

# **AA - Project 1**

## **Bee Subspecies Classification using Machine Learning Methods over Bee Images**

**Presented by:**

Beatriz Costa nºmec 109657

Miguel Carvalho nºmec 80169

# Introduction



# Introduction

---

- Bees play a major role in the conservation and sustainability of the world's flora and fauna.
- Although far from being extinct, such a scenario would bring devastating consequences.

# Introduction

---

- To prevent it, solutions to resist threats to the bee populations, such as invasive species, must be created.
- Our classifier intends to assist those same solutions.

# Dataset



# Dataset

---

- Kaggle's "The BeelImage Dataset: Annotated Honey Bee Images" was selected as our dataset.

# Dataset

---

- Provides 5172 images of bees, and the subspecies they belong to, alongside additional information such as the location of the hive it was sighted at, date of sight, and more.


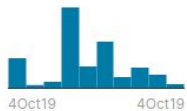
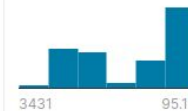

# Dataset

---

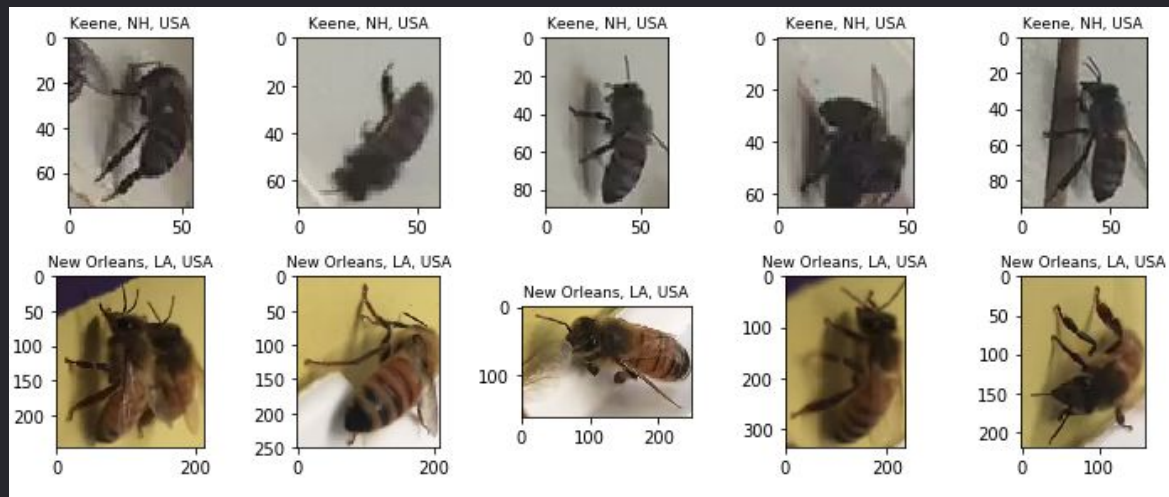
- Provided bee subspecies:
  - **Italian Honey Bee;**
  - **Russian Honey Bee;**
  - **Carniolian Honey Bee;**
  - **Mixed Local Stock;**
  - **-1 (later renamed to 'Unknown subspecies');**
  - **VSH Italian Honey Bee;**
  - **Western Honey Bee;**



# Dataset

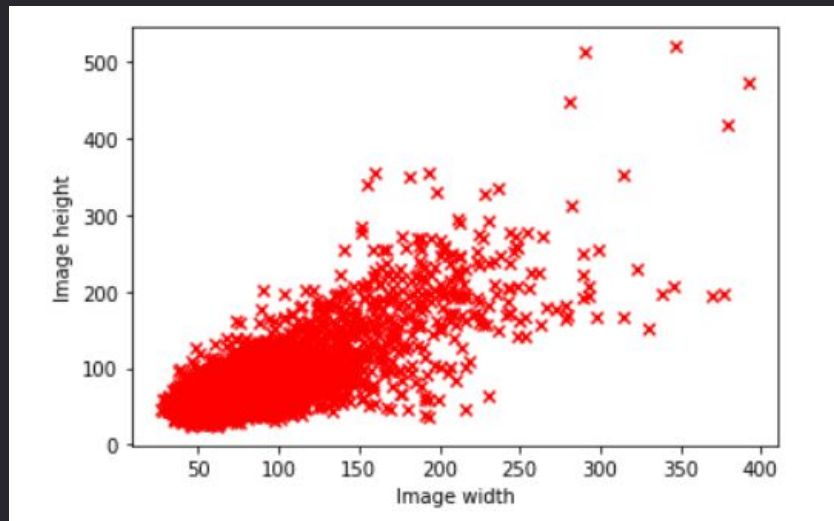
file	date	time	location	# zip code	subspecies	health	pollen_carrying	caste
File name in bee_imgs folder	Date of video captures	Time of day of video capture (military time)	Location (city, state, country)	Zip Code to numerically describe location	Subspecies of Apis mellifera species	Health of a bee	Presence of pollen on the bee's legs	Worker, Drone, or Queen bee
5172 unique values	 2Jul18 8Sep18	 4Oct19 4Oct19	Saratoga, CA, USA 39% Des Moines, IA, USA 19% Other (2199) 43%	 3431 95.1k	Italian honey bee 58% Russian honey bee 10% Other (1637) 32%	healthy 65% few varrao, hive be... 11% Other (1209) 23%	 true 18 0% false 5154 100%	1 unique value
041_066.png	8/28/18	16:07	Alvin, TX, USA	77511	-1	hive being robbed	FALSE	worker
041_072.png	8/28/18	16:07	Alvin, TX, USA	77511	-1	hive being robbed	FALSE	worker
041_073.png	8/28/18	16:07	Alvin, TX, USA	77511	-1	hive being robbed	FALSE	worker
041_067.png	8/28/18	16:07	Alvin, TX, USA	77511	-1	hive being robbed	FALSE	worker
041_059.png	8/28/18	16:07	Alvin, TX, USA	77511	-1	hive being robbed	FALSE	worker
041_071.png	8/28/18	16:07	Alvin, TX, USA	77511	-1	hive being robbed	FALSE	worker
041_065.png	8/28/18	16:07	Alvin, TX, USA	77511	-1	hive being robbed	FALSE	worker
041_064.png	8/28/18	16:07	Alvin, TX, USA	77511	-1	hive being robbed	FALSE	worker
041_070.png	8/28/18	16:07	Alvin, TX, USA	77511	-1	hive being robbed	FALSE	worker
041_058.png	8/28/18	16:07	Alvin, TX, USA	77511	-1	hive being robbed	FALSE	worker
041_074.png	8/28/18	16:07	Alvin, TX, USA	77511	-1	hive being robbed	FALSE	worker
041_060.png	8/28/18	16:07	Alvin, TX, USA	77511	-1	hive being robbed	FALSE	worker

# Dataset



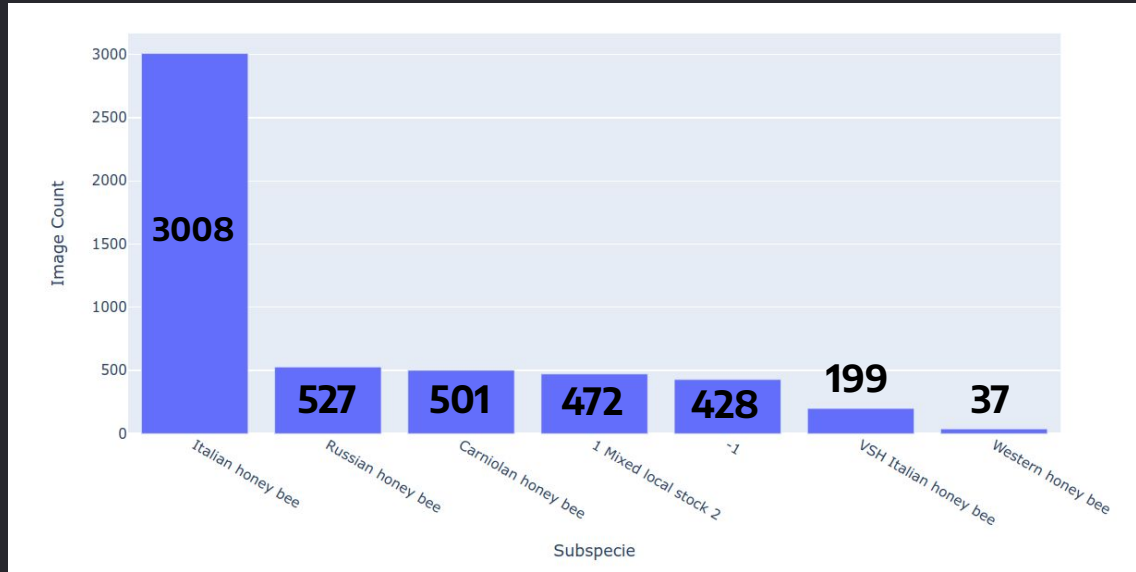
# Dataset - Issues

- Images exhibit different sizes:



# Dataset - Issues

- Image count per class differs:



# Dataset - Data Manipulation

---

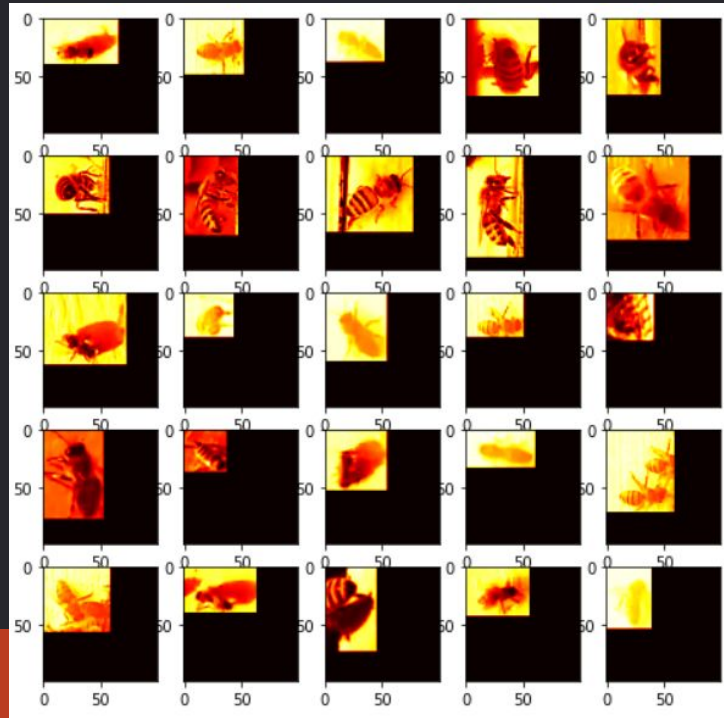
- Resizing Algorithms:
  - **Image Extending** - Images below the desired image size are filled with black colored pixels until the target size is reached, while the others are discarded.
  - **Image Rescaling** - Every image is rescaled to the desired height and width, regardless of their initial size.

# Dataset - Data Manipulation

Strategy	Advantages(+)/Disadvantages(-)
Image extending	<ul style="list-style-type: none"><li>+ Faster image processing</li><li>- Lower number of usable images</li><li>- Additional black pixels may mislead the classifier</li></ul>
Image rescaling	<ul style="list-style-type: none"><li>+ Most dataset images are usable</li><li>- Significant slower image processing</li></ul>

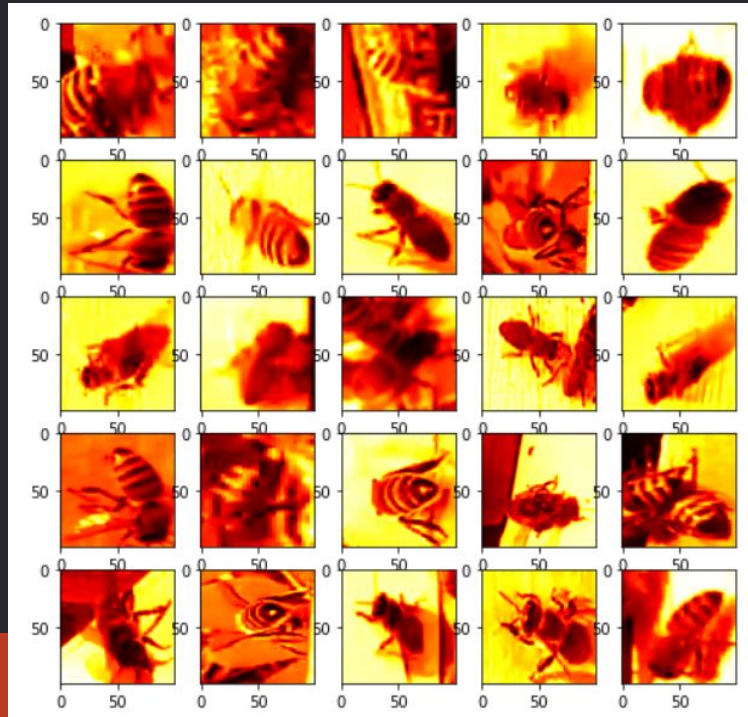
# Dataset - Issues

- Image Extending Algorithm:



# Dataset - Issues

- Image Rescaling Algorithm:





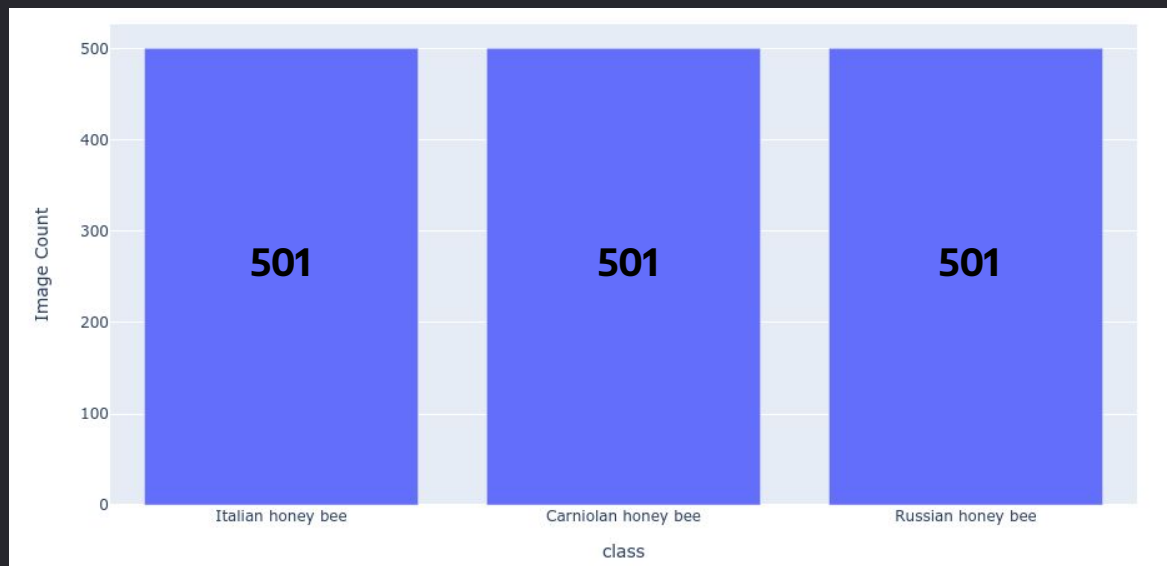
# Dataset - Data Manipulation

---

- Balancing Algorithm:
  - Minimum desired number of images per class as an input parameter
  - All classes with fewer images than the minimum are discarded
  - Number of images selected from each class is equal to the image count of the smallest remaining class

# Dataset - Data Manipulation

- Balancing Algorithm (500 images as minimum):



# Machine Learning Models



# ML Models

---

- Logistic Regression
- Neural Network

# ML Models - Load of Data

---

- Classes Mapping from Dictionary
- Holdout Method (80%-20%)
- Three-way split (60%-20%-20%)

# ML Models - Logistic Regression

---

- Implemented from the library *scikit-learn*
- Cross validation methods
  - K-Fold cross-validation
  - Stratified K-Fold cross-validation
  - Leave One Out cross-validation

# ML Models - Neural Networks

---

- Manually implemented and adapted from labs
- Hyper-parameters selected after evaluating metrics

# ML Models - Neural Networks

- Four phases:

	1st Phase	2nd Phase	3rd Phase	4th Phase
# Classes	4	2,3,4	3	3
Image Size	50x50	100x100 and 256x256	100x100	100x100
Data split	80%-20% and 70%-30%	Three-Way Split method	80%-20%	80%-20%
Method	Image Extension	Image Extension	Image Extension	Image Rescaling



# Results



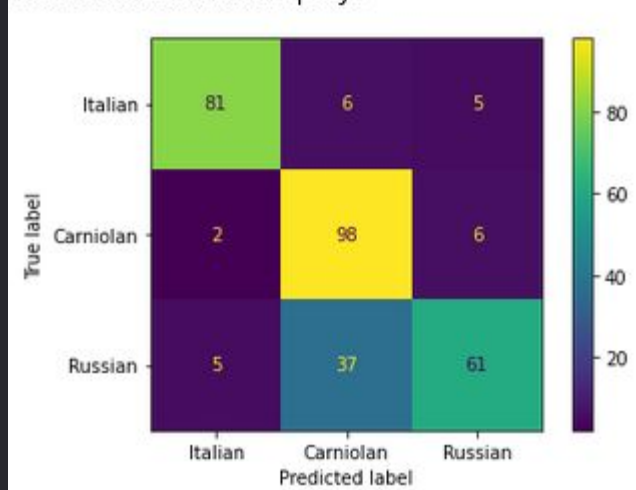
# Performance comparison - Metrics

---

- Accuracy;
- Confusion Matrix;
- Precision and Recall;
- F1-score;
- Cost loss function (not a metric);

# Results - LR Model

Confusion Matrix Display:



- Training Set accuracy: 100 %;
- Testing Set accuracy: 79.734 %;

Classification Report:

	precision	recall	f1-score	support
Italian honey bee	0.92	0.88	0.90	92
Carniolan honey bee	0.70	0.92	0.79	106
Russian honey bee	0.85	0.59	0.70	103
accuracy			0.80	301
macro avg	0.82	0.80	0.80	301
weighted avg	0.82	0.80	0.79	301

## Results - K-Fold cross-validation

---

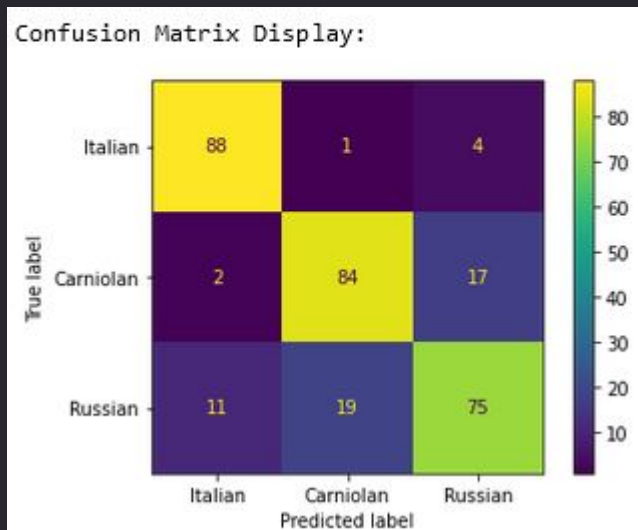
- Average accuracy: 76.133%
- Minimum accuracy: 71.667%
- Maximum accuracy: 80.663%
- Standard deviation: 0.03860

## Results - Stratified K-Fold cross-validation

---

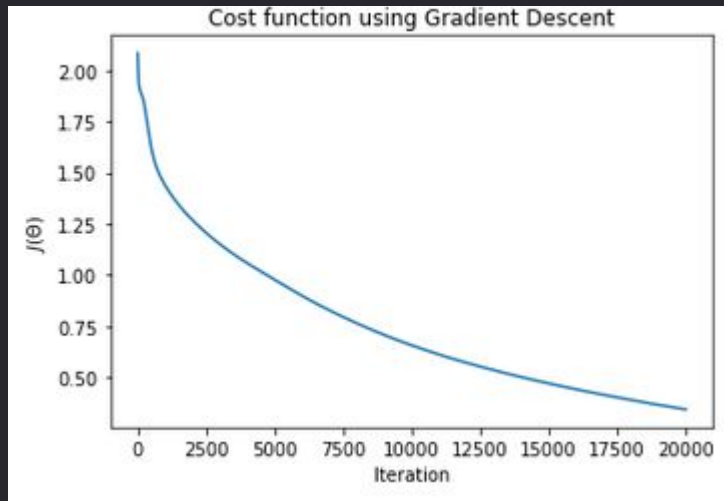
- Average accuracy: 76.023%
- Minimum accuracy: 71.111%
- Maximum accuracy: 79.005%
- Standard deviation: 0.03155

# Results - NN Model

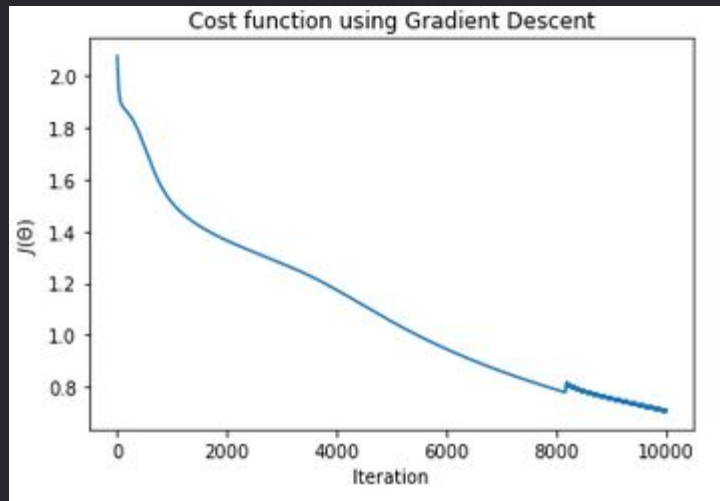


- **Training Set accuracy:** 100 %;
- **Testing Set accuracy:** 79.734 %;
- **Precision:** 0.946 (*Italian*), 0.816 (*Carniolan*), 0.714 (*Russian*)
- **Recall:** 0.946 (*Italian*), 0.808 (*Carniolan*), 0.781 (*Russian*)
- **F1-Score:** 0.946 (*Italian*), 0.812 (*Carniolan*), 0.746 (*Russian*)

# Results - Cost loss functions



- 3rd Phase



- 4th Phase

# Conclusion





# Conclusions

---

- **Better training set accuracy for LR**
- **Better testing set for the NN**
- **Results not fit from the same data in the LR and NN**
- **Phase three with best results**
- **Future approach with more tests for phase four and with higher number of iterations**
- **Balance algorithm should first oversample before discarding classes**
- **Tools enabling faster model training times should be used**

# The End

