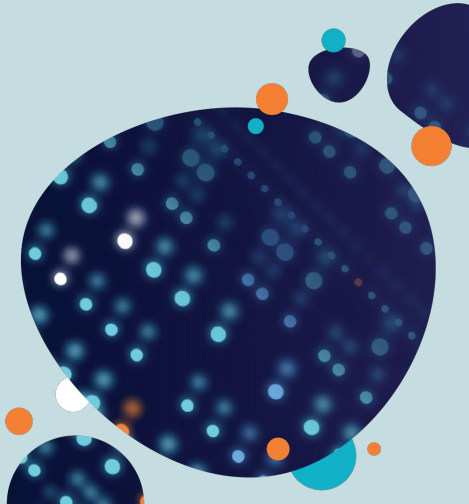
data science student society

Please check in!

Workshop Objectives

- Introduce Scikit-Learn
- Data Preprocessing
- Feature Engineering
- Introduce Common Machine Learning Models

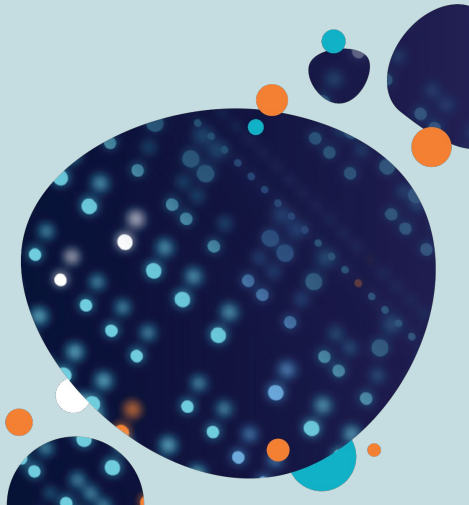
https://github.com/couchsnail/ds3-workshops.git



What is SciKit-Learn?

- SciKit-Learn (sklearn) is a Python library designed for machine learning
- It includes many different packages, including both regression and classification algorithms
- It can also be used to process data and create features

https://scikit-learn.org/stable/



Workshop Repo

- Please clone the Github repo here: https://github.com/couchsnail/ds3-workshops.git
- Here's the first 5 rows of the data we'll be looking at:

	#	Name	Type 1	Type 2	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
0	1	Bulbasaur	Grass	Poison	45	49	49	65	65	45	1	False
1	2	Ivysaur	Grass	Poison	60	62	63	80	80	60	1	False
2	3	Venusaur	Grass	Poison	80	82	83	100	100	80	1	False
3	3	VenusaurMega Venusaur	Grass	Poison	80	100	123	122	120	80	1	False
4	4	Charmander	Fire	NaN	39	52	43	60	50	65	1	False

https://github.com/couchsnail/ds3-workshops.git

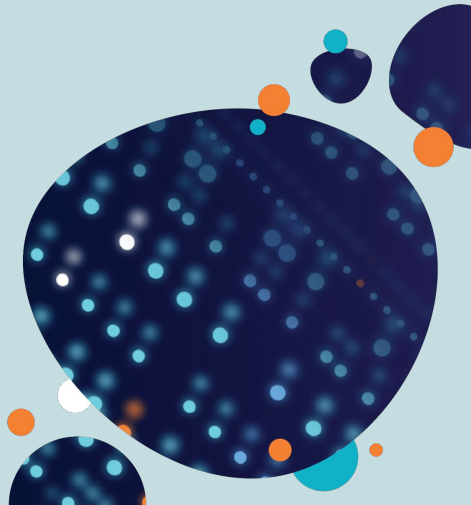
Features in Data Science

- In machine learning, you want to use information in the data to make predictions
  - Input information is called **features**
  - For classification tasks, output is called a **label**
  - For regression tasks, output is **numerical**
- **Features** of a dataset are the input data for predictions
- For example, say you're trying to build a model for predicting penguin species:
  - Features: beak length, wing length, foot size
  - Label: Gentoo, Adelie, Chinstrap

https://github.com/couchsnail/ds3-workshops.git

Feature Preprocessing

- Sklearn provides a variety of tools for preprocessing data to create features
- **StandardScaler -** standardizes training data around mean and standard deviation
- **Principal Component Analysis (PCA)** - dimensionality reduction



https://scikit-learn.org/1.5/data_transforms.html

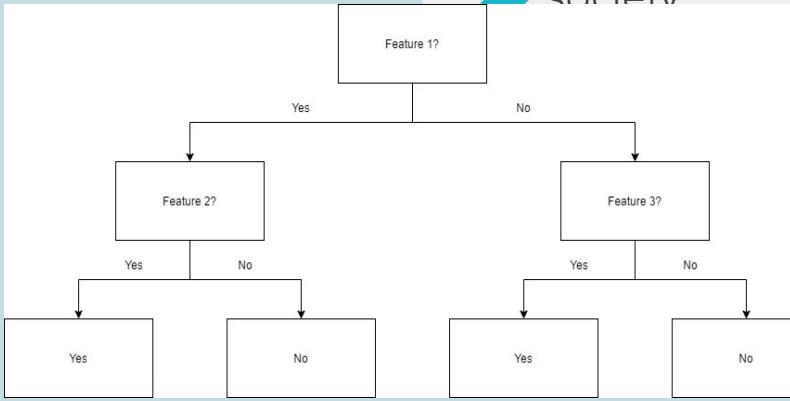
Feature Engineering

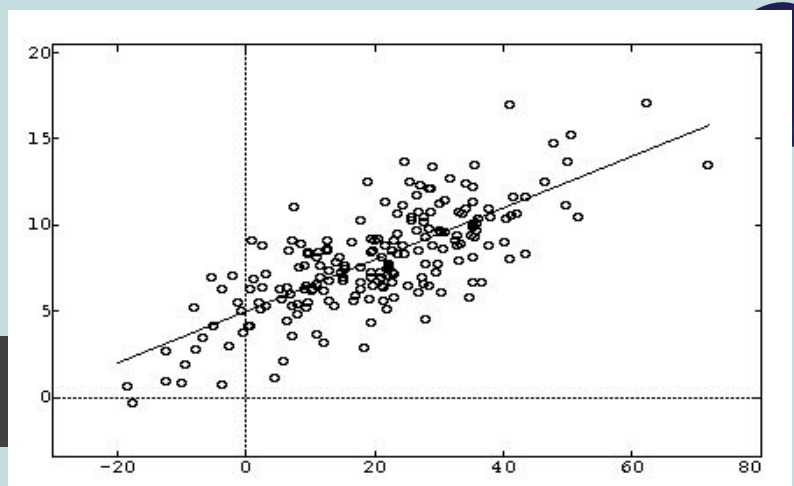
- Sklearn also provides tools to create features out of existing data structures
- **OneHotEncoder** - turns categorical data into binary values
  - Ex: Male or Female could have Female be 0 and Male be 1
- **OrdinalEncoder** - converts categorical data into numerical by assigning them a number by category
  - Ex: 1 for freshman, 2 for sophomore, etc.
- **CountVectorizer, TfidfVectorizer** - turn words into numerical data
  - Used for NLP, LLMs

https://scikit-learn.org/1.5/data_transforms.html

Some Models

- **LinearRegression -** fits a line through the independent and dependent variables
- **Logistic Regression** - Models the probability of a binary outcome using a logistic function to map inputs to probabilities
- **DecisionTree, RandomForest** - Decision Trees are like a choose-your-own-adventure game of conditions, while Random Forest combines many trees for better predictions

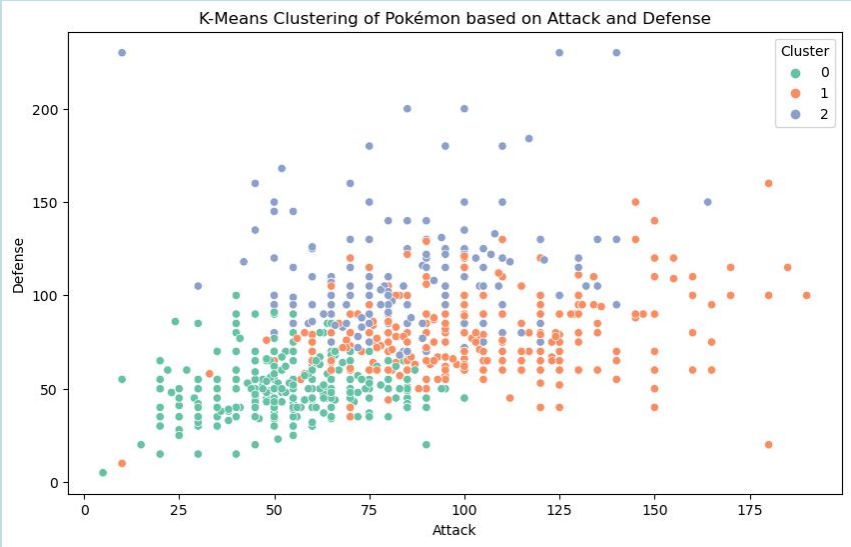




https://scikit-learn.org/stable/supervised_learning.html

More Models

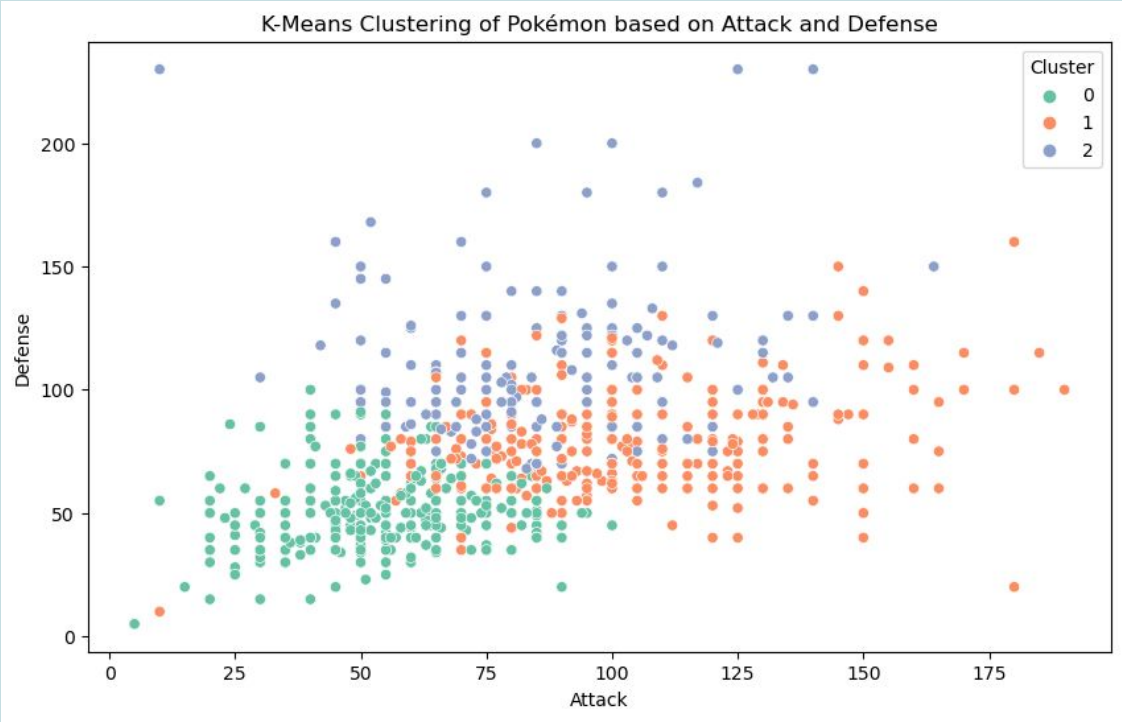
- **K-Nearest Neighbors (KNN)** - labels an input based on the majority vote of k-nearest labels closest to it
- **Support Vector Machines (SVM)** - Finds the optimal hyperplane to separate classes by maximizing the margin between them.
- **Naive Bayes** - Classifies data using Bayes' theorem, assuming all features are independent of each other



https://youtu.be/21u_pdYI7VI?feature=shared

Aside: K-Means Clustering

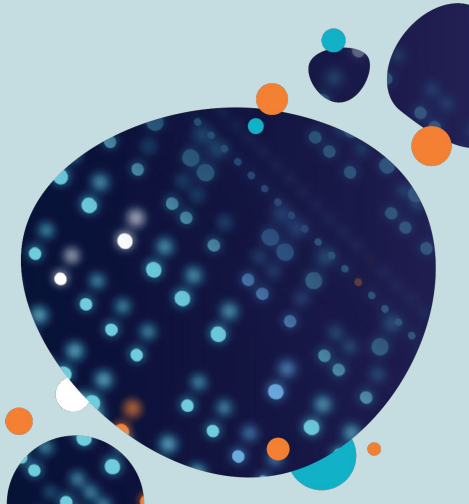
- **K-Means Clustering** - Groups similar items together into k-number of clusters
- See Intro to EDA Workshop for more information



Key Ideas

- Scikit-learn is a powerful library for machine learning
- It provides tools for feature preprocessing, feature engineering, and various machine learning models

https://github.com/couchsnail/ds3-workshops.git



data science student society

Leave your feedback here!