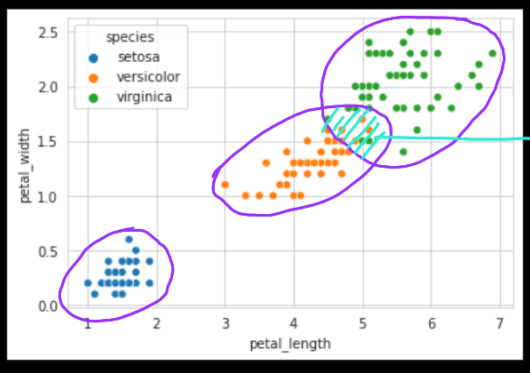
Introduction to probability and statistics:

First lets see why we need to understand probability and statistics

In the below plot we can classify the flowers into their class when there is no intersection.



→ But in this section we cannot exactly say weather the query point is Versicolor or virginica.

↓

But instead of saying which class x_q belong to we can say that x_q belongs to a class by some probility.

This is where we use probability in ML.

Probability: Probability is simply how likely something is to happen

In rolling dice

→ $P(x=1) = P(x=2) = P(x=3) = P(x=4) = P(x=5) = P(x=6) = \frac{1}{6}$

→ $P(x \text{ is even}) = 1/2$ → $P(x \text{ is odd}) = 1/2$

Random variable: A random variable is a numerical description of the outcome of statistical experiment.

For example r.v for an experiment of throwing a dice is

$X = \{1, 2, 3, 4, 5, 6\}$

r.v for tossing a coin $\Longrightarrow X = \{H, T\}$

→ If the variables are countable in a finite time then it is discrete variable and the outcomes of such experiments are **discrete random variable**

Ex: Rolling a dice. → finite set of outcomes

→ Continuous variables would literally take forever to count and the outcomes of such experiments are **continuous random variable.**

Ex: Measuring height of a student. → Infinite set of outcomes.

Outliers:

Y: height of a student

{122.2, 146.4, 132.5, ..., 12.26, 156.23}

All values of this R.V are in a range except this single value and this is an outlier

Outliers may occur due to many reasons

1. Human error
2. Observation errors
3. Genuine outliers and many more causes.

Outliers can corrupt your signal [mean and variance are impacted by outliers]