**1 – Introduction to IRIS dataset**

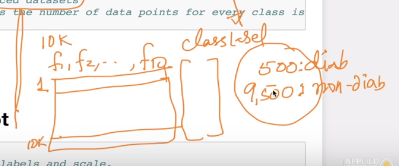
1. **Balnced v/s imbalanced datasets**

**Balanced datasets:** These datasets have approximately equal (not exactly, slight difference is considered) number of counts for each independent variables’ value.

Ex: for IRIS data set we have 50 values for each species (setossa, [versicolor](https://en.wikipedia.org/wiki/Iris_versicolor), [virginica](https://en.wikipedia.org/wiki/Iris_virginica))

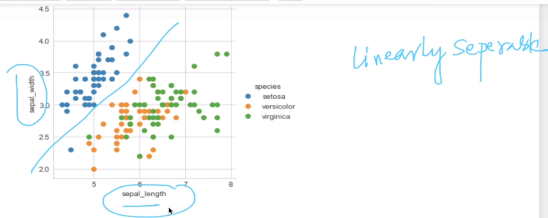
**Imbalanced datasets:** These datasets have equal number of counts for each independent variables’ value.

Ex: for hospital dataset there are total 10k vectors, and in that only 500 have diabetes and 9,500 don’t have diabetes.



1. **Linearly seperable:**

If data points can be seperable by just drawing a line then that data sets is known to be linearly seperable.



Here we can see blue points are separated from yellow and green so it is **linearly seperable.**

1. For  **Seaborn:** <https://jakevdp.github.io/PythonDataScienceHandbook/04.14-visualization-with-seaborn.html>
2. what is **hue** in seaborn ex: **hue="species"**

here the hue stands as the categorical feature based on which we want to saperate the scatter points. for eg. on the iris dataset we want to saperate the scatter points based on their species. we can also saperate them based on any other features like petal length/petal width etc

1. Limitation of **Pair plots**

Suppose if there are large number of features like 100, 200 then it will be very difficult to bring out observation from them, so in such datasets with large features we use **PCA, t-SNE.**