Darwin's Finches

Anushka Pradhan, Paige Ellingson, Yumi Chen, Kate Chen, Thomas Dinunzio

Introduction

- Dataset:

  Darwin's Finches Evolution Dataset[3]

- Parameters:
  - Years: 1975, 2012
  - Beak Length
  - Beak Depth
  - Species: Fortis, Scanden
  - Heredity: “In scientific terms, heritability is a statistical concept (represented as h^2) that describes how much of the variation in a given trait can be attributed to genetic variation.”[4]; similar to R^2
- We are interested in classifying species and year from beak length and beak depth data.
- Welch’s t-test for mean beak depth between offspring and parents found:
  - ○ Fortis p-value = 0.00019
  - ○ Scandens p-value = 0.0028

Fortis[1]

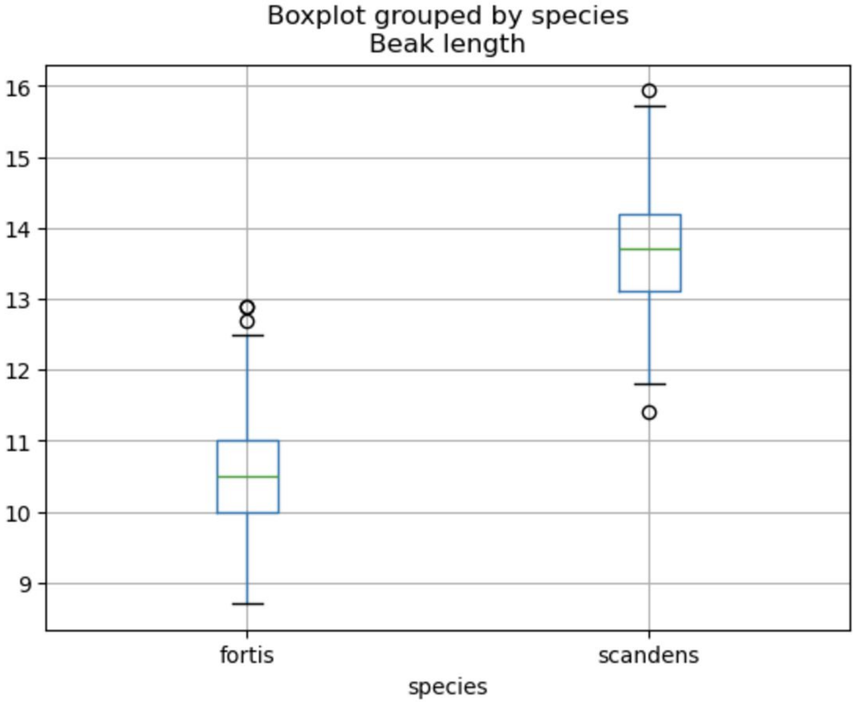
Scandens[2]

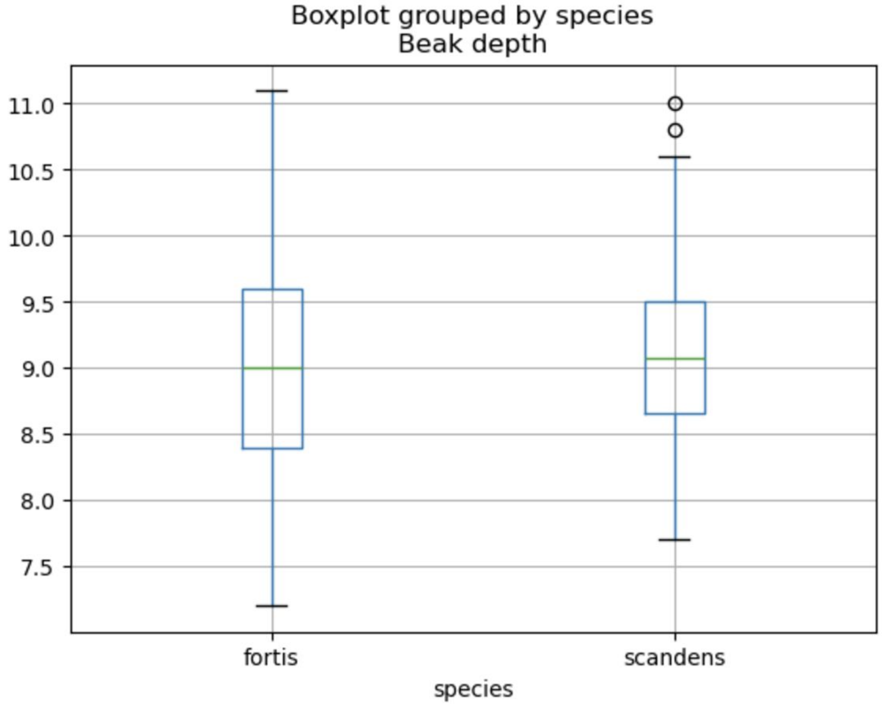
Methods

- Oversampling to overcome imbalanced data set in terms of species(more of one species than the other)
  - 413 fortis, 130 scandens
- Combine and format data from 2 years (1975, 2012) and 2 species
- t-test on sample means
- Decision tree classification - species
- Logistic regression to predict log-odds of year, log-odds of species
- kNN classification to predict log-odds of year, log-odds of species
- Unsupervised learning using PCA to separate the two species
  - with & without feature standardization
  - Incorporating data from both 1975 and 2012 as an additional dimension

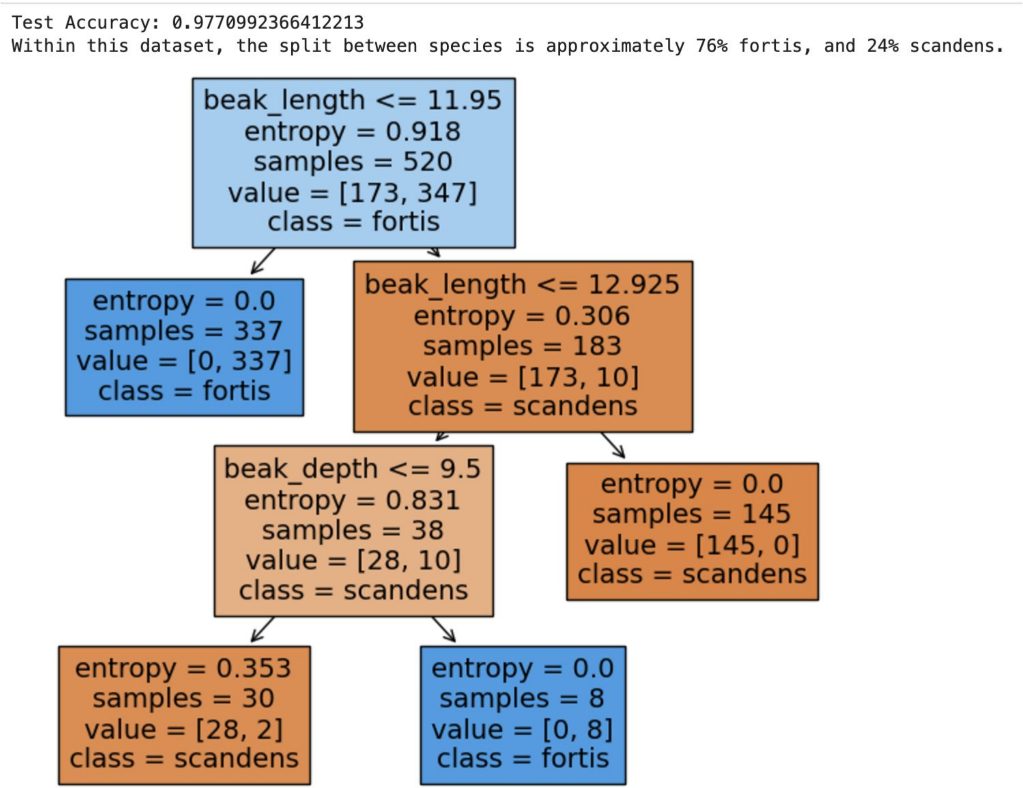
Data Distribution

Comparing the the distributions of the “Beak Length” and “Beak Depth” of the 2 species, it is apparent that “Beak Length” is more different among the 2 species.





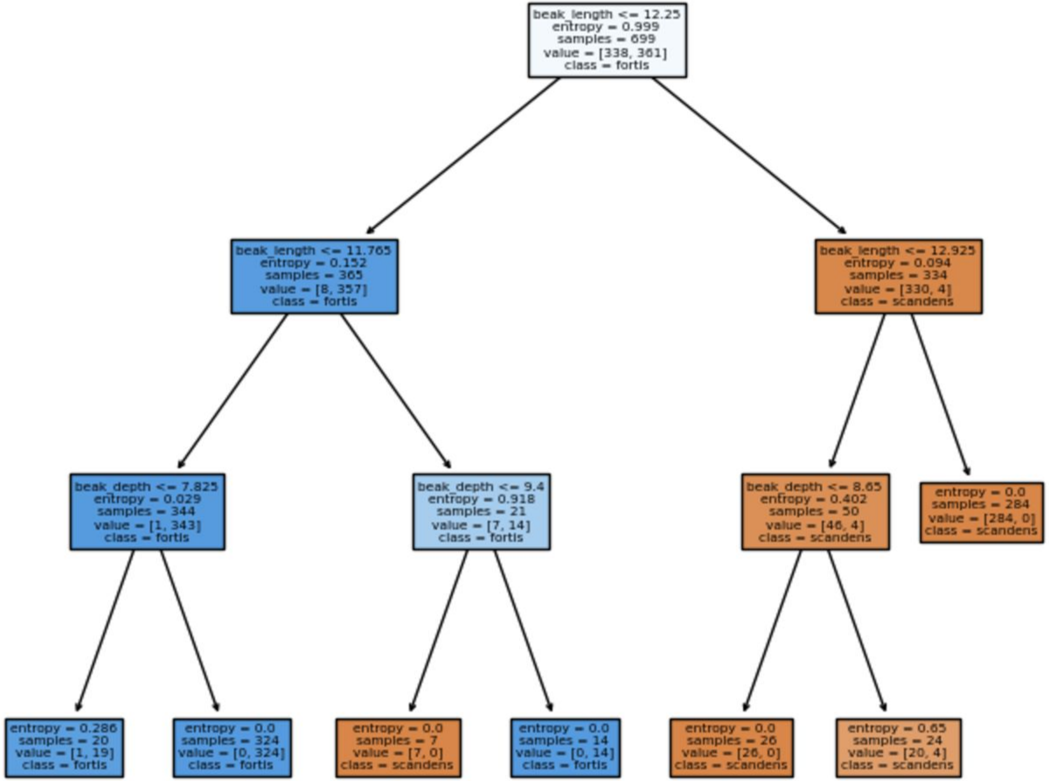
Decision Tree



```
Test Accuracy: 0.9770992366412213
Within this dataset, the split between species is approximately 76% fortis, and 24% scandens.
```

- Train Test Split
- max_depth = 3
- y = species
- X = beak_length, beak_depth
- Test Accuracy: .977
- Model might be overfitting
  - 76% fortis
  - 24% scandens

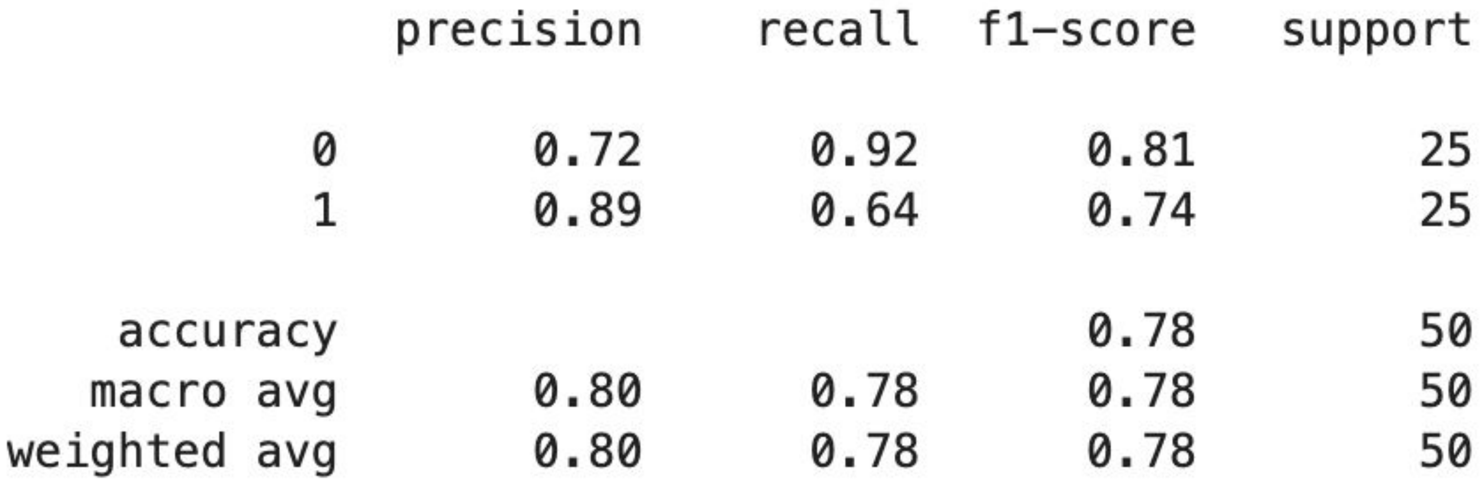
Decision Tree with Oversampling



- Train Test Split
- RandomOverSampler
- max_depth = 3
- y = species
- X = beak_length, beak_depth
- Test Accuracy: .994

Logistic Regression - Classification of fortis and scandans (Without Over-sampling)

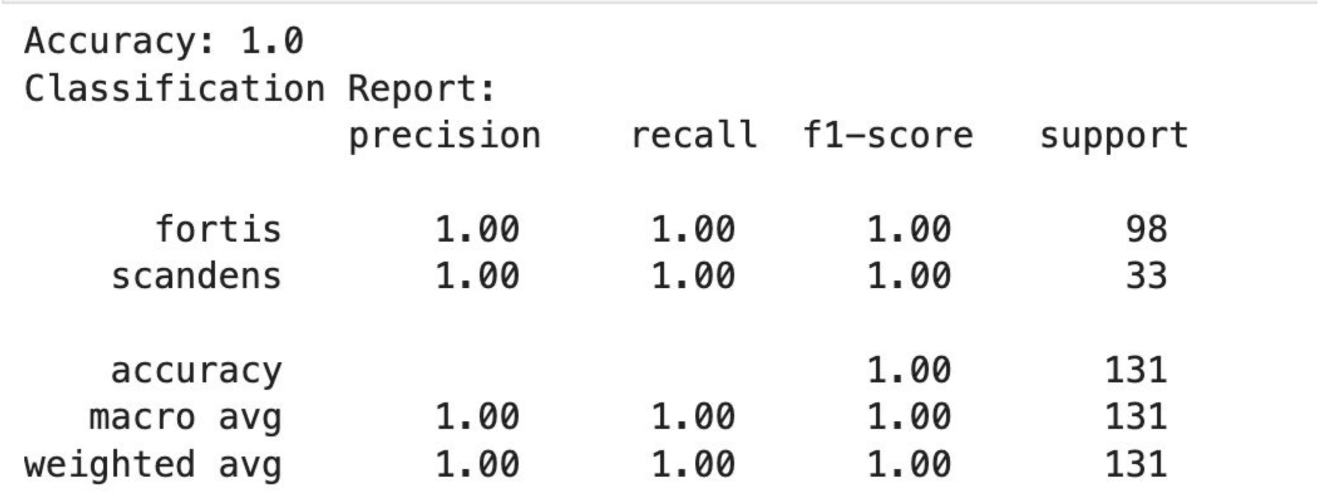
- Imbalance data: 437 fortis v.s. 214 scandans
- Train: 80%, 10% validation, 10% test
- X: beak length, beak depth, y: species
- coefficient = [-0.2015458   1.73670421], intercept = [0.0650701], training score=0.828



```
              precision    recall  f1-score   support

           0       0.72      0.92      0.81        25
           1       0.89      0.64      0.74        25

    accuracy                           0.78        50
   macro avg       0.80      0.78      0.78        50
weighted avg       0.80      0.78      0.78        50
```

Logistic Regression - Classification of fortis and scandans (With Over-sampling)

- Imbalance data: 437 fortis v.s. 214 scandans
- Oversampling to address the problem: 339 fortis v.s. 339 scandens
- Train: 80%, Test: 20%
- X: beak length, beak depth
- y: species



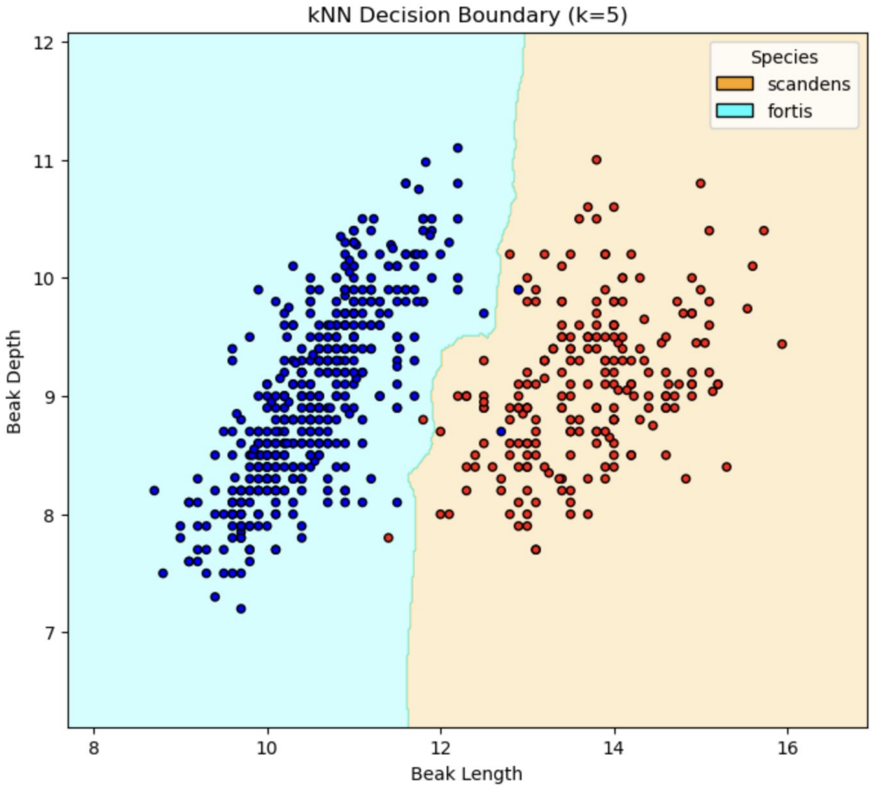
```
Accuracy: 1.0
Classification Report:
              precision    recall  f1-score   support

      fortis       1.00      1.00      1.00        98
    scandens       1.00      1.00      1.00        33

    accuracy                           1.00       131
   macro avg       1.00      1.00      1.00       131
weighted avg       1.00      1.00      1.00       131
```

k-Nearest Neighbors - Classification of fortis and scandens

Without Oversampling

- X: beak length, beak depth; y: species
- Train: 80%, Test: 20%
- k = 5
- 437 Fortis v.s. 214 Scandens
- Test Accuracy: 0.9771

```
Within this dataset, the split between species is approximately 76% fortis, and 24% scandens.
```

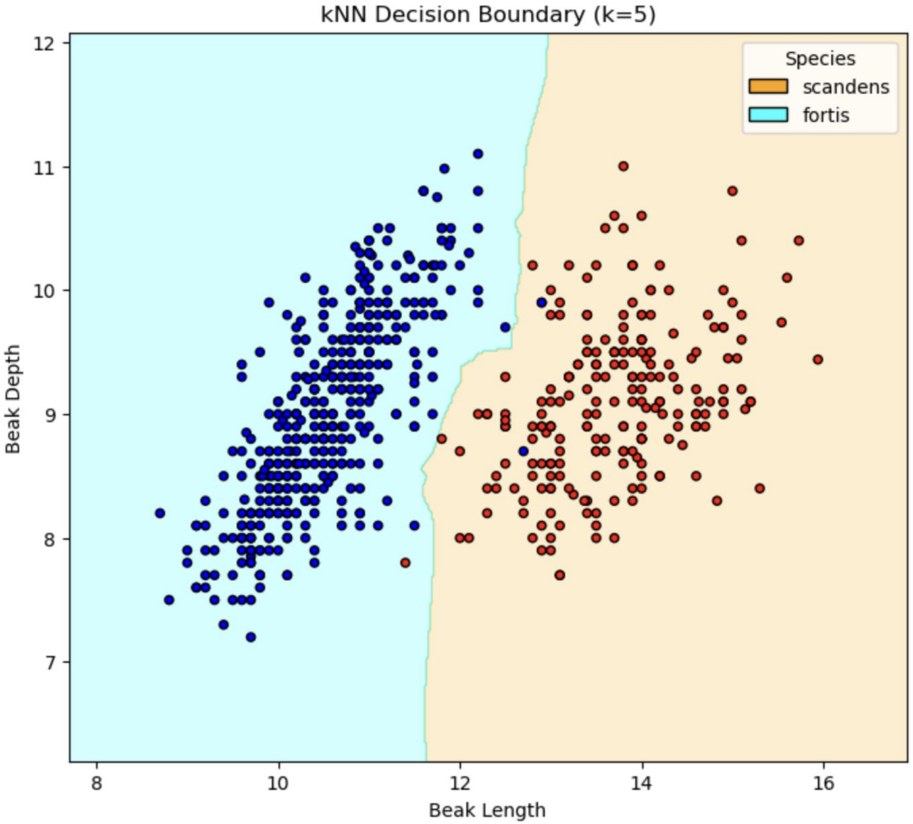


kNN Decision Boundary (k=5)

k-Nearest Neighbors - Classification of fortis and scandens

With Oversampling

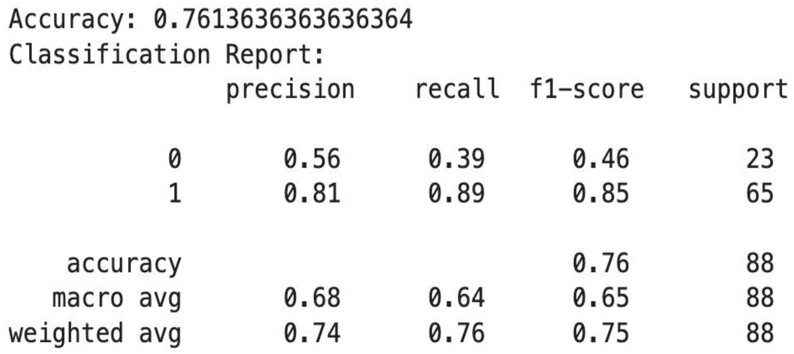
- X: beak length, beak depth; y: species
- Train: 80%, Test: 20%
- k = 5
- RandomOverSampler
- 347 Fortis v.s. 347 Scandens
- Test Accuracy: 0.9847



Logistic Regression - Classification of year 1975 and 2012 for fortis

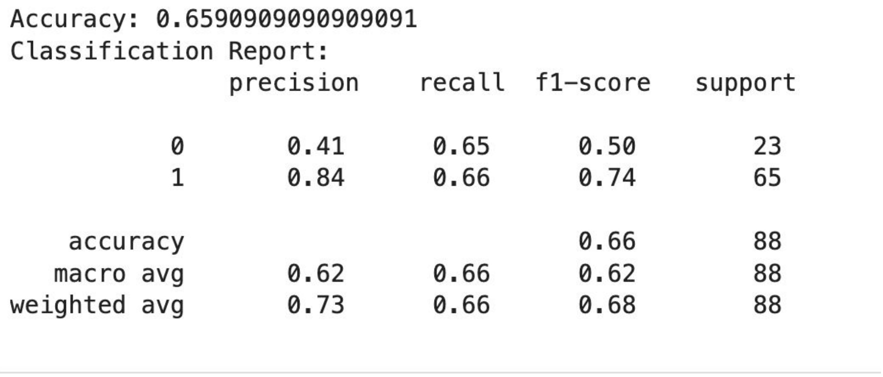
- Imbalance data for fortis - Fortis: 316 1975 v.s. 121 2012
- Oversampling to address the problem: Fortis: 251 1975 v.s. 251 2012
- Train: 80%, Test: 20%
- X: beak length, beak depth; y: year

Classify which year the fortis from without oversampling



```
Accuracy: 0.7613636363636364
Classification Report:
               precision    recall  f1-score   support

           0       0.56      0.39      0.46        23
           1       0.81      0.89      0.85        65

    accuracy                           0.76        88
   macro avg       0.68      0.64      0.65        88
weighted avg       0.74      0.76      0.75        88
```

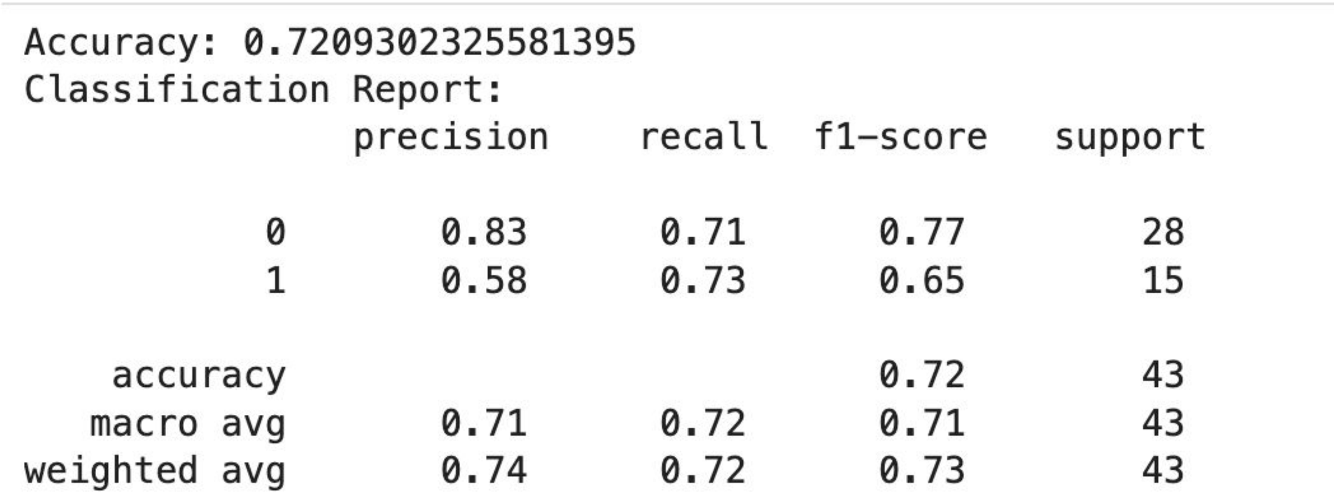
Classify which year the fortis from with oversampling (better recall for 2012 at the cost of accuracy)



```
Accuracy: 0.6590909090909091
Classification Report:
               precision    recall  f1-score   support

           0       0.41      0.65      0.50        23
           1       0.84      0.66      0.74        65

    accuracy                           0.66        88
   macro avg       0.62      0.66      0.62        88
weighted avg       0.73      0.66      0.68        88
```

Logistic Regression - Classification of year 1975 and 2012 for scandens

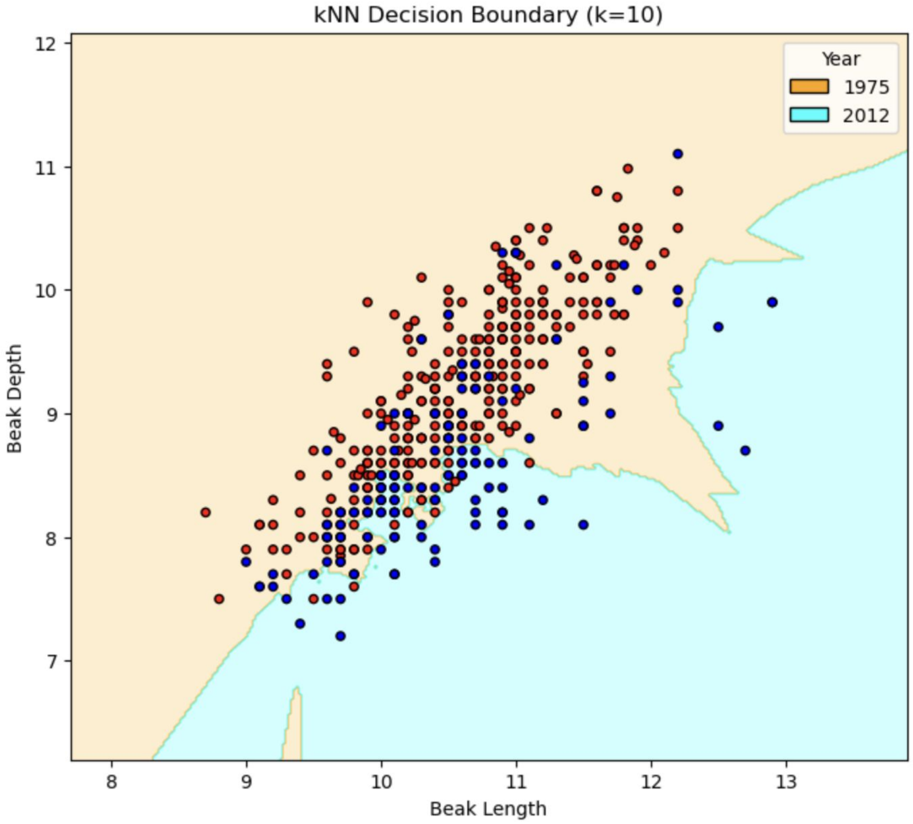
- Relatively balanced data: 87 1975 vs 127 2012
- Train: 80%, Test: 20%
- X: beak length, beak depth; y: year



```
Accuracy: 0.7209302325581395
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.71      0.77        28
           1       0.58      0.73      0.65        15

    accuracy                           0.72        43
   macro avg       0.71      0.72      0.71        43
weighted avg       0.74      0.72      0.73        43
```

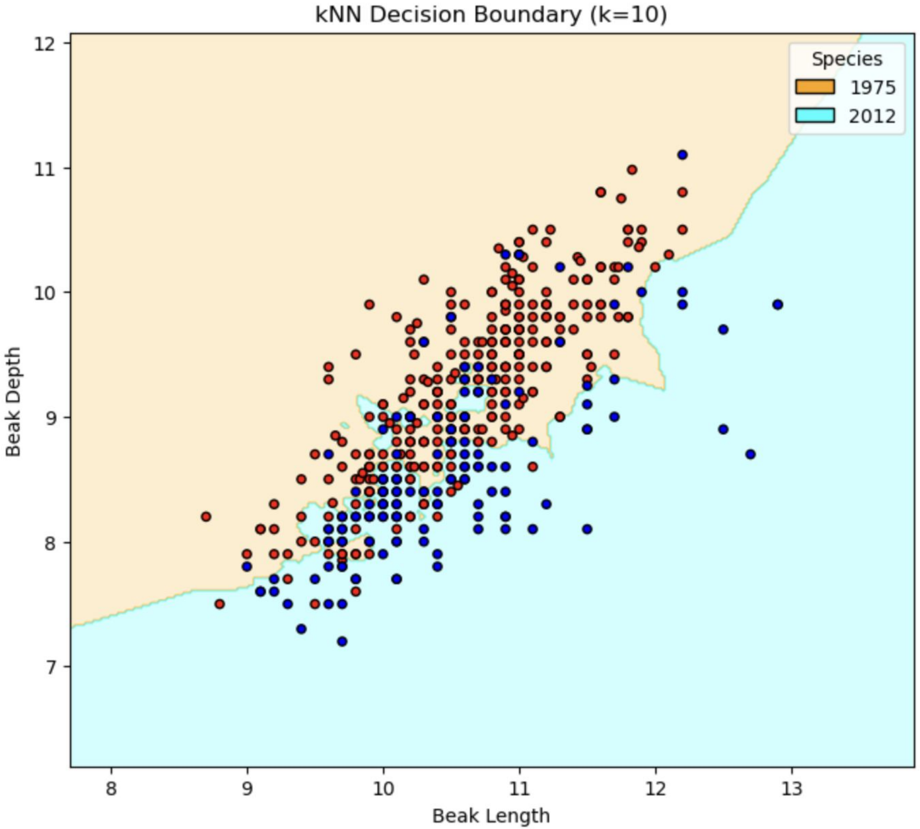
k-Nearest Neighbors - Classification of year 1975 and 2012 for fortis (Without Oversampling)

- X: beak length, beak depth; y: year
- Train: 80%, Test: 20%
- k = 10
- Fortis: 316 1975 v.s. 121 2012
- Test Accuracy: 0.8182
- 1975 Recall: 0.96
- 2012 Recall: 0.28



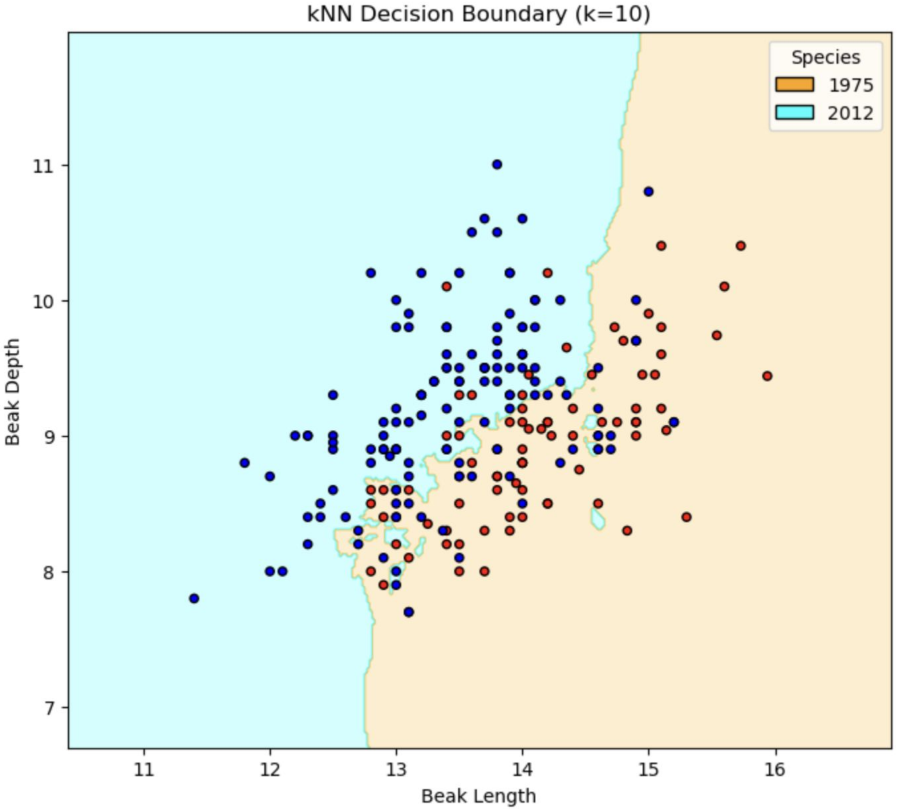
k-Nearest Neighbors - Classification of year 1975 and 2012 for fortis (With Oversampling)

- X: beak length, beak depth; y: year
- Train: 80%, Test: 20%
- k = 10
- RandomOverSampler
- Fortis: 246 1975 v.s. 246 2012
- Test Accuracy: 0.6704
- 1975 Recall: 0.70
- 2012 Recall: 0.56



k-Nearest Neighbors - Classification of year 1975 and 2012 for scandens

- X: beak length, beak depth; y: year
- Train: 80%, Test: 20%
- k = 10
- Scandens: 127 1975 v.s. 87 2012
- Test Accuracy: 0.7907
- 1975 Recall: 0.86
- 2012 Recall: 0.76



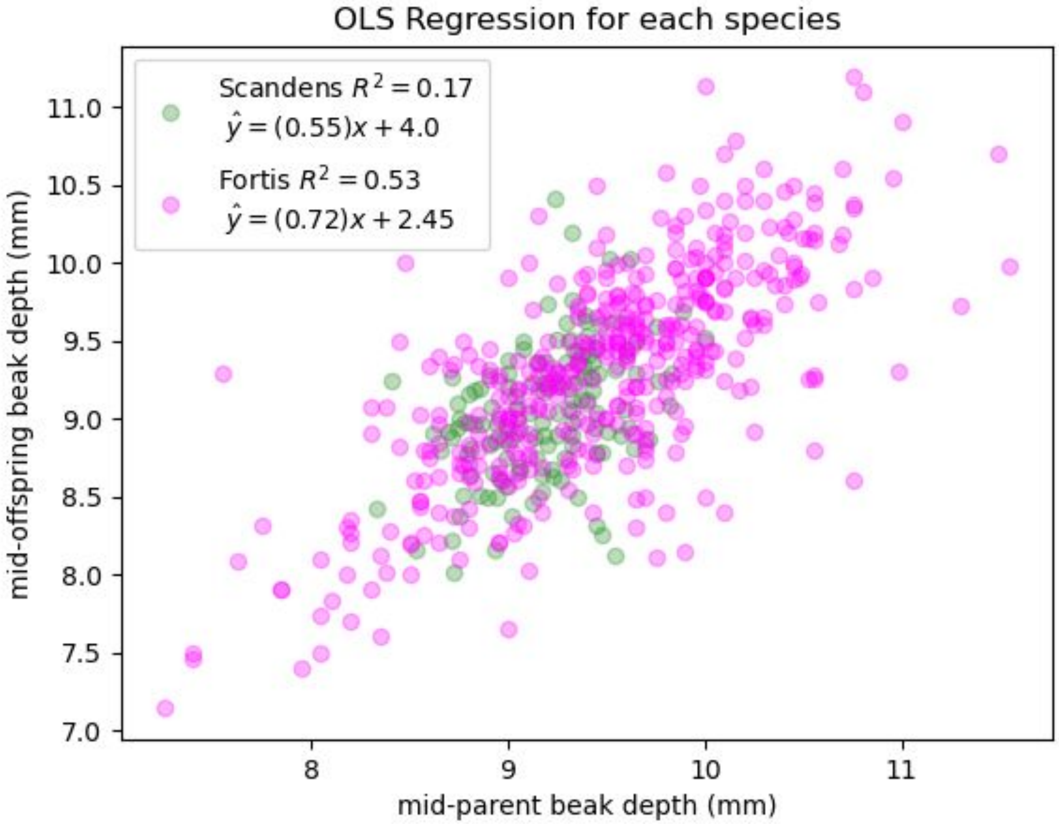
Linear Regression - Heredity index score for fortis

Scandens: R^2 = 0.17

- 17% of the variance in the offspring's beak depth can be explained by the variance in the mid-parent's beak depth.

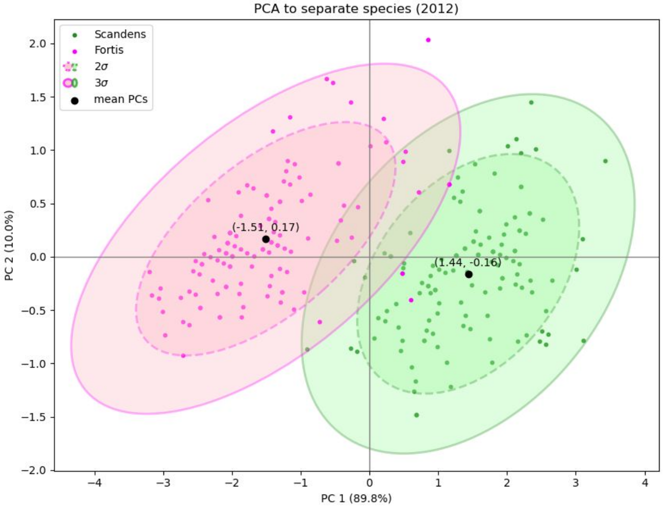
Fortis: R^2 = 0.53

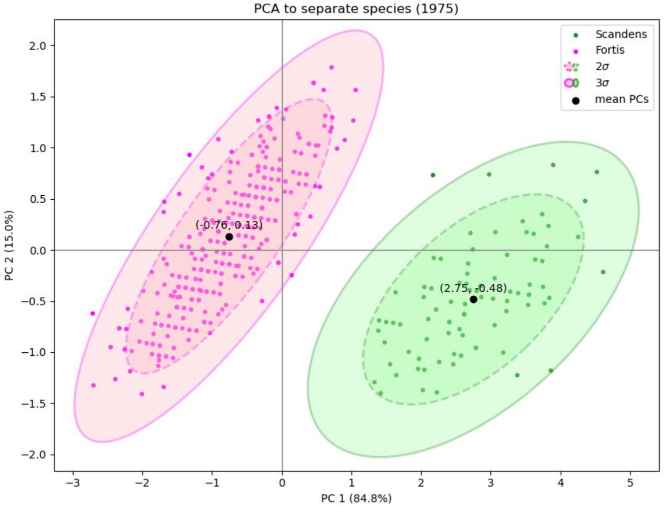
- 53% of the variance in the offspring's beak depth can be explained by the variance in the mid-parent's beak depth.



Unsupervised learning: Principal Component Analysis

Performed dimensionality reduction (orthogonalization) on beak depth and beak length data, colored by species to visualize separation



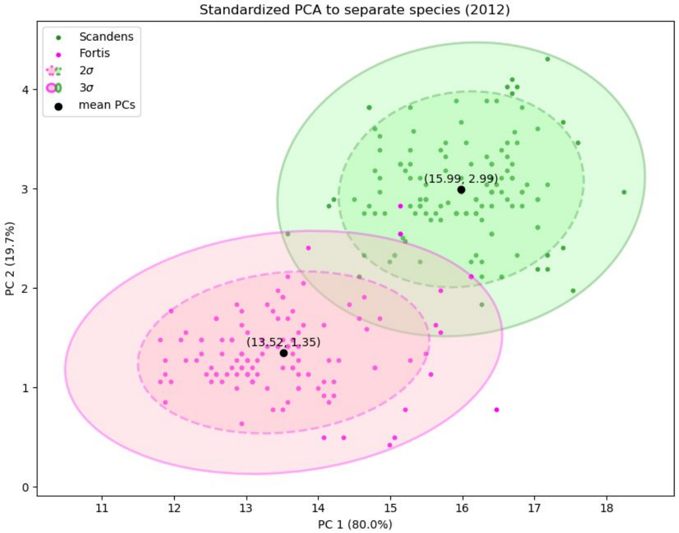


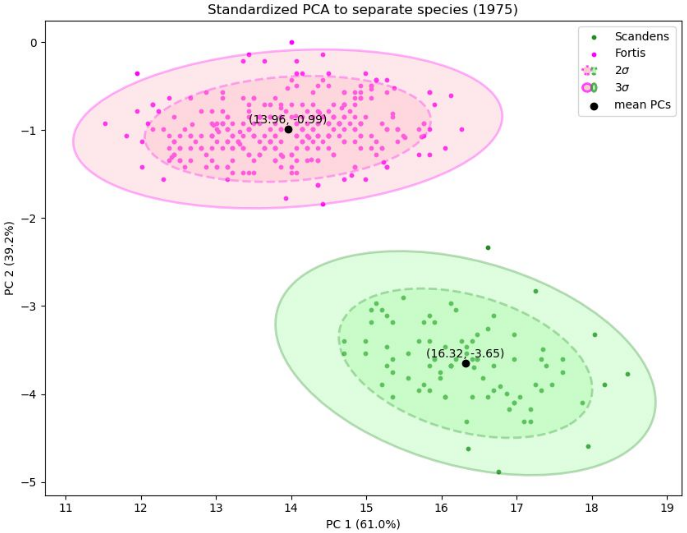
PC 1: 89.8%
PC 2: 10.0%

PC 1: 84.8%
PC 2: 15.0%

PCA is sensitive to standardization

Standardizing each feature randomly chooses an axis to be the first component, because all have σ=1.





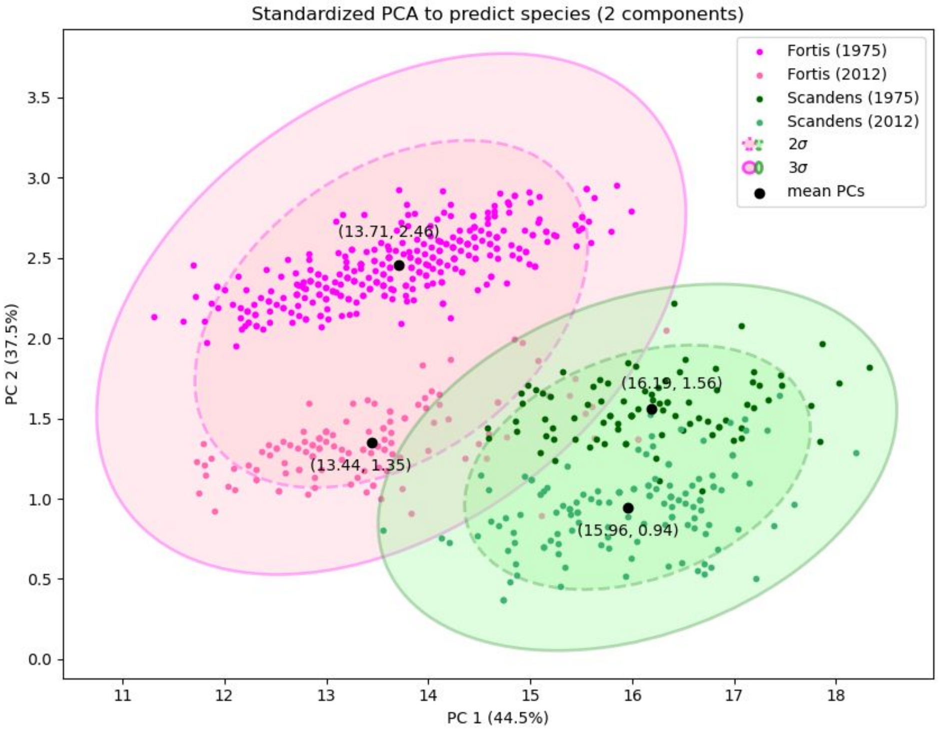
PC 1: 80.0%
PC 2: 19.7%

PC 1: 61.0%
PC 2: 39.2%

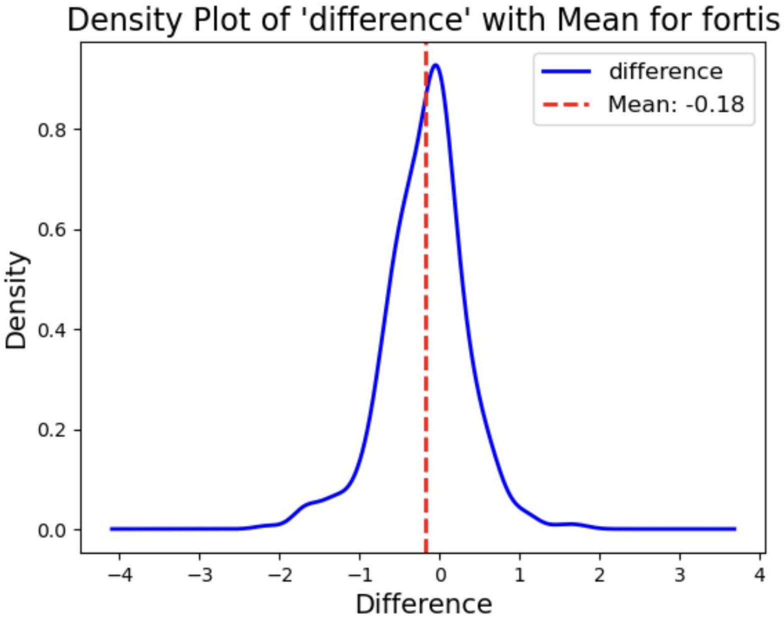
Standardized PCA: data from two years

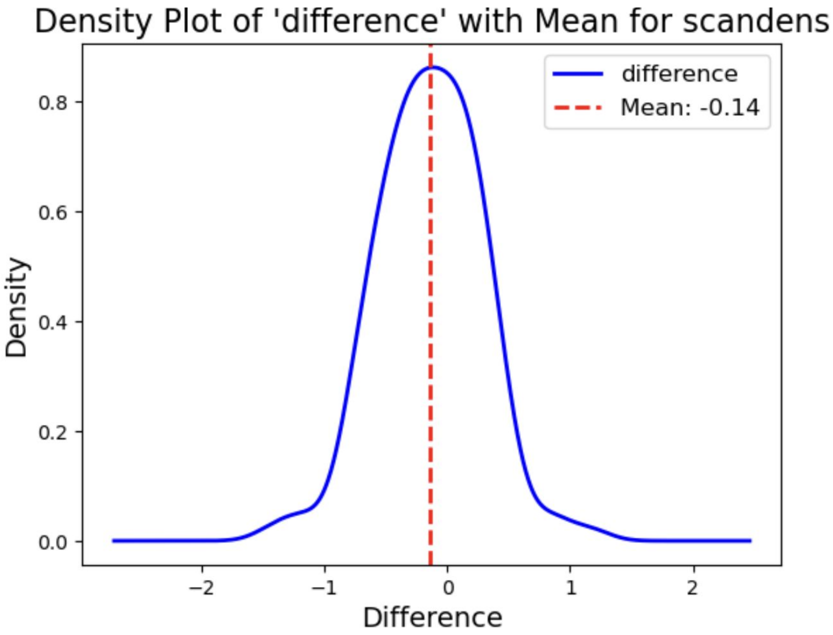
PC	% exp. var.	Equation
1	44.5%	$0.735131\ x_1 + 0.663541\ x_2 - 0.138911\ x_3$
2	37.5%	$-0.226035\ x_1 + 0.433087\ x_2 + 0.87255\ x_3$
3	18.0%	$0.639133\ x_1 - 0.610041\ x_2 + 0.468359\ x_3$

PC1 & PC2 account for 82.0% of the variance in the data, so we lost 18% through data compression.



Density plot: Difference of bead depth between mid parent and mid offsprings





Outcome

- The proportion of Fortis is 78% in 1975 dataset and 49% in 2012 dataset, and 67% in the combined dataset of both year, so the proportion is not significantly close to 1 or 0.
- We found that the beak length is a better predictor for species than the beak depth
- We can predict species with 97.7% or 100% accuracy on unseen data using Decision Tree or Logistic Regression model
- We can predict log-odds of year (1975 or 2012) given species with accuracy 65.91% for fortis and 72.09% for scandens on unseen data using logistic regression.
- PCA component #1 had ~88% variance explained, component #2 had ~10%
- Heredity indices gave r^2 = 53% for Fortis, r^2 = 17% for Scandens.
- There is significant difference between mean beak depth between mid parent and mid offspring for both fortis and scandens.
- Both low heredity indices and significant difference suggests that offspring's beak depth is influenced by the parents' beak depth for both fortis and scandens, but other factors (such as environmental influences, measurement error, or genetic interactions) play a significant role in determining the trait.