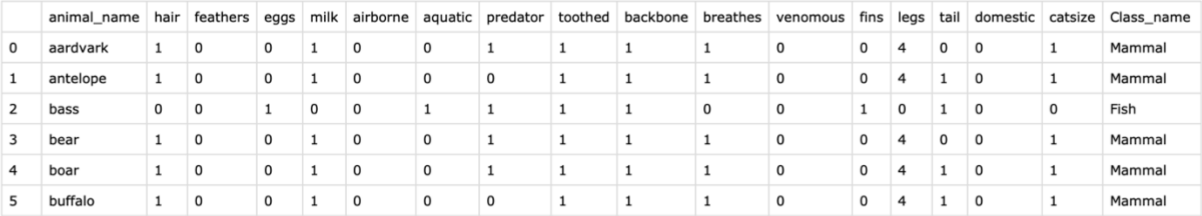
# Dataset selection

The "Zoo Animal Classification" dataset (5/14/1990) is used for the task of classifying animals into one of seven classes based on various features and attributes of the animals. The dataset contains 101 animals from the zoo, belonging to 7 classes of animals. The classes are: *Mammal, Bird, Reptile, Fish, Amphibian, Bug,* and *Invertebrate*. The data includes various attributes of the animals, such as whether they have hair, feathers, fins, eggs, milk and other characteristics. 15 binary and 3 categorical (animal_name, legs and Class_name) columns.



| | animal_name | hair | feathers | eggs | milk | airborne | aquatic | predator | toothed | backbone | breathes | venomous | fins | legs | tail | domestic | catsize | Class_name |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | aardvark | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 4 | 0 | 0 | 1 | Mammal |
| 1 | antelope | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 4 | 1 | 0 | 1 | Mammal |
| 2 | bass | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | Fish |
| 3 | bear | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 4 | 0 | 0 | 1 | Mammal |
| 4 | boar | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 4 | 1 | 0 | 1 | Mammal |
| 5 | buffalo | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 4 | 1 | 0 | 1 | Mammal |

*Table 1. First 5 rows form "Zoo Animal Classification" dataset*

The dataset typically comes in a .csv format and consists of rows and columns. Each row represents an animal and each column represents an attribute of the animal.

There is an interesting row in this dataset named girl and a twice repeated frog. I don't know if this is a mistake or on purpose))

For better usability, I replaced the last column “type”, which showed the class in numbers from 1 to 7 with a "Class_name" column, where each class name is shown as a word.

The "animal_name" attribute in next example’s data is specified only for better understanding of the characteristics and is NOT used in any way in the calculations.

(Source)

# Probabilistic classification

I have chosen, in my opinion, some of the most interesting examples for this classification. They demonstrate the features of their class perfectly. I have limited it to 5 attributes for a simpler and clearer calculation.

So let's apply
a naïve Bayes classifier with m-value constant m=2 to the test sample **x** = [1,0,0,2,1]:

| | Animal | Class | Hair | Feathers | Fins | Legs | Breathes |
|---|---|---|---|---|---|---|---|
| **1.** | Bear | Mammal | 1 | 0 | 0 | 4 | 1 |
| **2.** | Dolphin | Mammal | 0 | 0 | 1 | 0 | 1 |
| **3.** | Tuna | Fish | 0 | 0 | 1 | 0 | 0 |
| **4.** | Tortoise | Reptile | 0 | 0 | 0 | 4 | 1 |
| **5.** | Chicken | Bird | 0 | 1 | 0 | 2 | 1 |
| **6.** | Housefly | Bug | 1 | 0 | 0 | 6 | 1 |
| **7.** | Gnat | Bug | 0 | 0 | 0 | 6 | 1 |

*Table 2. Selected examples from the dataset for Probabilistic classification*

**1)** First, let's calculate the probability of choosing each class:

$$P(Mammal) = P(Bug) = \frac{2}{7}$$

$$P(Fish) = P(Reptile) = P(Bird) = \frac{1}{7}$$

**2)** Next, let's calculate the probabilities of having the required data in the example:

$$\boldsymbol{P(x|y) = \frac{n' + mp}{n + m}}$$

$\boldsymbol{n}$ - total number of examples in class *y*;
$\boldsymbol{n'}$ - total number of examples in class *y* that match our condition;
***p*** - the prior estimate of the probability;
***m*** - equivalent sample size, which determines how heavily to weight *p* relative to the observed data. (i.e. adding *m* "virtual" examples distributed according to *p*);

$$m = 2, \qquad p = [\frac{1}{2}, \frac{1}{4}]$$

*m* is just a constant, and p is the probability of choice in ***x*** -> there are only two choices [0,1] in $x_1, x_2, x_3, x_5$, so $p = \frac{1}{2}$, but in $x_4$ choices 4 = [0,2,4,6], so $p = \frac{1}{4}$:

$$P(x_1|y) = P(x_2|y) = P(x_3|y) = P(x_5|y) = \frac{n' + 2 * \frac{1}{2}}{n + 2} = \frac{n' + 1}{n + 2}$$

$$P(x_4|y) = \frac{n' + 2 * \frac{1}{4}}{n + 2} = \frac{n' + \frac{1}{2}}{n + 2}$$

**3)** Now we can compute the class probabilities for the test data:

| Class (y) | Hair ($x_1$) | Feathers ($x_2$) | Fins ($x_3$) | Legs ($x_4$) | Breathes ($x_5$) |
|---|---|---|---|---|---|
| ? | 1 | 0 | 0 | 2 | 1 |

*Table 3. Test Data*

$$P(x|Mammal) =$$
$$= P(x_1|Mammal) * P(x_2|Mammal) * P(x_3|Mammal) *$$
$$* P(x_4|Mammal) * P(x_5|Mammal) =$$
$$= \frac{1+1}{2+2} * \frac{2+1}{2+2} * \frac{1+1}{2+2} * \frac{0+\frac{1}{2}}{2+2} * \frac{2+1}{2+2} = \frac{\mathbf{9}}{\mathbf{152}} \approx 0{,}0175781$$

$$P(x|Fish) = P(x_1|Fish) * P(x_2|Fish) * P(x_3|Fish) * P(x_4|Fish) * P(x_5|Fish) =$$
$$= \frac{0+1}{1+2} * \frac{1+1}{1+2} * \frac{0+1}{1+2} * \frac{0+\frac{1}{2}}{1+2} * \frac{0+1}{1+2} = \frac{\mathbf{1}}{\mathbf{243}} \approx 0{,}004115$$

$$P(x|Reptile) = P(x_1|Reptile) * P(x_2|Reptile) * P(x_3|Reptile) * P(x_4|Reptile) *$$
$$* P(x_5|Reptile) = \frac{0+1}{1+2} * \frac{1+1}{1+2} * \frac{1+1}{1+2} * \frac{0+\frac{1}{2}}{1+2} * \frac{1+1}{1+2} = \frac{\mathbf{4}}{\mathbf{243}} \approx$$
$$\approx 0{,}0164609$$

$$P(x|Bird) = P(x_1|Bird) * P(x_2|Bird) * P(x_3|Bird) * P(x_4|Bird) * P(x_5|Bird) ==$$
$$= \frac{0+1}{1+2} * \frac{0+1}{1+2} * \frac{1+1}{1+2} * \frac{1+\frac{1}{2}}{1+2} * \frac{1+1}{1+2} = \frac{\mathbf{2}}{\mathbf{81}} \approx 0{,}24691358$$

$$P(x|Bug) = P(x_1|Bug) * P(x_2|Bug) * P(x_3|Bug) * P(x_4|Bug) * P(x_5|Bug) ==$$
$$= \frac{1+1}{2+2} * \frac{2+1}{2+2} * \frac{2+1}{2+2} * \frac{0+\frac{1}{2}}{2+2} * \frac{2+1}{2+2} = \frac{\mathbf{27}}{\mathbf{1024}} \approx 0{,}0263672$$

**4)** And here, we insert all the data we obtained earlier into the Naïve Bayes formula:

$$P(y|x) = \frac{P(y) * P(x|y)}{\sum_{v \in V} P(v)P(x|v)}$$

*V = {Mammal, Fish, Reptile, Bird ,Bug}*

$$P(Mammal|x) = \frac{P(Mammal) * P(x|Mammal)}{\sum_{v \in V} P(v)P(x|v)} =$$

$$= \frac{\frac{2}{7} * \frac{9}{152}}{\frac{2}{7} * \frac{9}{152} + \frac{1}{7} * \frac{1}{234} + \frac{1}{7} * \frac{4}{243} + \frac{1}{7} * \frac{2}{81} + \frac{2}{7} * \frac{27}{1024}} = \frac{3639168}{6655703} \approx 0{,}55$$

$$P(Fish|x) = \frac{P(Fish) * P(x|Fish)}{\sum_{v \in V} P(v)P(x|v)} =$$

$$= \frac{\frac{1}{7} * \frac{1}{234}}{\frac{2}{7} * \frac{9}{152} + \frac{1}{7} * \frac{1}{234} + \frac{1}{7} * \frac{4}{243} + \frac{1}{7} * \frac{2}{81} + \frac{2}{7} * \frac{27}{1024}} = \frac{131328}{6655703} \approx 0{,}02$$

$$P(Reptile|x) = \frac{P(Reptile) * P(x|Reptile)}{\sum_{v \in V} P(v)P(x|v)} =$$

$$= \frac{\frac{1}{7} * \frac{4}{243}}{\frac{2}{7} * \frac{9}{152} + \frac{1}{7} * \frac{1}{234} + \frac{1}{7} * \frac{4}{243} + \frac{1}{7} * \frac{2}{81} + \frac{2}{7} * \frac{27}{1024}} = \frac{505856}{6655703} \approx 0{,}08$$

$$P(Bird|x) = \frac{P(Bird\,) * P(x|Bird\,)}{\sum_{v \in V} P(v)P(x|v)} =$$

$$= \frac{\frac{1}{7} * \frac{2}{81}}{\frac{2}{7} * \frac{9}{152} + \frac{1}{7} * \frac{1}{234} + \frac{1}{7} * \frac{4}{243} + \frac{1}{7} * \frac{2}{81} + \frac{2}{7} * \frac{27}{1024}} = \frac{758784}{6655703} \approx 0{,}11$$

$$P(Bug|x) = \frac{P(Bug) * P(x|Bug\,)}{\sum_{v \in V} P(v)P(x|v)} =$$

$$= \frac{\frac{2}{7} * \frac{27}{1024}}{\frac{2}{7} * \frac{9}{152} + \frac{1}{7} * \frac{1}{234} + \frac{1}{7} * \frac{4}{243} + \frac{1}{7} * \frac{2}{81} + \frac{2}{7} * \frac{27}{1024}} = \frac{1620567}{6655703} \approx 0{,}24$$

**5)** And finally, let's summarize:

| Class | Probability (P(y\|x) * 100) |
|---|---|
| Mammal | 55% |
| Fish | 2% |
| Reptile | 8% |
| Bird | 11% |
| Bug | 24% |
| **Summary** | **100%** |

*Table 4. Summary table*

As we can see from the results in the table on the left, the sample turned out to be 55% - Mammal. To be honest, when I was guessing the data, I was thinking about *kangaroos*, so I'm satisfied with the results.

# Decision tree

For the next task, I have chosen 10 different animals with 7 attributes-characteristics. Attribute "Predator" - will be the target of this task. So, using the ID3 algorithm, lets build a decision tree:

| | Animal name | Hair | Feathers | Eggs | Milk | Airborne | Aquatic | Predator |
|---|---|---|---|---|---|---|---|---|
| 1. | **Elephant** | 1 | 0 | 0 | 1 | 0 | 0 | **0** |
| 2. | **Giraffe** | 1 | 0 | 0 | 1 | 0 | 0 | **0** |
| 3. | **Dolphin** | 0 | 0 | 0 | 1 | 0 | 1 | **1** |
| 4. | **Lark** | 0 | 1 | 1 | 0 | 1 | 0 | **0** |
| 5. | **Octopus** | 0 | 0 | 1 | 0 | 0 | 1 | **1** |
| 6. | **Penguin** | 0 | 1 | 1 | 0 | 0 | 1 | **1** |
| 7. | **Honeybee** | 1 | 0 | 1 | 0 | 1 | 0 | **0** |
| 8. | **Tuatara** | 0 | 0 | 1 | 0 | 0 | 0 | **1** |
| 9. | **Flamingo** | 0 | 1 | 1 | 0 | 1 | 0 | **0** |
| 10. | **Antelope** | 1 | 0 | 0 | 1 | 0 | 0 | **0** |

*Table 5. Dataset "Zoo"*

1) What is Entropy and who is Gain?

*Entropy* comes from information theory.
The higher the entropy the more the information content.

$$Entropy\ H(S) = \sum_{x \in X} - p(x) \log_2 p(x)$$

***S*** - the data set for which the entropy is calculated.
***X*** - the set of classes in *S*.
***p (x)*** - the ratio of the number of elements in class *x* to the number of elements in the set *S*.

Then we want to determine which attribute in a given set of training feature vectors is most useful for discriminating between the classes to be learned.
*Information gain* tells us how important a given attribute of the feature vectors is.

$$Gain = Entropy(S) - \frac{Attribute\ Y's\ Instaces}{All\ Instances} Entropy(S_{Attribute\ Y})$$

2) Let's find Predator Entropy:

We have 4 positive(1) and 6 negative (0) decisions in "Predator" column, so put them in P:
**P** = [4+; 6-]

$$Entropy\ (P) = -\frac{4}{4+6}\log_2\frac{4}{4+6} - \frac{6}{4+6}\log_2\frac{6}{4+6} \approx \mathbf{0,970951}$$

3) Now, we need to find entropy and gain of each attribute:

Attribute: Hair -> {0,1}

$$P_{(Hair=1)} = [\,0+;4\,-]$$

$$Entropy\left(P_{(Hair=1)}\right) = -\frac{0}{0+4}\log_2\frac{0}{0+4} - \frac{4}{0+4}\log_2\frac{4}{0+4} = \mathbf{0}$$

$$P_{(Hair=0)} = [\,4+;2-]$$

$$Entropy\left(P_{(Hair=0)}\right) = -\frac{4}{4+2}\log_2\frac{4}{4+2} - \frac{2}{2+4}\log_2\frac{2}{2+4} = \mathbf{0,918296}$$

$$\begin{aligned}&Gain\,(P, Hair) = \\ &= Entropy\,(P) - \frac{0+4}{10}Entropy\left(P_{(Hair=1)}\right) - \frac{4+2}{10}Entropy\left(P_{(Hair=0)}\right) \\ &= 0{,}970951 - \frac{4}{10}*0 - \frac{6}{10}*0{,}918296 \approx \mathbf{0,419973}\end{aligned}$$

Attribute: Feathers -> {0,1}

$$P_{(\text{Feathers } = 1)} = [\,1+;2\,-]$$

$$Entropy\left(P_{(\text{Feathers } = 1)}\right) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = \mathbf{0,918296}$$

$$P_{(\text{Feathers}=0)} = [\,3+;4-]$$

$$Entropy\left(P_{(\text{Feathers}=0)}\right) = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = \mathbf{0,985228}$$

$$\begin{aligned}&Gain\,(P, \text{Feathers}) = \\ &= Entropy\,(P) - \frac{3}{10}Entropy\left(P_{(\text{Feathers}=1)}\right) - \frac{7}{10}Entropy\left(P_{(\text{Feathers}=0)}\right) \\ &= 0{,}970951 - \frac{3}{10}*0{,}918296 - \frac{7}{10}*0{,}985228 \approx \mathbf{0,0058026}\end{aligned}$$

Attribute: Egg -> {0,1}

$$P_{(\mathrm{Egg=1})} = [\,3+; 3\,-]$$

$$Entropy\,\left(P_{(\mathrm{Egg=1})}\right) = \,-\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} \,= \mathbf{1}$$

$$P_{(\mathrm{Egg=0})} = [\,1+; 3-]$$

$$Entropy\,\left(P_{(\mathrm{Egg=0})}\right) = \,-\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} \,= \mathbf{0,811278}$$

$$Gain\,(P, \mathrm{Egg}) = Entropy\,(P) - \frac{6}{10}Entropy\left(P_{(\mathrm{Egg=1})}\right) - \frac{4}{10}Entropy\left(P_{(\mathrm{Egg=0})}\right)$$
$$= 0{,}970951 - \,\frac{6}{10} * 1 - \,\frac{4}{10} * 0{,}811278 \,\approx \mathbf{0,0464398}$$

Attribute: Milk -> {0,1}

$$P_{(\mathrm{Milk\,=1})} = [\,1+; 3\,-]$$

$$Entropy\,\left(P_{(\mathrm{Milk\,=1})}\right) = \,-\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} \,= \mathbf{0,811278}$$

$$P_{(\mathrm{Milk=0})} = [\,3+; 3-]$$

$$Entropy\,\left(P_{(\mathrm{Milk=0})}\right) = \,-\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} \,= \mathbf{1}$$

$$Gain\,(P, \mathrm{Milk}) = Entropy\,(P) - \frac{4}{10}Entropy\left(P_{(\mathrm{Milk=1})}\right) - \frac{6}{10}Entropy\left(P_{(\mathrm{Milk=0})}\right) =$$
$$= 0{,}970951 - \,\frac{4}{10} * 0{,}811278 - \,\frac{6}{10} * 1 \,\approx \mathbf{0,0464398}$$

Attribute: Airborne -> {0,1}

$$P_{(\text{Airborne}=1)} = [\,0+;3\,-]$$

$$Entropy\,\left(P_{(\text{Airborne}=1)}\right) = \;-\frac{0}{3}\log_2\frac{0}{3} - \frac{3}{3}\log_2\frac{3}{3} \;= \mathbf{0}$$

$$P_{(\text{Airborne}=0)} = [\,4+;3-]$$

$$Entropy\,\left(P_{(\text{Airborne}=0)}\right) = \;-\frac{4}{7}\log_2\frac{4}{7} - \frac{3}{7}\log_2\frac{3}{7} \;= \mathbf{0,985228}$$

$$\begin{aligned}&Gain\;(P,\text{Airborne}) =\\&\quad= Entropy\;(P) - \frac{3}{10}Entropy\left(P_{(\text{Airborne}=1)}\right) - \frac{7}{10}Entropy\left(P_{(\text{Airborne}=0)}\right)\\&\quad= 0{,}970951 - \frac{3}{10} * 0 - \frac{7}{10} * 0{,}985228 \approx \mathbf{0,281291}\end{aligned}$$

Attribute: Aquatic -> {0,1}

$$P_{(Aquatic=1)} = [\,3+;0\,-]$$

$$Entropy\,\left(P_{(\text{Aquatic}=1)}\right) = \;-\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3} \;= \mathbf{0}$$

$$P_{(\text{Aquatic}=0)} = [\,1+;6-]$$

$$Entropy\,\left(P_{(\text{Aquatic}=0)}\right) = \;-\frac{1}{7}\log_2\frac{1}{7} - \frac{6}{7}\log_2\frac{6}{7} \;= \mathbf{0,591673}$$

$$\begin{aligned}&Gain\;(P,\text{Aquatic}) =\\&\quad= Entropy\;(P) - \frac{3}{10}Entropy\left(P_{(\text{Aquatic}=1)}\right) - \frac{7}{10}Entropy\left(P_{(\text{Aquatic}=0)}\right) =\\&\quad= 0{,}970951 - \frac{3}{10} * 0 - \frac{7}{10} * 0{,}591673 \;\approx \mathbf{0,55678}\end{aligned}$$

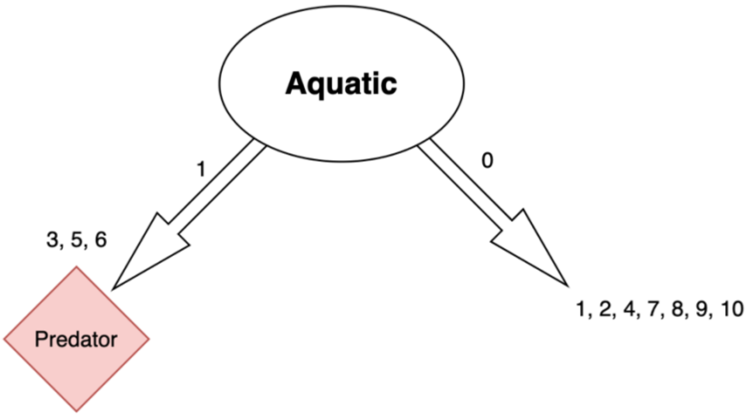
4) Summarize all gains:

| Attribute | Value |
|---|---|
| $Gain\ (P, Hair)$ | 0,419 |
| $Gain\ (P, Feathers)$ | 0,006 |
| $Gain\ (P, Egg)$ | 0,046 |
| $Gain\ (P, Milk)$ | 0,046 |
| $Gain\ (P, \text{Airborn})$ | 0,281 |
| $Gain\ (P, \text{Aquatic})$ | $\mathbf{0,557}$ |

*Table 6. Summary Table*

From the table on the left, we can see that *Gain(P, Aquatic)* has the most weight on our data, so the tree will start with it.

5) Separate Aquatic:



| | Animal name | Hair | Feathers | Eggs | Milk | Airborne | Predator |
|---|---|---|---|---|---|---|---|
| 3. | **Dolphin** | 0 | 0 | 0 | 1 | 0 | **1** |
| 5. | **Octopus** | 0 | 0 | 1 | 0 | 0 | **1** |
| 6. | **Penguin** | 0 | 1 | 1 | 0 | 0 | **1** |

*Table 7. Predators found*

As we can see, all 3 samples are predators, so this branch ( with 3,5,6) ends here (tab. 7). We have found 3 predators and the last one remains, so we continue to work with the existing examples (1,2,4,7,8,9,10). Column "Aquatic" we can delete (tab. 8).

6) Continue calculation:

| | Animal name | Hair | Feathers | Eggs | Milk | Airborne | Predator |
|---|---|---|---|---|---|---|---|
| 1. | **Elephant** | 1 | 0 | 0 | 1 | 0 | **0** |
| 2. | **Giraffe** | 1 | 0 | 0 | 1 | 0 | **0** |
| 4. | **Lark** | 0 | 1 | 1 | 0 | 1 | **0** |
| 7. | **Honeybee** | 1 | 0 | 1 | 0 | 1 | **0** |
| 8. | **Tuatara** | 0 | 0 | 1 | 0 | 0 | **1** |
| 9. | **Flamingo** | 0 | 1 | 1 | 0 | 1 | **0** |
| 10. | **Antelope** | 1 | 0 | 0 | 1 | 0 | **0** |

*Table 8. Remaining data*

Attribute: Hair -> {0,1}

$$P_{(Hair=1)} = [\,0+;4\,-]$$

$$Entropy\,\left(P_{(Hair=1)}\right) = \,-\frac{0}{0+4}\log_2\frac{0}{0+4}\,-\frac{4}{0+4}\log_2\frac{4}{0+4}\,=\mathbf{0}$$

$$P_{(Hair=0)} = [\,1+;2-]$$

$$Entropy\,\left(P_{(Hair=0)}\right) = \,-\frac{1}{3}\log_2\frac{1}{3}-\frac{2}{3}\log_2\frac{2}{3}\,=\mathbf{0,918296}$$

$$Gain\,(P,Hair) == Entropy\,(P)-\frac{4}{7}Entropy\left(P_{(Hair=1)}\right)-\frac{3}{7}Entropy\left(P_{(Hair=0)}\right)$$
$$=0{,}970951-\,\frac{4}{7}*0-\,\frac{3}{7}*0{,}918296\,\approx\mathbf{0,577396}$$

Attribute: Feathers -> {0,1}

$$P_{(\text{Feathers }=1)} = [\,0+;2\,-]$$

$$Entropy\,\left(P_{(\text{Feathers }=1)}\right) = \,-\frac{0}{2}\log_2\frac{0}{2}\,-\frac{2}{2}\log_2\frac{2}{2}\,=\mathbf{0}$$

$$P_{(\text{Feathers}=0)} = [\,1+;4-]$$

$$Entropy\,\left(P_{(\text{Feathers}=0)}\right) = \,-\frac{1}{5}\log_2\frac{1}{5}\,-\frac{4}{5}\log_2\frac{4}{5}\,=\mathbf{0,721928}$$

$$Gain\,(P,\text{Feathers}) =$$
$$=Entropy\,(P)-\frac{2}{7}Entropy\left(P_{(\text{Feathers}=1)}\right)-\frac{5}{7}Entropy\left(P_{(\text{Feathers}=0)}\right)$$
$$=0{,}970951-\,\frac{2}{7}*0-\,\frac{5}{7}*0{,}721928\approx\mathbf{0,455288}$$

Attribute: Egg -> {0,1}

$$P_{(\text{Egg}=1)} = [0+; 3-]$$

$$Entropy\ (P_{(\text{Egg}=1)}) = -\frac{0}{3}\log_2\frac{0}{3} - \frac{3}{3}\log_2\frac{3}{3} = \mathbf{0}$$

$$P_{(\text{Egg}=0)} = [\ 1+; 3-]$$

$$Entropy\ (P_{(\text{Egg}=0)}) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} = \mathbf{0{,}811278}$$

$$Gain\ (P, \text{Egg}) = Entropy\ (P) - \frac{3}{7}Entropy(P_{(\text{Egg}=1)}) - \frac{4}{7}Entropy(P_{(\text{Egg}=0)})$$
$$= 0{,}970951 - \frac{3}{7} * 0 - \frac{4}{7} * 0{,}811278 \approx \mathbf{0{,}507364}$$

Attribute: Milk -> {0,1}

$$P_{(\text{Milk}=1)} = [\ 0+; 3-]$$

$$Entropy\ (P_{(\text{Milk}=1)}) = -\frac{0}{3}\log_2\frac{0}{3} - \frac{3}{3}\log_2\frac{3}{3} = \mathbf{0}$$

$$P_{(\text{Milk}=0)} = [\ 1+; 3-]$$

$$Entropy\ (P_{(\text{Milk}=0)}) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} = \mathbf{0{,}811278}$$

$$Gain\ (P, \text{Milk}) = Entropy\ (P) - \frac{3}{7}Entropy(P_{(\text{Milk}=1)}) - \frac{4}{7}Entropy(P_{(\text{Milk}=0)}) =$$
$$= 0{,}970951 - \frac{3}{7} * 0 - \frac{4}{7} * 0{,}811278 \approx \mathbf{0{,}507364}$$

Attribute: Airborne -> {0,1}

$$P_{(\text{Airborne}=1)} = [\,0+;3\,-]$$

$$Entropy\left(P_{(\text{Airborne}=1)}\right) = -\frac{0}{3}\log_2\frac{0}{3} - \frac{3}{3}\log_2\frac{3}{3} = \mathbf{0}$$

$$P_{(\text{Airborne}=0)} = [\,1+;3-]$$

$$Entropy\left(P_{(\text{Airborne}=0)}\right) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} = \mathbf{0{,}811287}$$

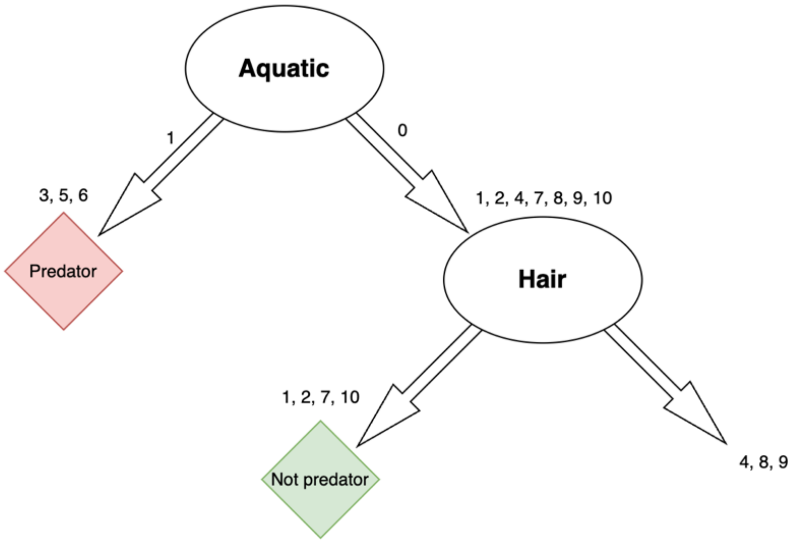
$$\begin{aligned} &Gain\ (P, \text{Airborne}) = \\ &\quad = Entropy\ (P) - \frac{3}{7}Entropy\left(P_{(\text{Airborne}=1)}\right) - \frac{4}{7}Entropy\left(P_{(\text{Airborne}=0)}\right) \\ &\quad = 0{,}970951 - \frac{3}{7} * 0 - \frac{4}{7} * 0{,}811287 \approx \mathbf{0{,}507358} \end{aligned}$$

7) Summarize all gains:

| Attribute | Value |
|---|---|
| $\boldsymbol{Gain\ (P, Hair)}$ | $\mathbf{0{,}57739}$ |
| $\boldsymbol{Gain\ (P, Feathers)}$ | 0,45528 |
| $\boldsymbol{Gain\ (P, Egg)}$ | 0,50736 |
| $\boldsymbol{Gain\ (P, Milk)}$ | 0,50736 |
| $\boldsymbol{Gain\ (P, \text{Airborn})}$ | 0,50735 |

*Table 9. Summary table*

Now, from the table on the left, we can see that *Gain(P, Hair)* has the most weight on our data, so the tree will continues with it and separate all examples.



8) After last iteration, we obtain next data:

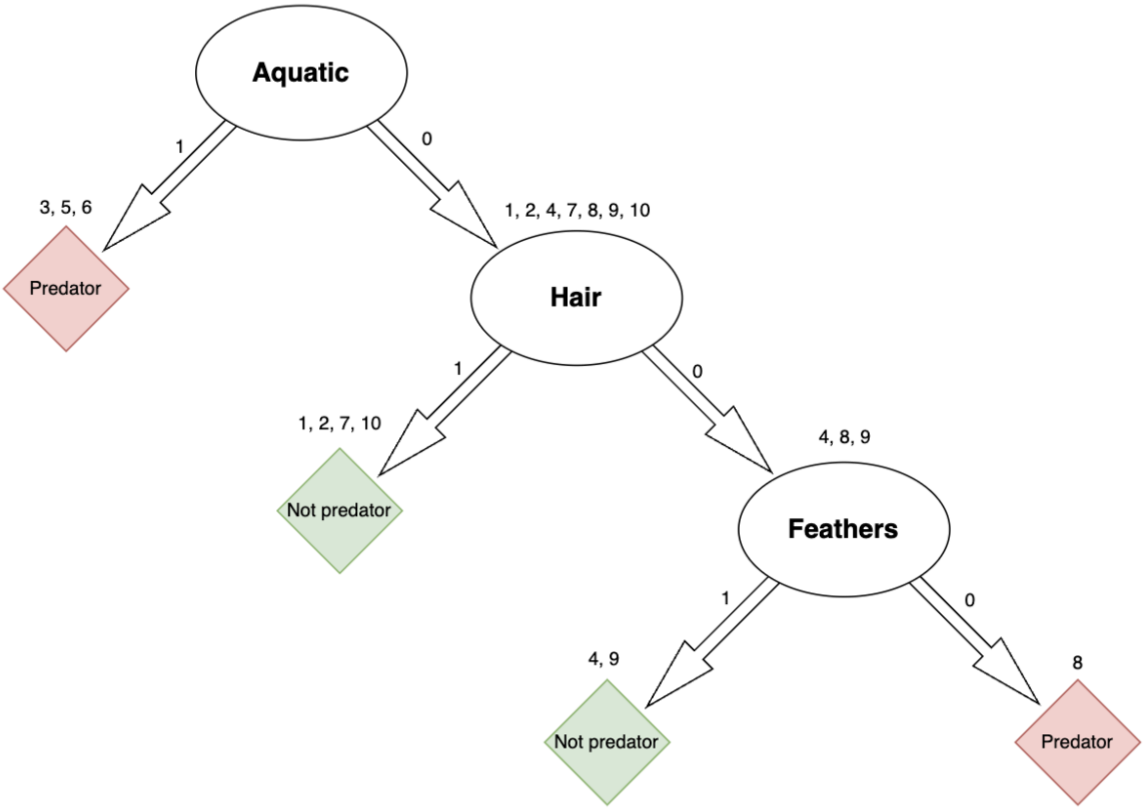
| | Animal name | Feathers | Eggs | Milk | Airborne | Predator |
|---|---|---|---|---|---|---|
| 4. | **Lark** | 1 | 1 | 0 | 1 | **0** |
| 8. | **Tuatara** | 0 | 1 | 0 | 0 | **1** |
| 9. | **Flamingo** | 1 | 1 | 0 | 1 | **0** |

At this point, I removed the "egg" and "milk" columns because they are identical and have no impact on the decision.

And finally, we have obtained data that can clearly identify the predator:

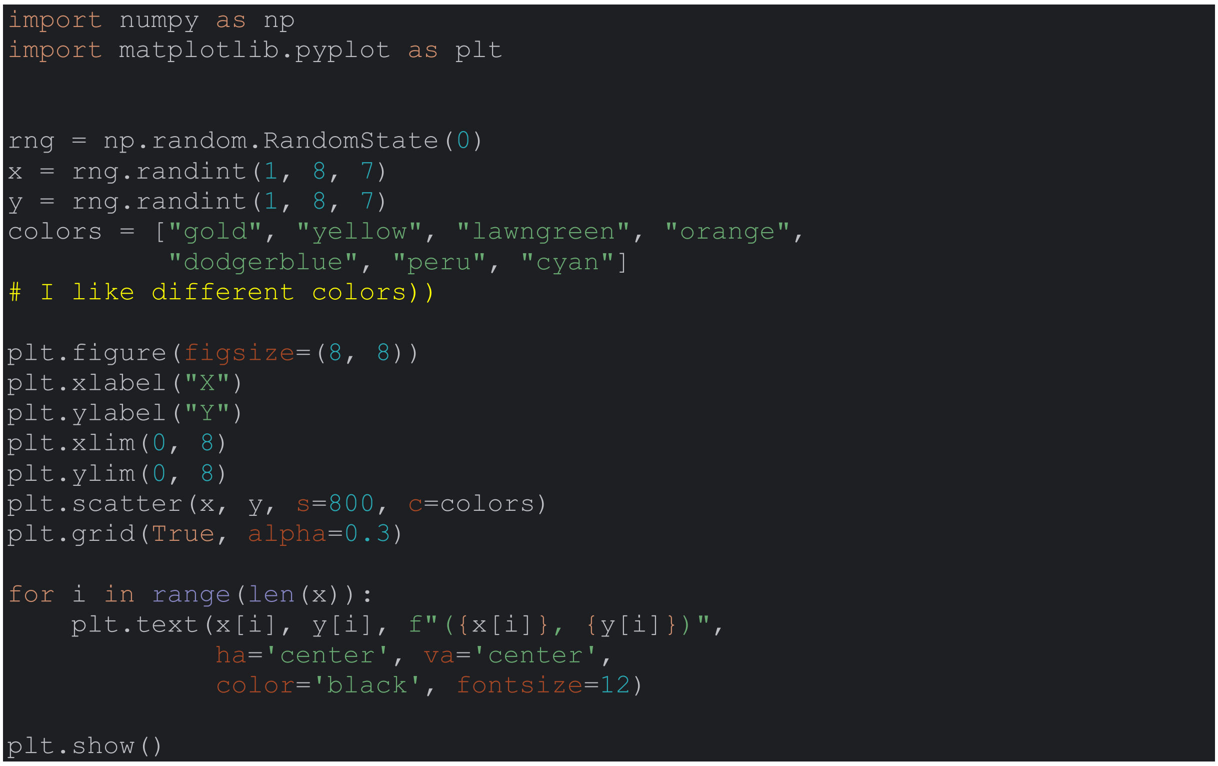
| | Animal name | Feathers | Airborne | Predator |
|---|---|---|---|---|
| 4. | **Lark** | 1 | 1 | **0** |
| 8. | **Tuatara** | 0 | 0 | **1** |
| 9. | **Flamingo** | 1 | 1 | **0** |

And here we can choose the next branch as "feathers" or "airborne" - it doesn't matter, I chose "feathers":



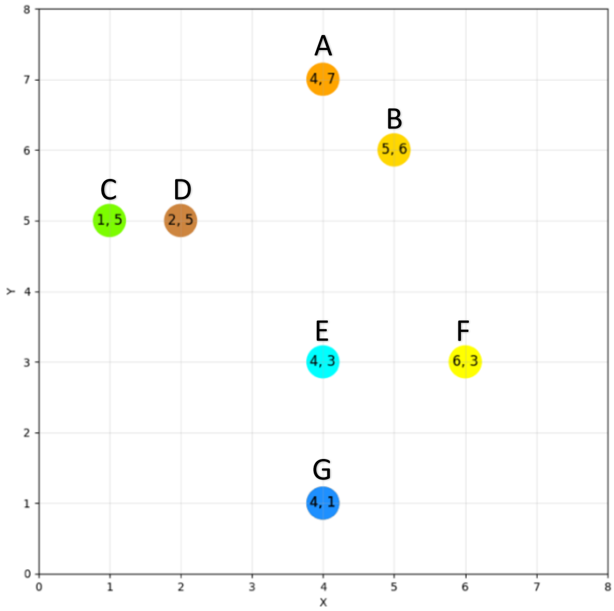
# Clustering

According to the following code, generate 7 random points and cluster them using the k-mean method, where **k = 3**:



```
import numpy as np
import matplotlib.pyplot as plt


rng = np.random.RandomState(0)
x = rng.randint(1, 8, 7)
y = rng.randint(1, 8, 7)
colors = ["gold", "yellow", "lawngreen", "orange",
          "dodgerblue", "peru", "cyan"]
# I like different colors))

plt.figure(figsize=(8, 8))
plt.xlabel("X")
plt.ylabel("Y")
plt.xlim(0, 8)
plt.ylim(0, 8)
plt.scatter(x, y, s=800, c=colors)
plt.grid(True, alpha=0.3)

for i in range(len(x)):
    plt.text(x[i], y[i], f"({x[i]}, {y[i]})",
             ha='center', va='center',
             color='black', fontsize=12)

plt.show()
```

Next output:



| POINT | AXES |
|---|---|
| **A** | (4; 7) |
| **B** | (5; 6) |
| **C** | (1; 5) |
| **D** | (2; 5) |
| **E** | (4; 3) |
| **F** | (6; 3) |
| **G** | (4; 1) |

| MEAN | AXES |
|---|---|
| **A** | (4; 7) |
| **C** | (1; 5) |
| **E** | (4; 3) |

**Randomly** select 3 (*k*) points.

Since we are given 7 points, the number of iterations $i$ will be also 7. During each iteration, we have to calculate the distance to each $mean_j$, where $j$ = (1:k). We will do the calculations using following formulas:

$$\begin{array}{cc} \boldsymbol{point_i} & \boldsymbol{mean_j} \\ (x_i, y_i) & (x_j, y_j) \end{array}$$

$$\rho(point_i, mean_j) = |x_j - x_i| + |y_j - y_i|$$

$\rho(point_A, mean_1) = |x_1 - x_A| + |x_1 - y_A| = |4 - 4| + |7 - 7| = 0 + 0 = \boldsymbol{0}$
(4; 7) (4; 7)
$\rho(point_A, mean_2) = |x_2 - x_A| + |x_2 - y_A| = |1 - 4| + |5 - 7| = 3 + 2 = \boldsymbol{5}$
(4; 7) (1; 5)
$\rho(point_A, mean_3) = |x_3 - x_A| + |y_3 - y_A| = |4 - 4| + |3 - 7| = 0 + 4 = \boldsymbol{4}$
(4; 7) (4; 3)

Let A = 1, C = 2, E = 3 -cluster, in this case, since the smallest number in [0, 5, 4] is 0 and A is 1st cluster, this point is located nearest to A, so it belongs to the first cluster:

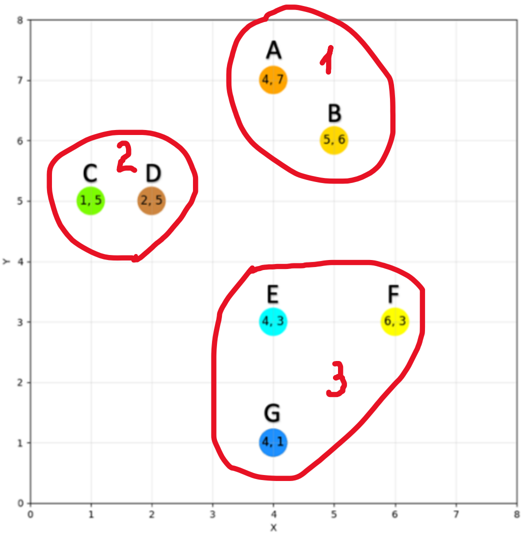
| | POINTS | A (4; 7) | C (1; 5) | E (4; 3) | CLUSTER |
|---|---|---|---|---|---|
| A | (4; 7) | **0** | 5 | 4 | 1 |

By applying this operation to all points, we get:

| | POINTS | A (4; 7) | C (1; 5) | E (4; 3) | CLUSTER |
|---|---|---|---|---|---|
| A | (4; 7) | **0** | 5 | 4 | 1 |
| B | (5; 6) | **2** | 5 | 4 | 1 |
| C | (1; 5) | 5 | **0** | 5 | 2 |
| D | (2; 5) | 4 | **1** | 4 | 2 |
| E | (4; 3) | 4 | 5 | **0** | 3 |
| F | (6; 3) | 6 | 7 | **2** | 3 |
| G | (4; 1) | 6 | 7 | **2** | 3 |

As we can see, 3 clusters have formed around 3 means[A,C,E], which contain all the nearest points.

The next step will be to find the new cluster centers (mean values). We do this by taking the average of all the points in each cluster.



For 1st cluster, we have 2 points A (4; 7), B (5; 6)

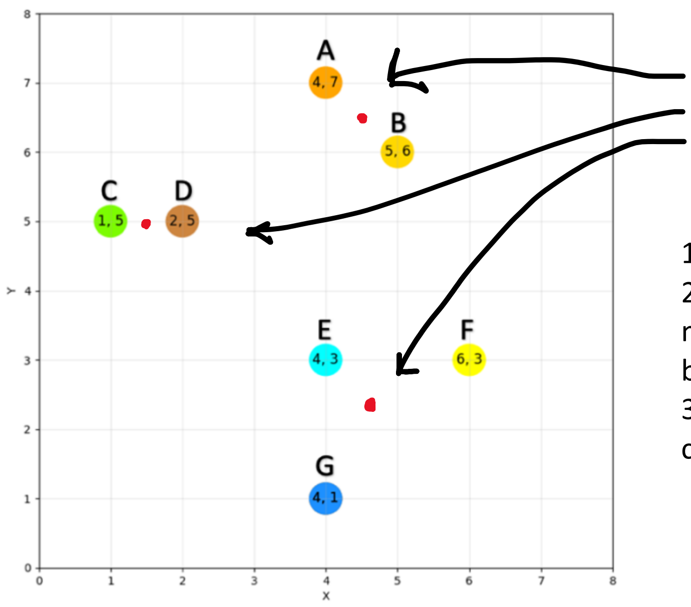
-> new 1st cluster center = ( (4+5)/2; (7+6)/2 ) = **(4,5; 6,5)**

For 2nd cluster, we have 2 points C (1; 5), D (2; 5) also

-> new 2nd cluster center = ( (1+2)/2; (5+5)/2 ) = **(1,5; 5)**

For 3rd cluster 3, we have 3 points E (4; 3), F (6; 3), G (4; 1)

-> new 3rd cluster center = ( (4+6+4)/3; (3+3+1)/3 ) = **(4,7; 2,3)**

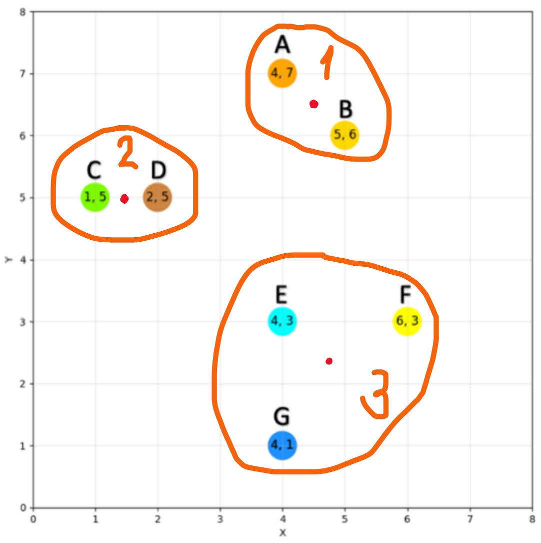


| CLUSTER | AXES |
|---|---|
| 1st | (4,5; 6,5) |
| 2nd | (1,5; 5) |
| 3rd | (4,7; 2,3) |

1. Placing new centers (red dots).
2. Doing the whole calculation mentioned at the beginning again, but with new mean centers.
3. Repeat the first two steps while changes are still happening.

Here are the new values and the result of the second iteration:

| | POINTS | (4,5; 6,5) | (1,5; 5) | (4,7; 2,3) | CLUSTER |
|---|---|---|---|---|---|
| A | (4; 7) | **1** | 4,5 | 5,4 | 1 |
| B | (5; 6) | **1** | 4,5 | 4 | 1 |
| C | (1; 5) | 5 | **0,5** | 6,4 | 2 |
| D | (2; 5) | 4 | **0,5** | 5,4 | 2 |
| E | (4; 3) | 4 | 4,5 | **1,4** | 3 |
| F | (6; 3) | 5 | 5,5 | **3** | 3 |
| G | (4; 1) | 6 | 6,5 | **2** | 3 |



Since the area of the current and previous iteration coincide, we can complete our clustering and observe the following result:

# Conclusions

**Data mining:**
Data mining involves finding patterns and relationships in large data sets to extract valuable information and make informed decisions.
**Naive Bayes Classification:**
Naive Bayes is a simple probabilistic classifier that predicts the likelihood of a given outcome based on available attributes, assuming that the factors are conditionally independent
**ID3 algorithm:**
ID3 is a decision tree algorithm used for classification that selects the best attributes to classify data sets based on the information received, thus creating a tree-like model for decision making.
**Clustering:**
Clustering is a method of grouping similar data points together based on certain characteristics, with the aim of identifying underlying patterns or patterns in the data without pre-defined categories.

Things that we have become used to, such as spam sorting in email or the world-famous interactive game Akinator, have long been based on these algorithms. Imagine my wonder when I caught myself thinking about this while writing this work.

For me, this work was quite interesting. I learned a lot about how the artificial intelligence algorithms mentioned above work. I invested a fair amount of time in this work. It was difficult at the beginning, but when I read a lot of theoretical information mentioned in the resources, I became very interested in implementing these methods in my dataset. The most tedious and time-consuming part was the calculations for each section, especially for the decision tree, but it was worth it. I liked the teacher's approach to these tasks, especially how we students should do these tasks, because this is how we really learn the material. Now I feel like a Guru in these topics.

# Resources

Dataset:

1. https://www.kaggle.com/datasets/uciml/zoo-animal-classification/

Naïve Bayes:

2. Classification slides, UMA, Campus Virtual, Intelligent Systems, Ezequiel López-Rubio and Enrique Domínguez - Link

3. https://stackoverflow.com/questions/34314277/what-should-be-taken-as-m-in-m-estimate-of-probability-in-naive-bayes

4. https://www2.cs.uh.edu/~arjun/courses/nlp/naive_bayes_keller.pdf

Iterative Dichotomiser 3:

5. Decision tree slides, UMA, Campus Virtual, Intelligent Systems, Ezequiel López-Rubio and Enrique Domínguez - Link

6. https://homes.cs.washington.edu/~shapiro/EE596/notes/InfoGain.pdf

Clustering:

7. Clustering slides, K-Means Clustering – Example, UMA, Campus Virtual, Intelligent Systems, Ezequiel López-Rubio and Enrique Domínguez - Link