

TECH ALPHARETTA

2972 Webb Bridge Rd, Alpharetta, GA 30009

**Project #2: Data vs. Wild**

This project will throw a curveball at you. Many of the other challenges in this prep kit will be fairly lengthy, but in this one, speed and efficiency are of the essence! (You'll see why.)

Specifically, this project is about how quickly and efficiently you can (1) train an effective model, (2) spot potential issues with it, and (3) address those issues.

**Tip:**In data science, oftentimes *less is more*. One way you can tell the difference between an amateur and a professional is how directly they tackle a problem.

Amateurs will tend to try every data visualization they can think of, jump around from model to model, and end up with their code looking like a pile of dirty laundry.

Professionals try to identify *exactly* what they need to answer the objective and then avoid too many tangents. This keeps their code clean and error-proof.

**Background**

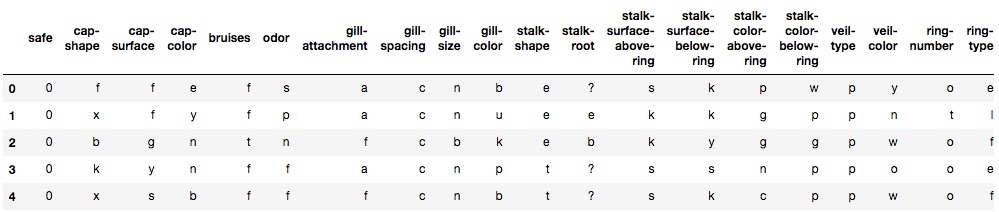
What do you do if you’re lost in the woods, you’re starving, and your only source of food is potentially poisonous mushrooms?

* Panic. Run around in circles. (*Outcome: run into an angry bear.*)
* Screw it. Grab the juiciest mushroom and dig in! (*Outcome: nasty food poisoning.*)
* Psh. Who needs to eat? (*Outcome: starve.*)
* None of the above.

Of course, we’re going with (D)! After all, we’re data scientists, right? And any good data scientist should carry 2 things at all times: A laptop and *a dataset of poisonous/non-poisonous mushrooms* :-)

**Data**

We have one table called **data\_vs\_wild.csv**, and it contains 1,000,000 observations from mushrooms in the wild, with information on features such as color and shape. Each one is labeled as safe or poisonous.



***Note****: This dataset was adapted from a dataset provided under the CC0 license*[*here*](https://archive.ics.uci.edu/ml/datasets/Mushroom)*. It has been significantly altered to better reflect what you’d see on a real take-home challenge. (In other words, do not expect any previous papers or scripts written for the original dataset to apply to this one.)*

***Data Dictionary:***

* **safe** – 1=safe to nom,0=poisonous
* **cap-shape** – bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
* **cap-surface** – fibrous=f,grooves=g,scaly=y,smooth=s
* **cap-color** – brown=n,buff=b,cinnamon=c,gray=g,green=r, pink=p,purple=u,red=e,white=w,yellow=y
* **bruises** – bruises=t,no=f
* **odor** – almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s
* **gill-attachment** – attached=a,descending=d,free=f,notched=n
* **gill-spacing** – close=c,crowded=w,distant=d
* **gill-size** – broad=b,narrow=n
* **gill-color** – black=k,brown=n,buff=b,chocolate=h,gray=g, green=r,orange=o,pink=p,purple=u,red=e, white=w,yellow=y
* **stalk-shape** – enlarging=e,tapering=t
* **stalk-root** – bulbous=b,club=c,cup=u,equal=e, rhizomorphs=z,rooted=r,missing=?
* **stalk-surface-above-ring** – fibrous=f,scaly=y,silky=k,smooth=s
* **stalk-surface-below-ring** – fibrous=f,scaly=y,silky=k,smooth=s
* **stalk-color-above-ring** – brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
* **stalk-color-below-ring** – brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
* **veil-type** – partial=p,universal=u
* **veil-color** – brown=n,orange=o,white=w,yellow=y
* **ring-number** – none=n,one=o,two=t
* **ring-type** – cobwebby=c,evanescent=e,flaring=f,large=l, none=n,pendant=p,sheathing=s,zone=z
* **spore-print-color** – black=k,brown=n,buff=b,chocolate=h,green=r, orange=o,purple=u,white=w,yellow=y
* **population** – abundant=a,clustered=c,numerous=n, scattered=s,several=v,solitary=y
* **habitat** – grasses=g,leaves=l,meadows=m,paths=p, urban=u,waste=w,woods=d

**Objectives**

* Train a model to predict whether a mushroom is poisonous or not. And do it quickly… (After all, we’re getting hungry!) This dataset has already been cleaned.