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 BeeImage Dataset: Annotated Honey Bee Images" was selected as our dataset.

Dataset

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

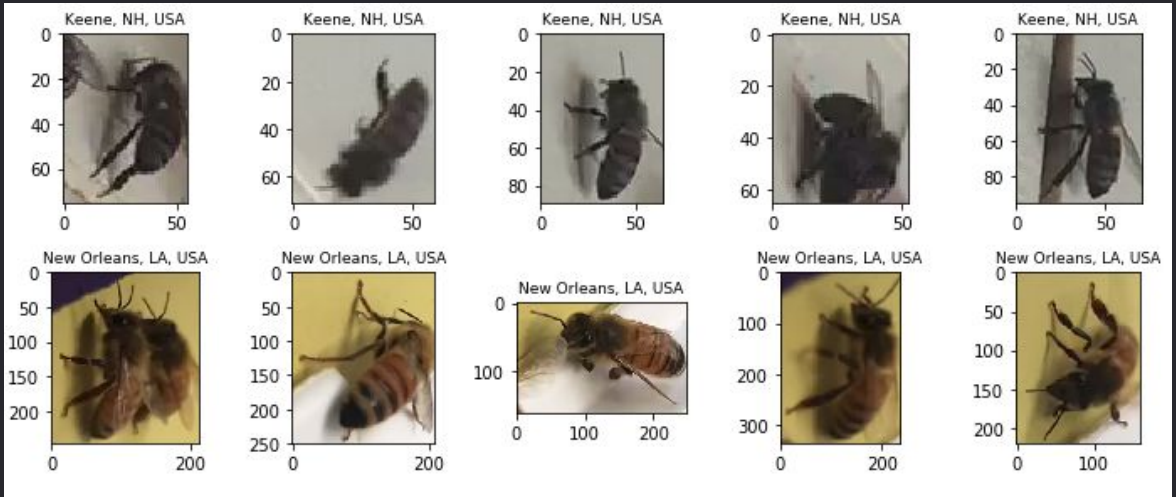
Dataset

- 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

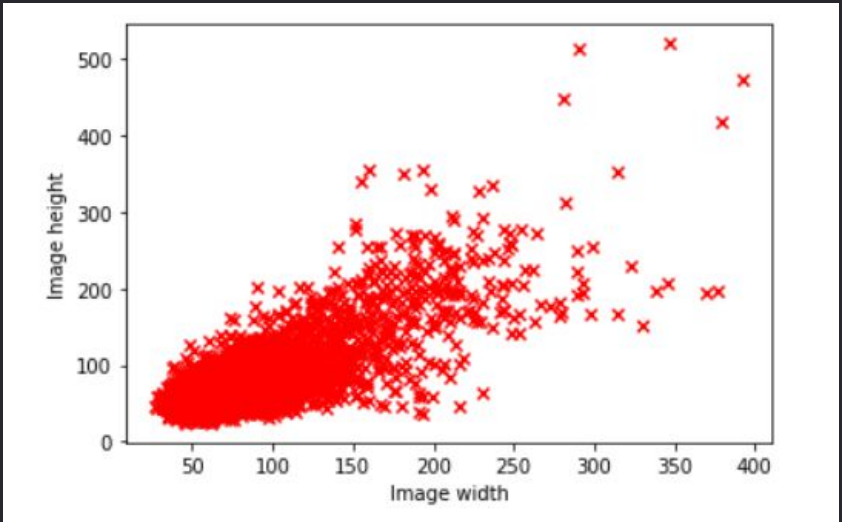
file	date	time	location	zip code	subspecies	health	pollen_carrying	caste
File name in bee_imgs folder	Date of video captures	Time of day of video capture (military time)	Location (city, state, country)	Zip Code to numerically describe location	Subspecies of Apis mellifera species	Health of a bee	Presence 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


Keene, NH, USA
Keene, NH, USA
Keene, NH, USA
Keene, NH, USA
Keene, NH, USA
New Orleans, LA, USA
New Orleans, LA, USA
New Orleans, LA, USA
New Orleans, LA, USA
New Orleans, LA, USA

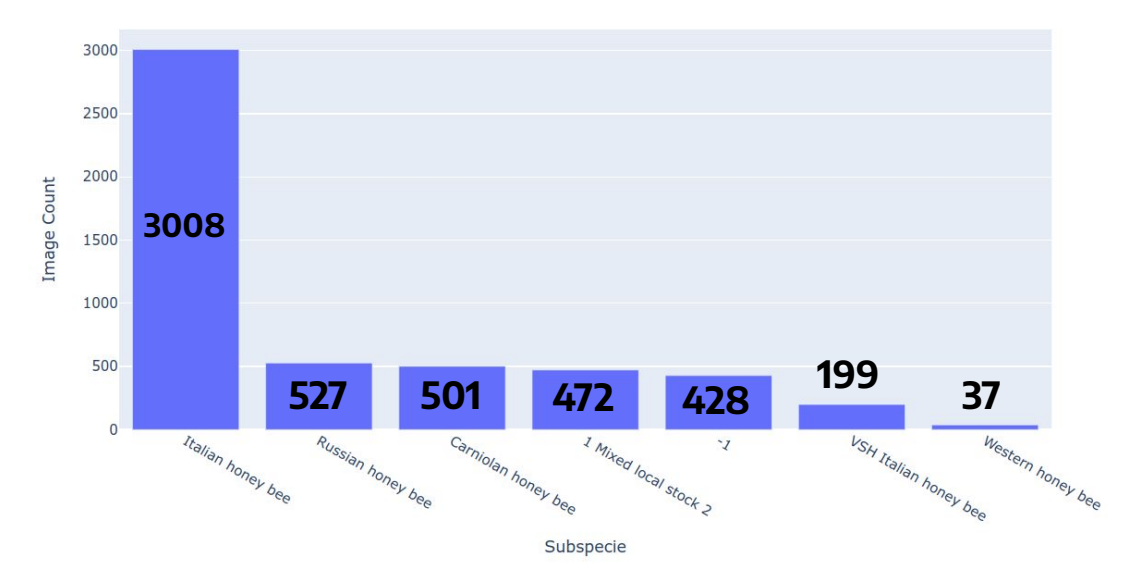
Dataset - Issues

- Images exhibit different sizes:



Dataset - Issues

- Image count per class differs:



Dataset - Data Manipulation

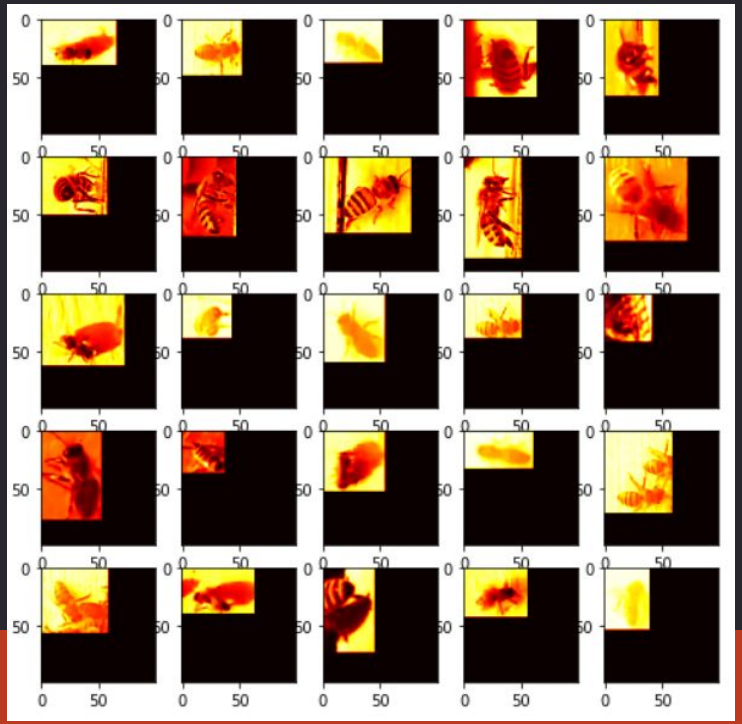
- Resizing Algorithms:
  - **Image Extending** - Images below the desired image size are filled with black colored pixels until the target size is reached, while the others are discarded.
  - **Image Rescaling** - Every image is rescaled to the desired height and width, regardless of their initial size.

Dataset - Data Manipulation

Strategy	Advantages(+)/Disadvantages(-)
Image extending	+ Faster image processing - Lower number of usable images - Additional black pixels may mislead the classifier
Image rescaling	+ Most dataset images are usable - Significant slower image processing

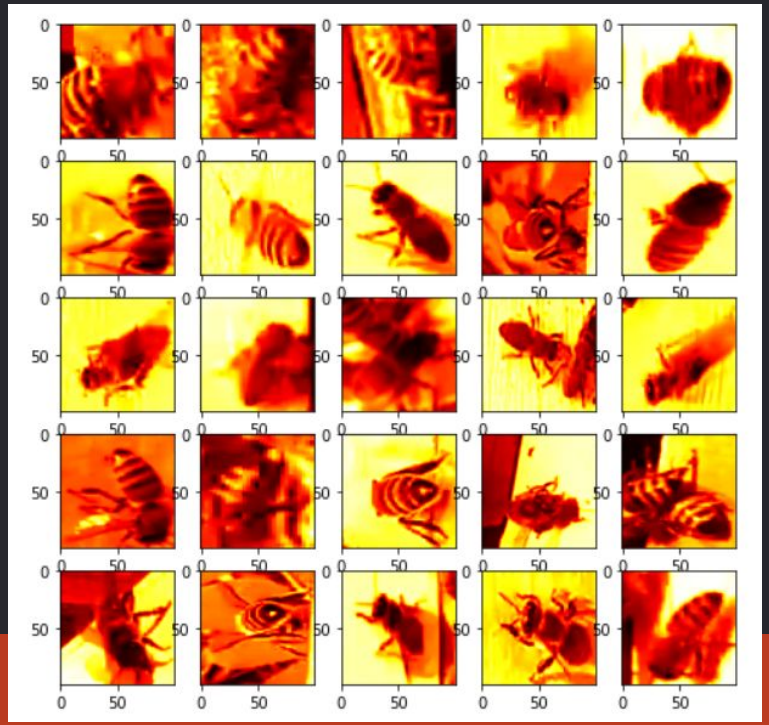
Dataset - Issues

- **Image Extending** Algorithm:



Dataset - Issues

- **Image Rescaling** Algorithm:

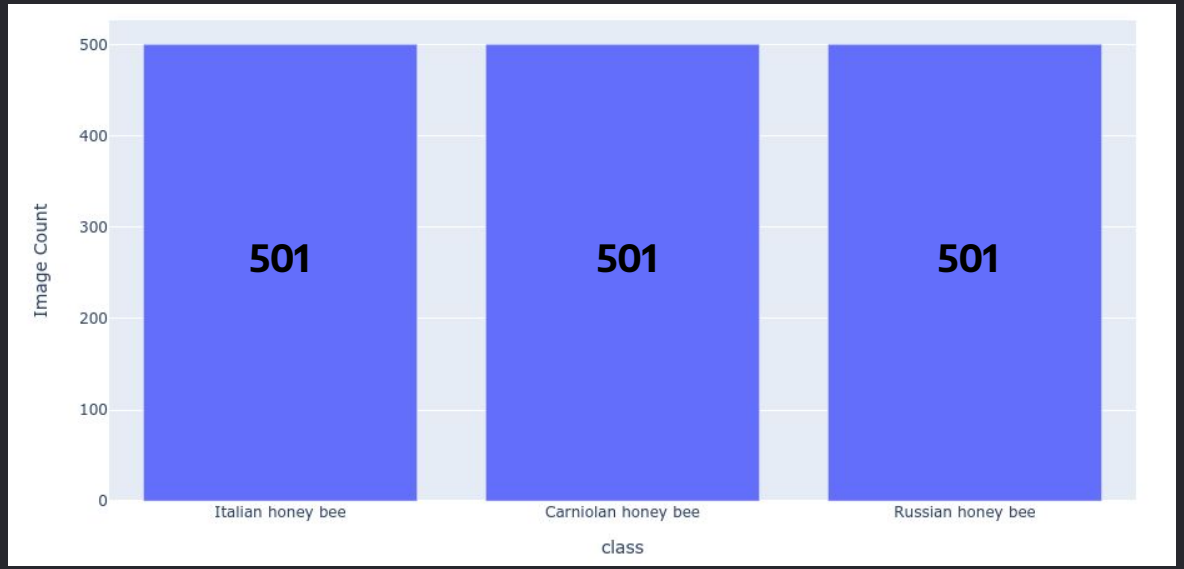


Dataset - Data Manipulation

- **Balancing Algorithm v.1**:
  - 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 v.1** (500 images as minimum):



Dataset - Data Manipulation

- **Balancing Algorithm v.2**:
  - Minimum desired number of images per class as an input parameter.
  - All classes with fewer images than **half of the minimum** are discarded.

Dataset - Data Manipulation

- **Balancing Algorithm v.2**:
  - Class with the least amount of images, from the classes with a number of images higher than the minimum threshold, limits the amount of images each class must have.

Dataset - Data Manipulation

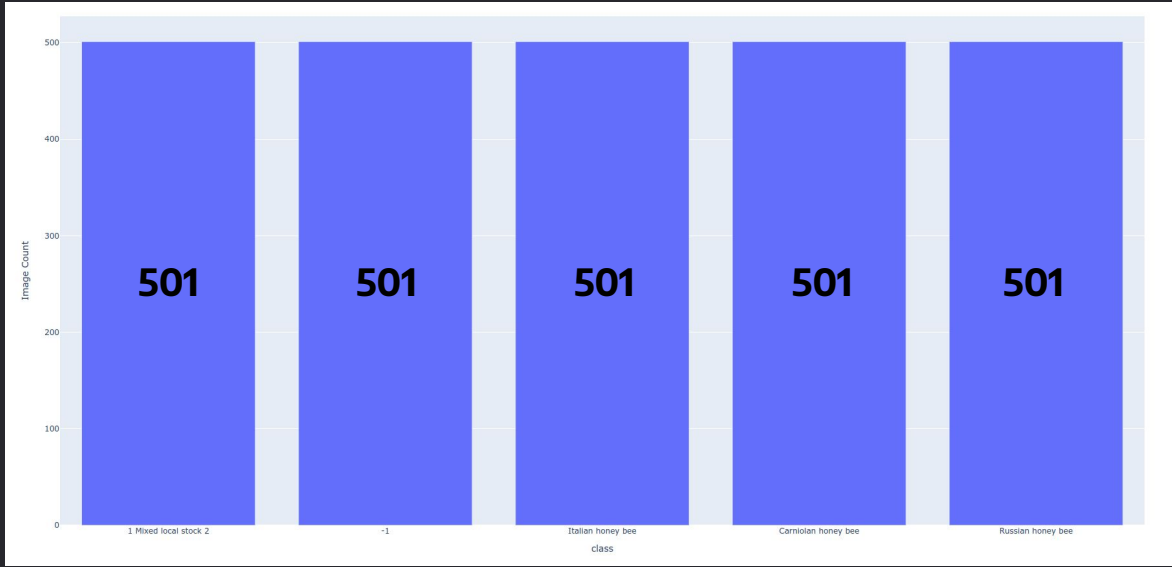
- **Balancing Algorithm v.2**:
  - Classes which display a number of images below the defined count but above half of it are oversampled until the limit amount is reached.

Dataset - Data Manipulation

- **Balancing Algorithm v.2**:
  - The oversampling process takes into account the amount of images needed to reach the limit and replicates and rotates (randomly between 90 and 270 degrees) that same number of images from that class.

Dataset - Data Manipulation

- **Balancing Algorithm v.2** (500 images as minimum):



Dataset - Data Manipulation

- The additional images and classes the **balancing algorithm v.2** is able to select were not used since this update was made in a later phase of the development of this project.

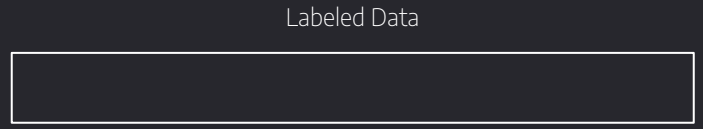
Machine Learning Models

ML Models

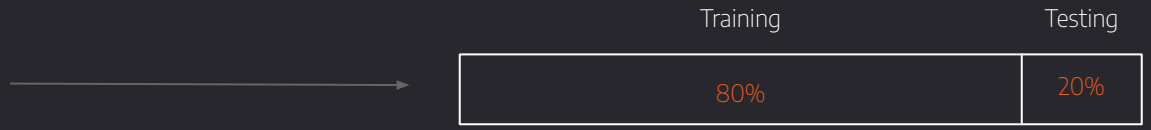
- Logistic Regression
- Neural Network

ML Models - Load of Data

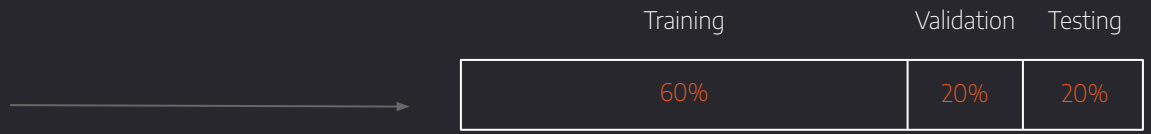
- Classes Mapping from Dictionary



- Holdout Method



- Three-way split



ML Models - Logistic Regression

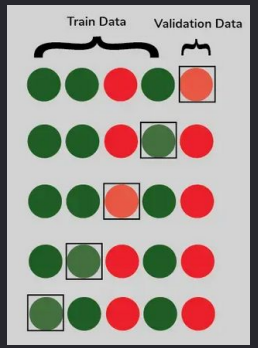
- Implemented from the library *scikit-learn*
- Cross validation methods
  - K-Fold cross-validation
  - Stratified K-Fold cross-validation
  - Leave One Out cross-validation

ML Models - Cross Validation

Example for 2 classes, K = 5

K-Fold

Stratified K-Fold



Leave One Out

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);

Obtained from classification report

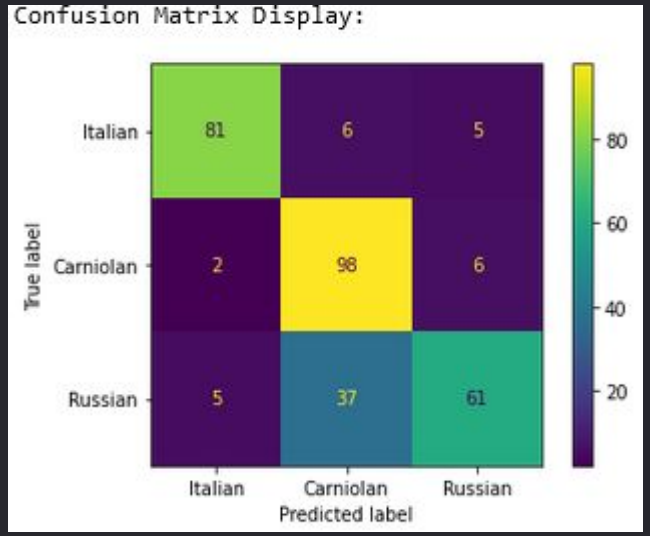
Classification Report:

	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

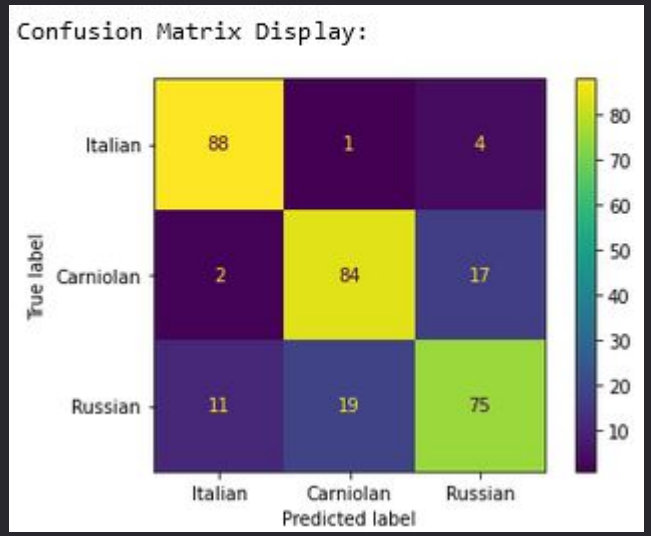
Results - LR ***vs*** NN Model Metrics

	LR	**NN**
Training Set accuracy	100 %	96.506%
Testing Set accuracy	79.734 %	82.060%
Precision	0.92 (*Italian*), 0.70(*Carniolan*), 0.85 (*Russian*)	0.946 (*Italian*), 0.816 (*Carniolan*), 0.714 (*Russian*)
Recall	0.88 (*Italian*), 0.92 (*Carniolan*), 0.59 (*Russian*)	0.946 (*Italian*), 0.808 (*Carniolan*), 0.781 (*Russian*)
F1-Score	0.90 (*Italian*), 0.79 (*Carniolan*), 0.70 (*Russian*)	0.946 (*Italian*), 0.812 (*Carniolan*), 0.746 (*Russian*)

Results - LR ***vs*** NN Model Confusion Matrix



Logistic Regression



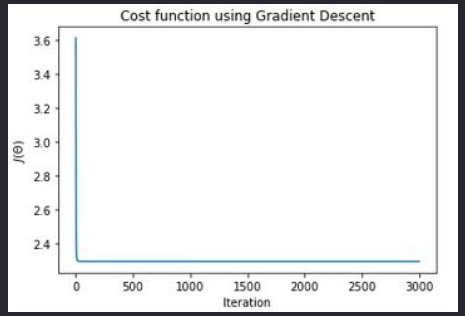
Neural Network

Results - K-Fold *vs* Stratified cross-validation

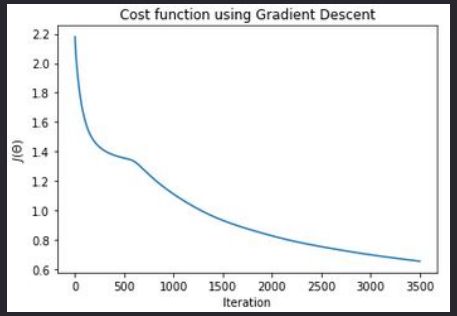
3 classes, K = 5

	K-Fold	**Stratified**
Average accuracy	76.133%	76.023%
Minimum accuracy	71.667%	71.111%
Maximum accuracy	80.663%	79.005%
Standard deviation	0.03860	0.03155

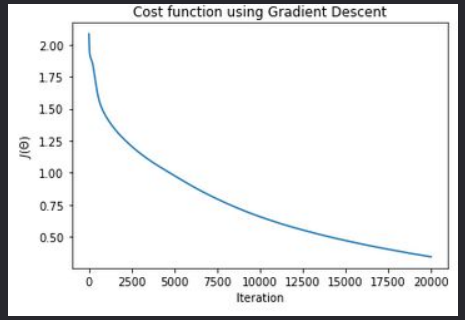
Results - Cost loss functions



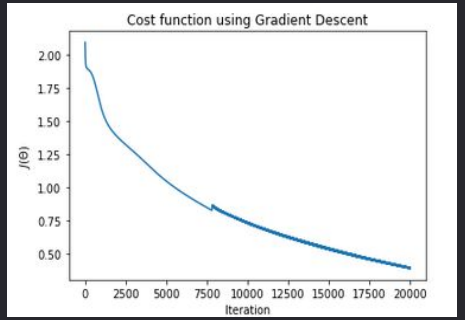
1st Phase



2nd Phase



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 four with the best result

Improvements

Improvements

- Improved balance algorithm should be used for model training
- Tools enabling faster model training times should be used

The End