PRÀCTICA 3

CAS KAGGLE

Serena Sánchez

APRENENTATGE COMPUTACIONAL

BASE DE DADES

Fish Market

7
Atributs

159
Mostres

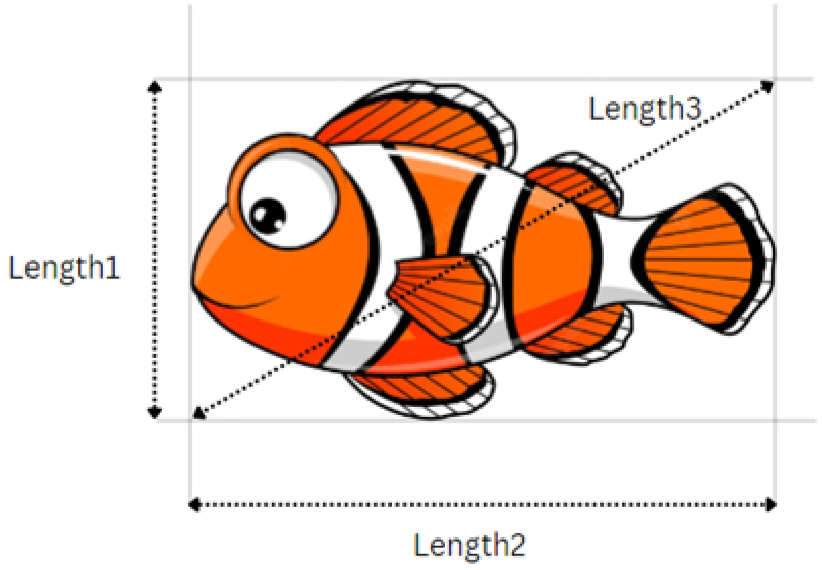
SPECIES
Atribut Objectiu

	Species	Weight	Length1	Length2	Length3	Height	Width
0	Bream	242.000	23.200	25.400	30.000	11.520	4.020
1	Bream	290.000	24.000	26.300	31.200	12.480	4.306
2	Bream	340.000	23.900	26.500	31.100	12.378	4.696

BREAM ROACH PIKE PEARCH WHITEFISH SMELT PARKKI

BASE DE DADES

ies	Weight	Length1	Length2	Length3	Height	Wi
eam	242.000	23.200	25.400	30.000	11.520	4.
eam	290.000	24.000	26.300	31.200	12.480	4.
eam	340.000	23.900	26.500	31.100	12.378	4.



PITÀGORES:

$$Length3^2 = Length2^2 + Length1^2$$

ies	Weight	len_diagonal	Height	Widt
am	242.000	30.000	11.520	4.02
am	290.000	31.200	12.480	4.3(
am	340.000	31.100	12.378	4.6(

PREPROCESSAMENT DE DADES

	Species	Weight	len_diagonal	Height	Width
40	Roach	0.000	22.800	6.475	3.352

VALOR sense INFORMACIÓ

INFERÈNCIA (KNNIMPUTER)

	Species	Weight	len_diagonal	Height	Width
40	Roach	119.636	22.800	6.475	3.352

label encoding

SPECIES	SPECIES
BREAM	1
ROACH	2
WHITEFISH	3

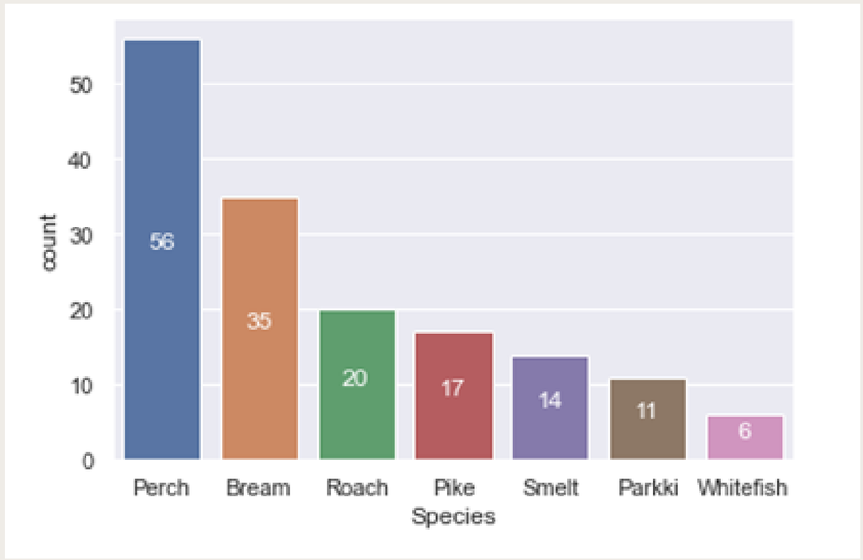
one hot encoding

SPECIES	BREAM	ROACH	WHITE FISH
BREAM	1	0	0
ROACH	0	1	0
WHITEFISH	0	0	1

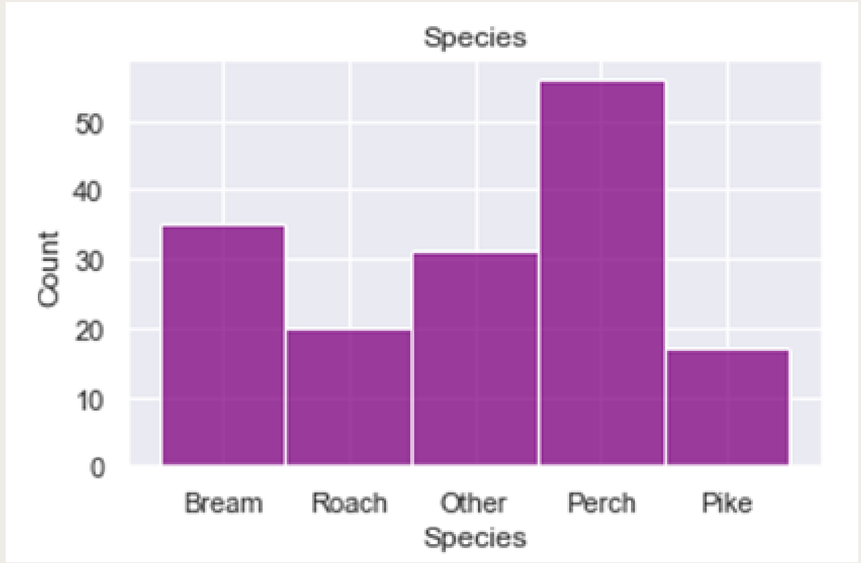
DADES CATEGÒRIQUES

PREPROCESSAMENT DE DADES

DATASET NO AGRUPAT

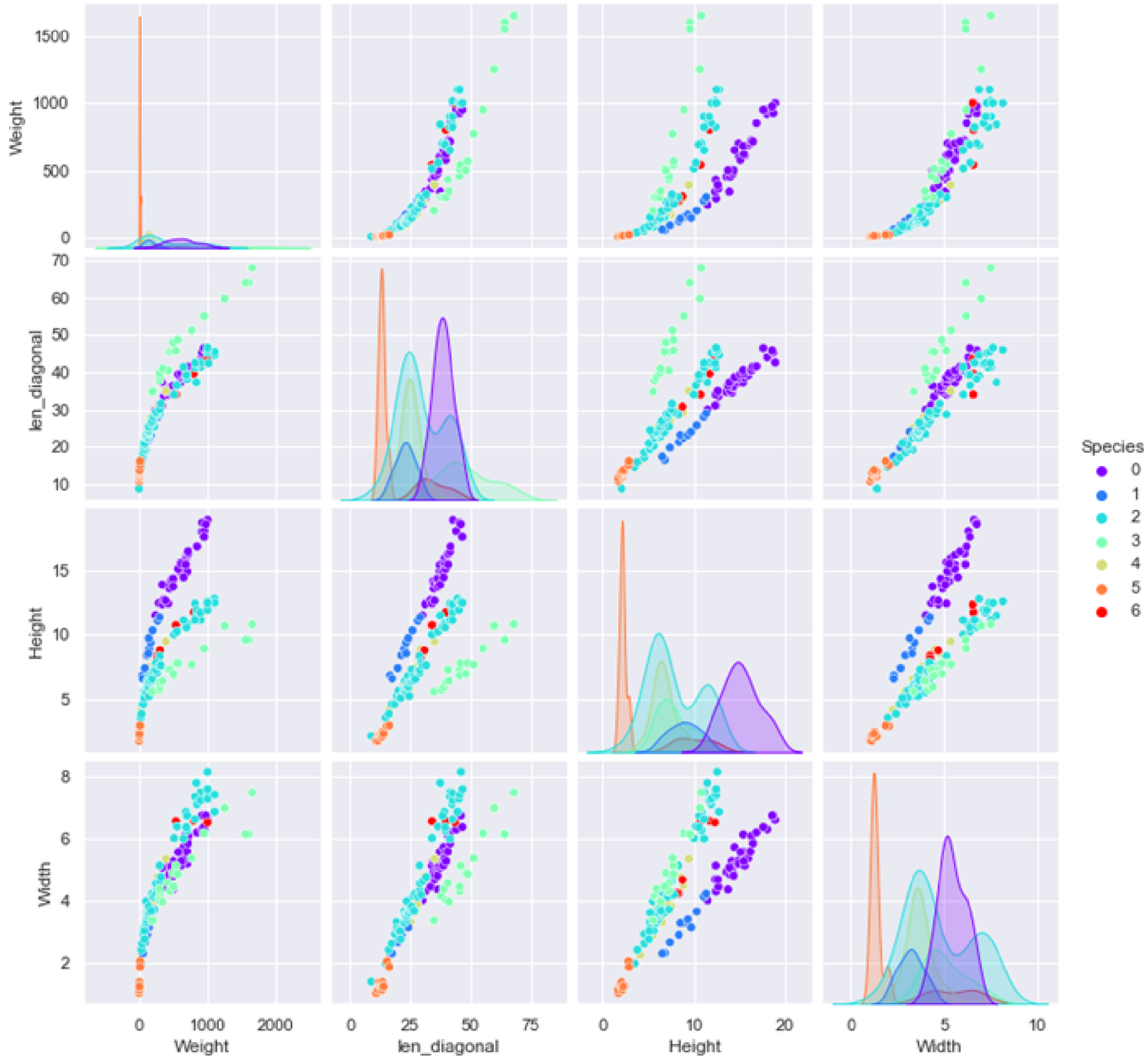


DATASET AGRUPAT



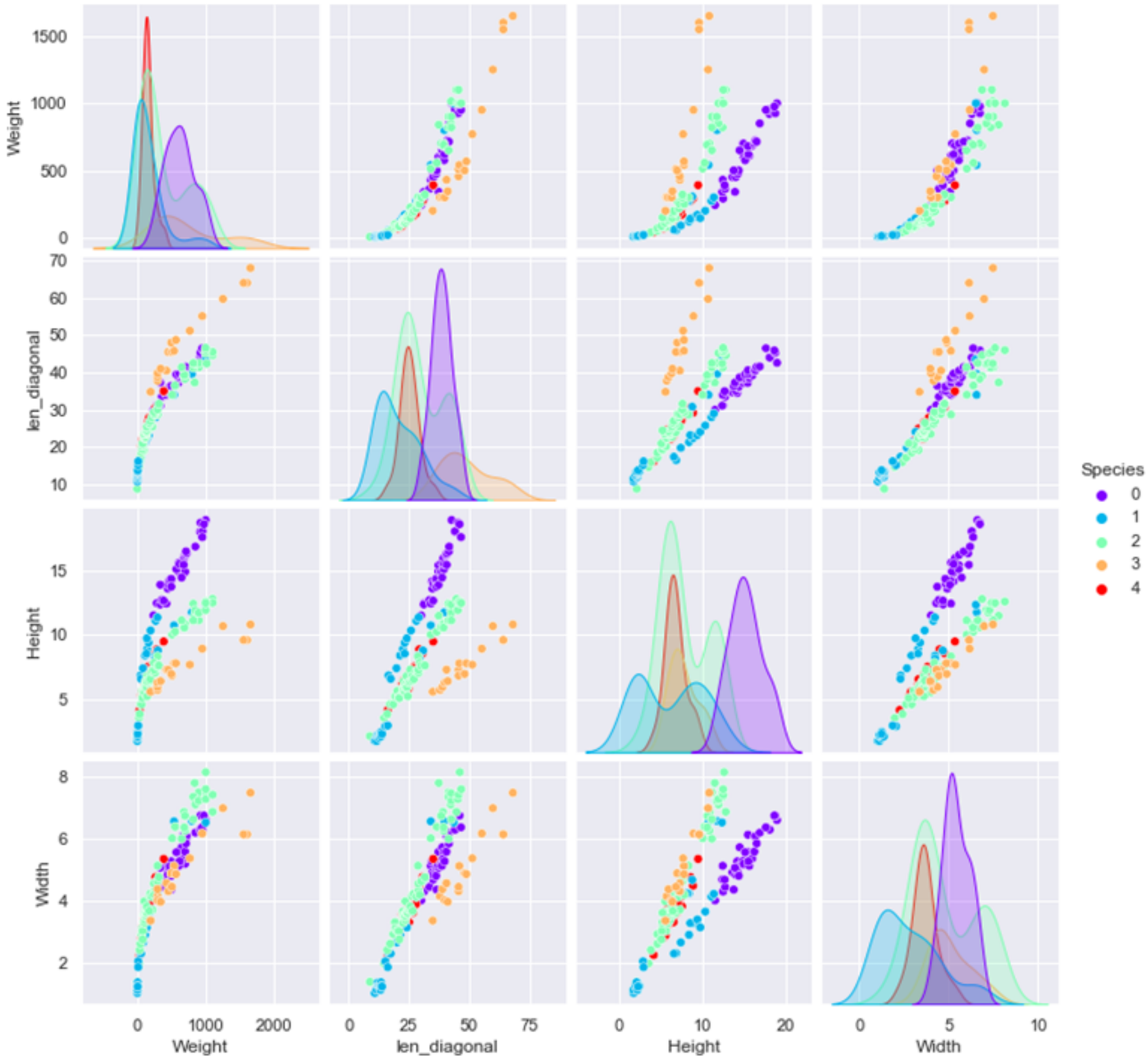
BALANCEJAMENT DE DADES

PREPROCESSAMENT DE DADES



DATASET NO AGRUPAT

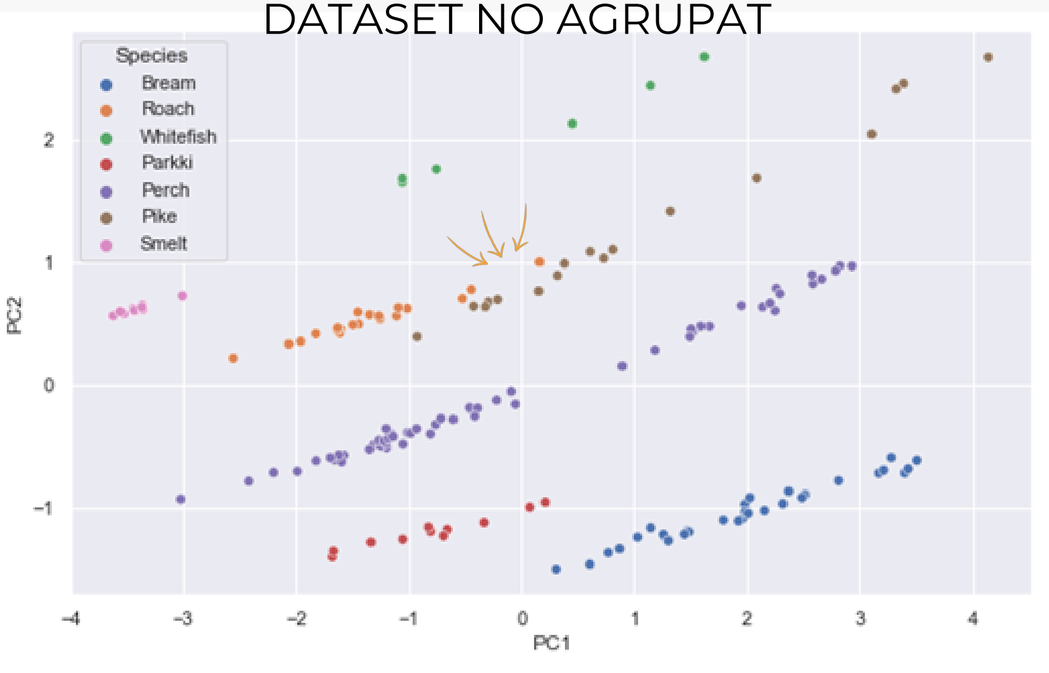
PREPROCESSAMENT DE DADES



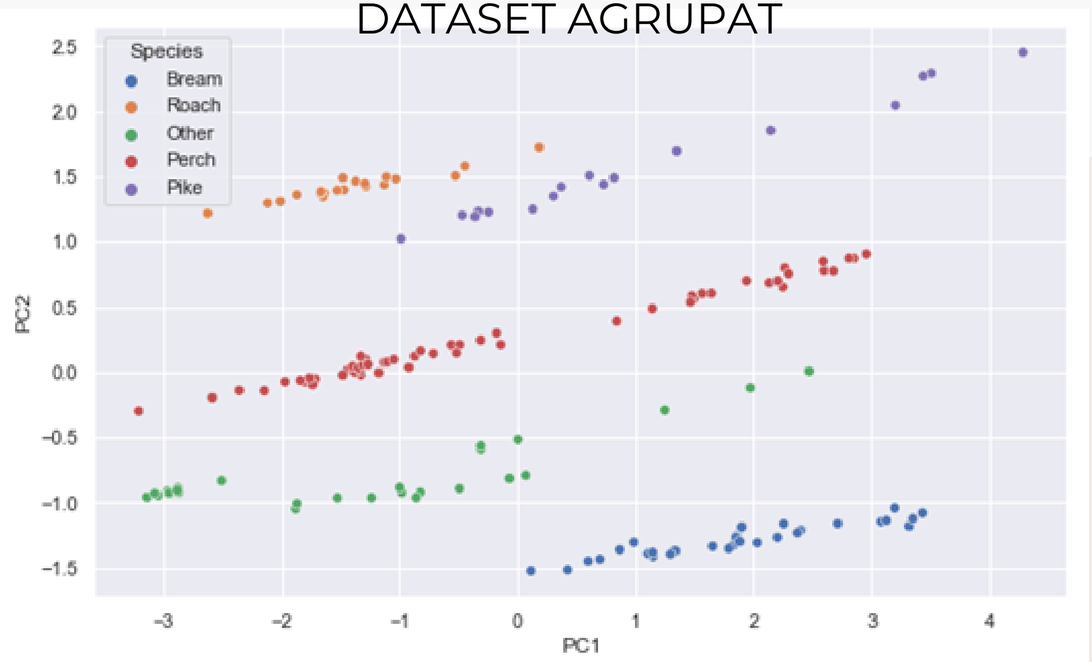
DATASET AGRUPAT

PREPROCESSAMENT DE DADES

DATASET NO AGRUPAT

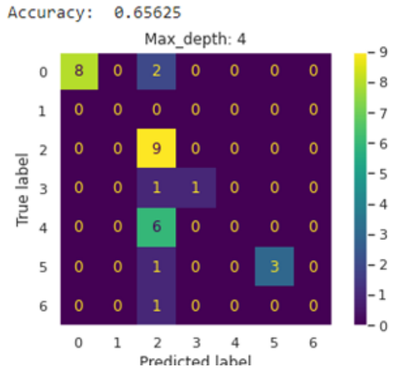


DATASET AGRUPAT

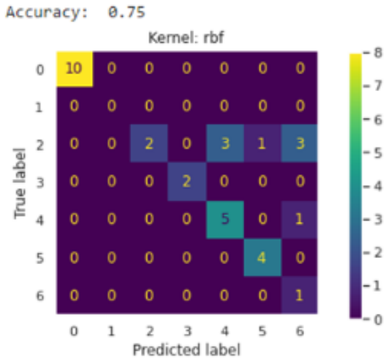


PCA
(normalització prèvia de les dades)

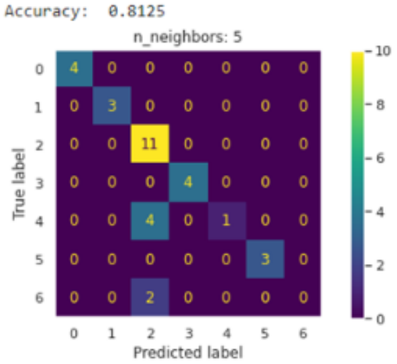
SELECCIÓ DEL MODEL



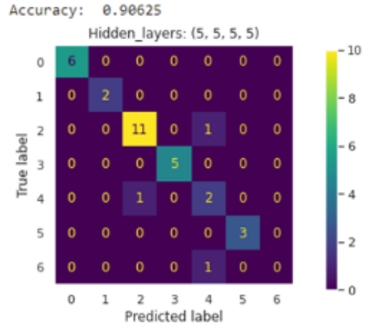
Arbre de decisió



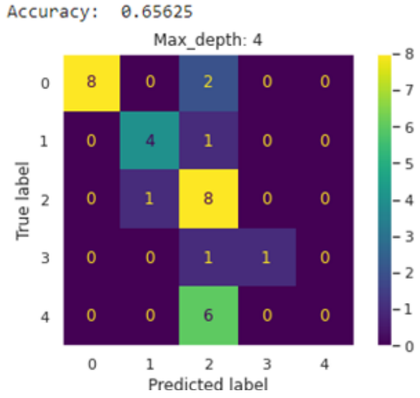
SVM

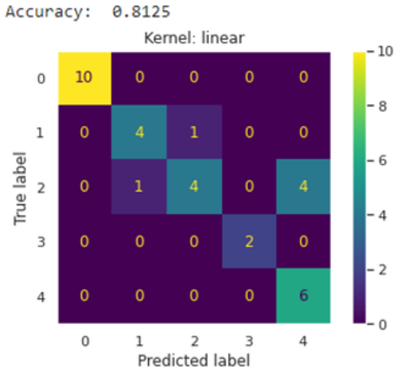


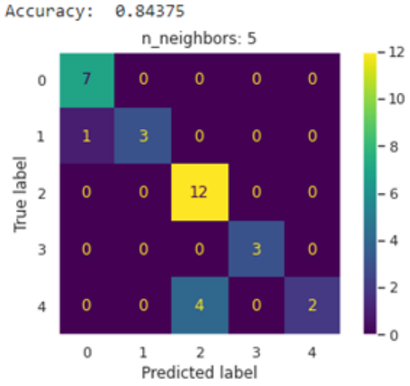
KNN

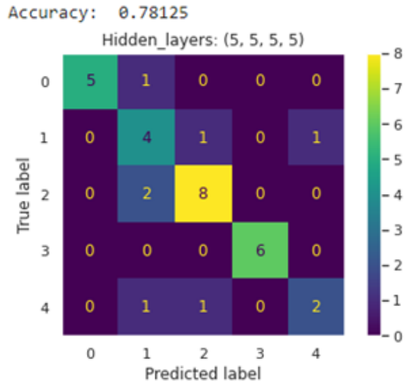


Xarxes Neuronals





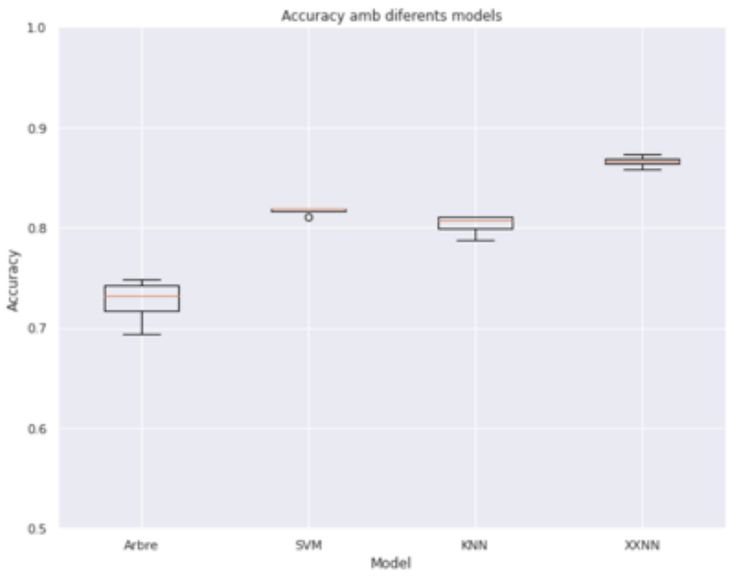




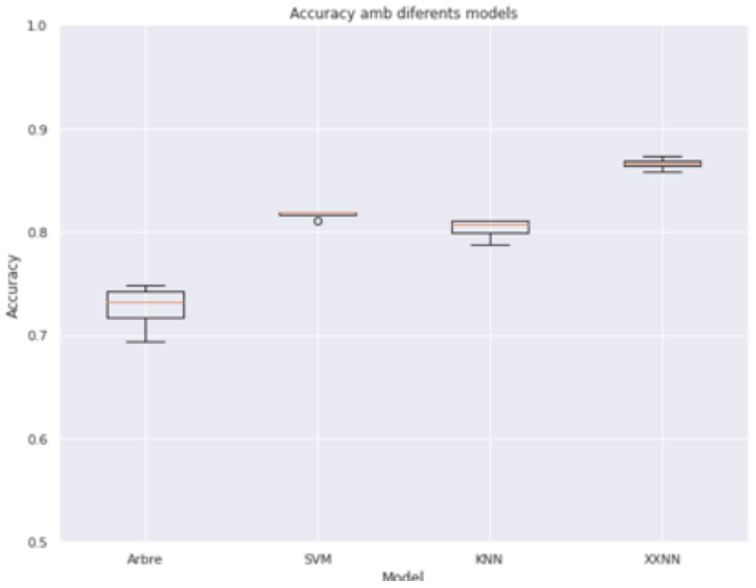
OPTIMITZACIÓ D'HYPERPARÀMETRES

Arbre de decisió	SVM	KNN	Xarxes Neuronals
criterion: entropy splitter: best max_depth: 3 class_weight: None	shape: ovo kernel: linear degree: 1 class_weight: None gamma: scale	n_neighbors: ? weights: uniform (7), distance (5) algorithm: auto	hidden_layers_size: ? activation: indentity, tanh solver: adam max_iter: 1000

Espècies no agrupades:



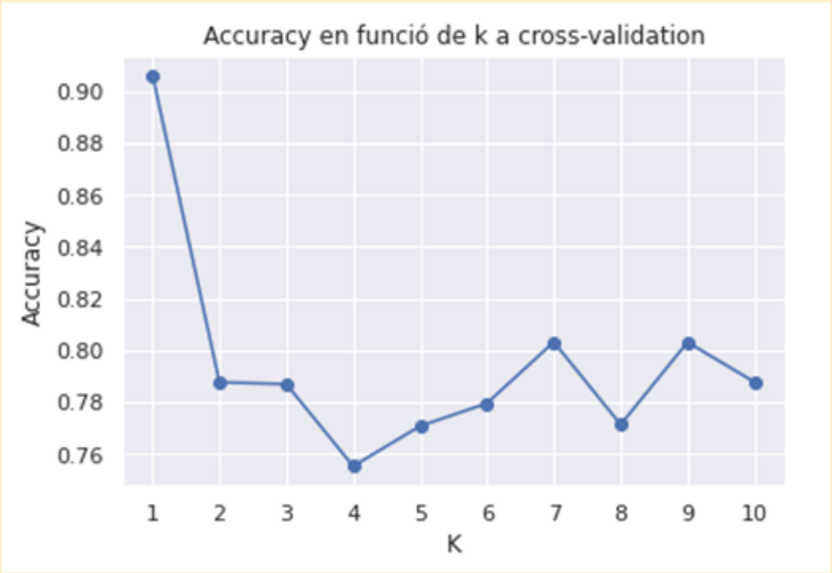
Espècies agrupades:



LAZY PREDICTOR

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
QuadraticDiscriminantAnalysis	0.91	0.94	None	0.91	0.01
KNeighborsClassifier	0.84	0.72	None	0.81	0.02
XGBClassifier	0.81	0.76	None	0.80	0.06
LinearDiscriminantAnalysis	0.78	0.69	None	0.77	0.02
SGDClassifier	0.78	0.68	None	0.74	0.01
CalibratedClassifierCV	0.78	0.67	None	0.70	0.14
Perceptron	0.78	0.67	None	0.70	0.01
RandomForestClassifier	0.78	0.62	None	0.75	0.18
DecisionTreeClassifier	0.78	0.83	None	0.78	0.03
PassiveAggressiveClassifier	0.75	0.65	None	0.67	0.02

QUADRATIC DISCRIMINANT ANALYSIS



COMPARATIVA AMB UN ALTRE TREBALL

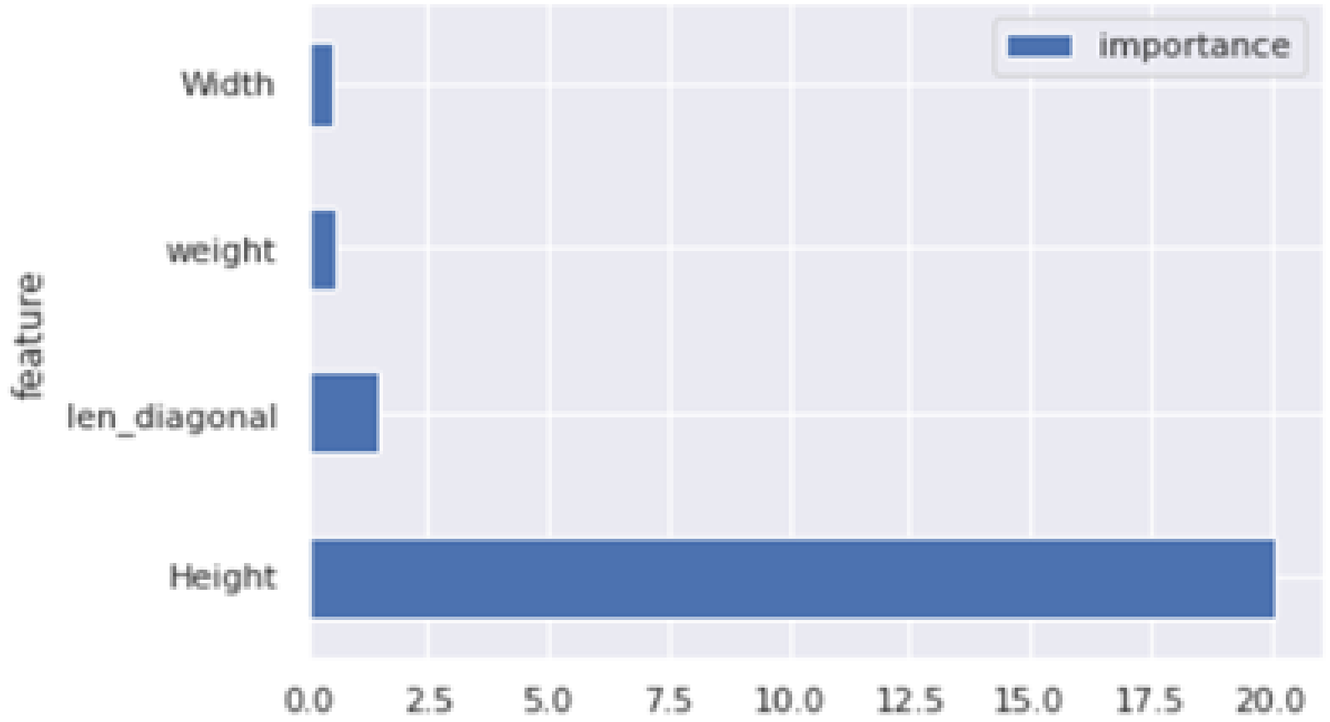
PCA

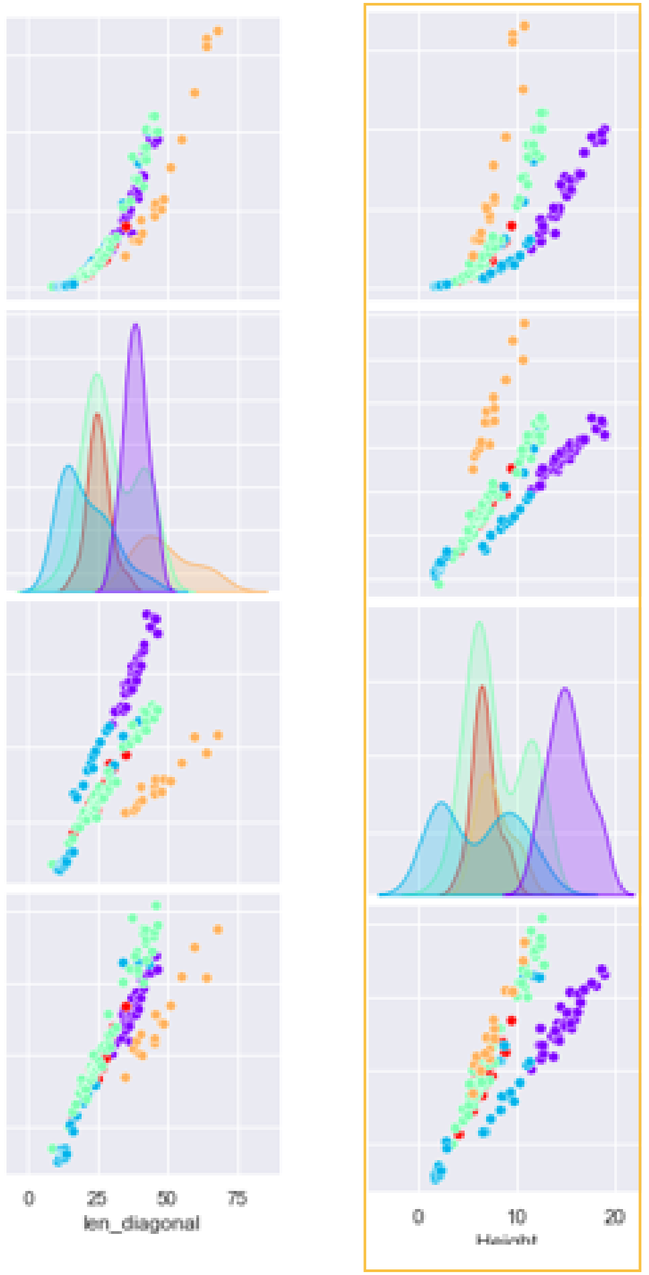
REGRESSIÓ LOGÍSTICA

SUGGERENCIES EN COL·LABORACIÓ

SUGGERÈNCIES

Feature Importance a Regressió Logísitica

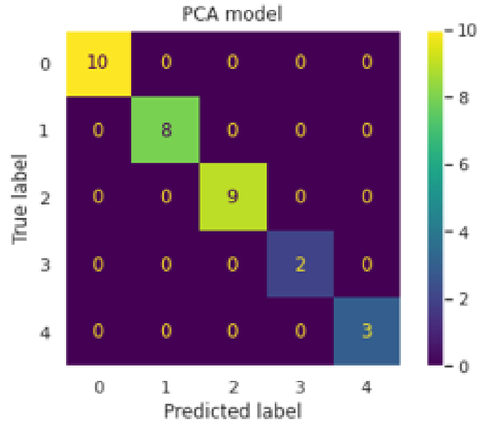


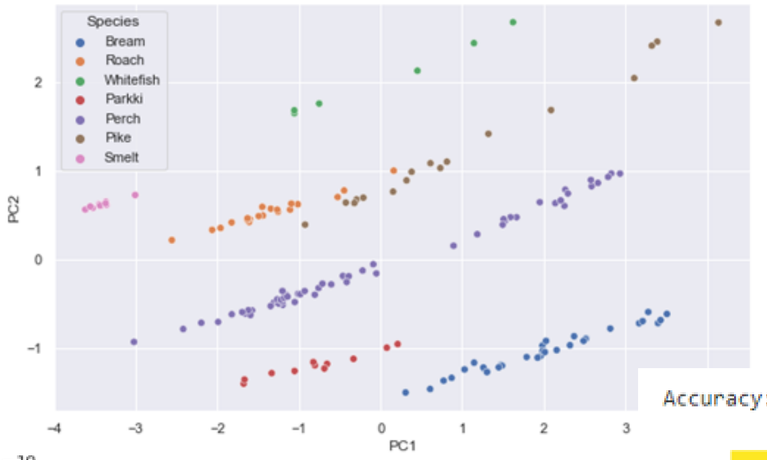


SUGGERÈNCIES

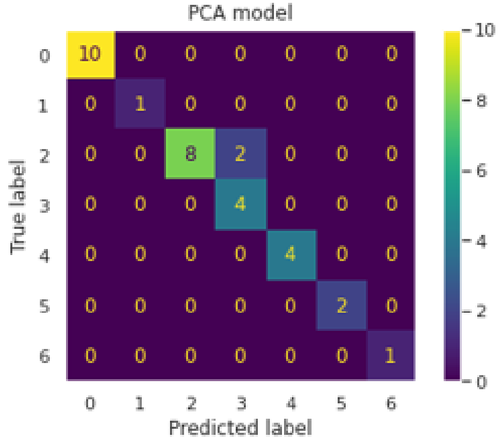
Regressió Logísitica amb components del PCA

Accuracy: 1.0





Accuracy: 0.9375



CONCLUSIONS

Reptes:

1

Saber on acabar el treball

2

Models hypersensibles a l'inicialització (poques dades)

Coneixements:

3

Objectius assolits satisfactòriament

PRÀCTICA 3

CAS KAGGLE

Serena Sánchez

APRENENTATGE COMPUTACIONAL