# Classification Over Small Datasets with Increasing Cardinality

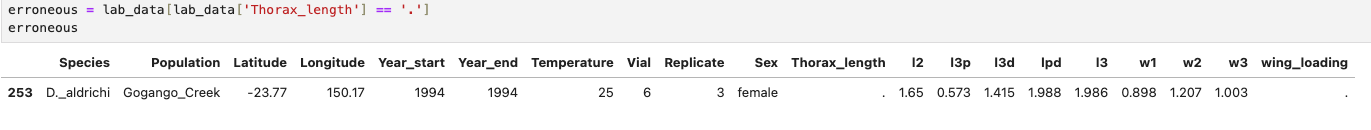
#### Brook Queree 48423384

## Data Exploration

TODO

## Data Cleaning

There is a singular row in the 83\_Loeschcke dataset that contains neither a Thorax\_length value, nor a wing\_loading value.Both simply contain the character “.”.



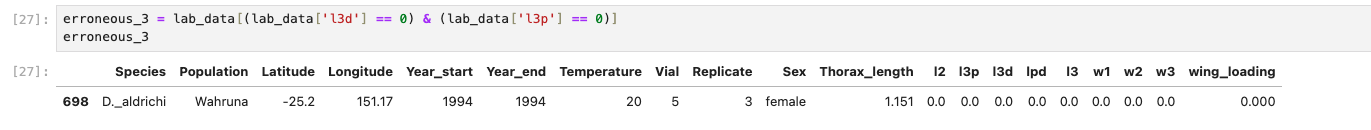
Due to the small nature of the data set and the fact that all other measurement values are available for this row, it makes sense to infer a sensible placeholder for the Thorax\_length and wing\_loading so as not to throw away data.

For this purpose, the median Thorax\_length across the 87 rows containing identical Species, Population, and Sex was used as an estimate of Thorax\_length for this specimen. From there, the wing\_loading estimate was calculated as per the definition specified in Loeschcke et. al [1].

This strategy gives us an estimated Thorax\_length of 1.141, and an estimated wing\_loading of 1.741 for the partially-complete row.

There are an additional two rows that were majority empty/invalid containing zero for seven and nine numeric columns respectively:





Similarly to above, the zero-value columns were estimated to be the median of the associated column across all rows of the same (Species, Population, Sex).

## Problem Description

The “83\_Loeschcke\_et\_al\_2000\_Thorax\_&\_wing\_traits\_lab pops” dataset contains three main classification columns:

* Species (D.\_aldrichi or D.\_buzzatii)
* Population (Binjour, Gogango\_Creek, Grandchester, Oxford\_Downs, Wahruna)
* Sex (Male or Female)

With a total of 1731 data points measuring various parts of the fruit fly anatomy. These anatomical

In order to simulate a high-cardinality classification these classification columns were combined into an appended “class” column with a total of 20 possible class vales.

## Decision Tree Classification

## Softmax Classification

## Neural Network Classification

## Lower Cardinality Classification

Single feature classification:

* Species (cardinality of 2)
* Sex (cardinality of 2)
* Population (cardinality of 5)

Double-feature classification:

* Species, sex (cardinality of 4)
* Species, population (cardinality of 10)
* Population, sex (cardinality of 10)

Decision tree classification

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Class Definition | Cardinality | Test Set Accuracy (min-leaf) | Class Medians | Class Standard Deviations | Class Population |
| Species | 2 | 0.657 | D.\_aldrichi:  D.\_buzzatii: | D.\_aldrichi:  D.\_buzzatii: |  |
| Sex | 2 | 0.811 |  |  |  |
| Species, Sex | 4 | 0.509 |  |  |  |
| Population | 5 | 0.233 |  |  |  |
| Species, Population | 10 | 0.164 |  |  |  |
| Population, Sex | 10 | 0.204 |  |  |  |
| Species, Population, Sex | 20 | 0.133 |  |  |  |

Softmax classification

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class Definition | Cardinality | Test Set Accuracy (Standardised) | Class Medians | Class Standard Deviations |
| Species | 2 | 0.719 |  |  |
| Sex | 2 | 0.830 |  |  |
| Species, Sex | 4 | 0.566 |  |  |
| Population | 5 | 0.260 |  |  |
| Species, Population | 10 | 0.204 |  |  |
| Population, Sex | 10 | 0.225 |  |  |
| Species, Population, Sex | 20 | 0.15 |  |  |

Neural Network classification

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class Definition | Cardinality | Test Set Accuracy after 1200 epochs | Class Medians | Class Standard Deviations |
| Species | 2 | 128: 0.686  256: 0.680  512: 0.678  1024: 0.8660 |  |  |
| Sex | 2 | 128: 0.825  256: 0.830  512: 0.819  1024: 0.834 |  |  |
| Species, Sex | 4 | 128: 0.557  256: 0.553  512: 0.549  1024: 0.530 |  |  |
| Population | 5 | 128: 0.256  256: 0.237  512: 0.191  1024: 0.177 |  |  |
| Species, Population | 10 | 128: 0.164  256: 0.187  512: 0.190  1024: 0.164 |  |  |
| Population, Sex | 10 | 128: 0.229  256: 0.225  512: 0.214  1024: 0.224 |  |  |
| Species, Population, Sex | 20 | 128: 0.150  256: 0.135  512: 0.123  1024: 0.135 |  |  |

## References