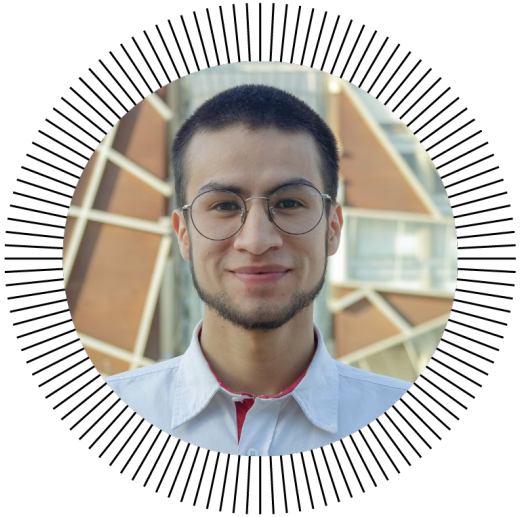
Categorization Challenge

Datarangers

Datarangers Team



Edgar Steven Baquero

Jorge Daniel Gutierrez

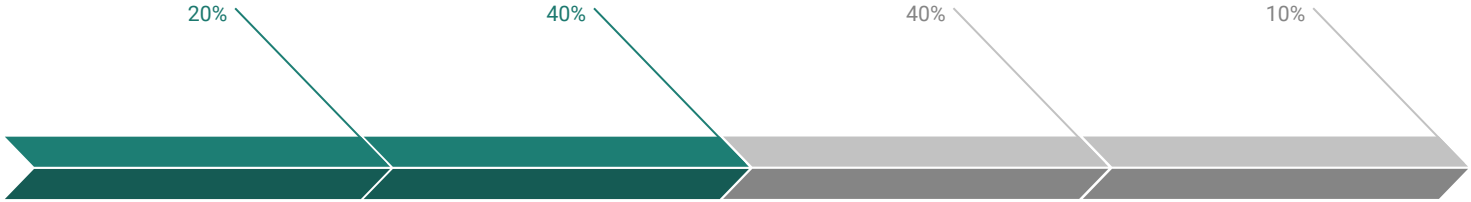
The story of Failures Learning...

A happy story with a sad ending:

- How we handle the problem
- Results of the happy part
- Results of the sad part
- Some conclusions

Handling the problem

Handling the problem - Initial insights



Work with sklearn dataset

Due to data upload time when working with Google Colab

Try some baseline models

Selecting best models

Parameters

Improve model selection

Go to Challenge's Data

Use the models and its parameters to train challenge's Data

Handling the problem - Sklearn dataset

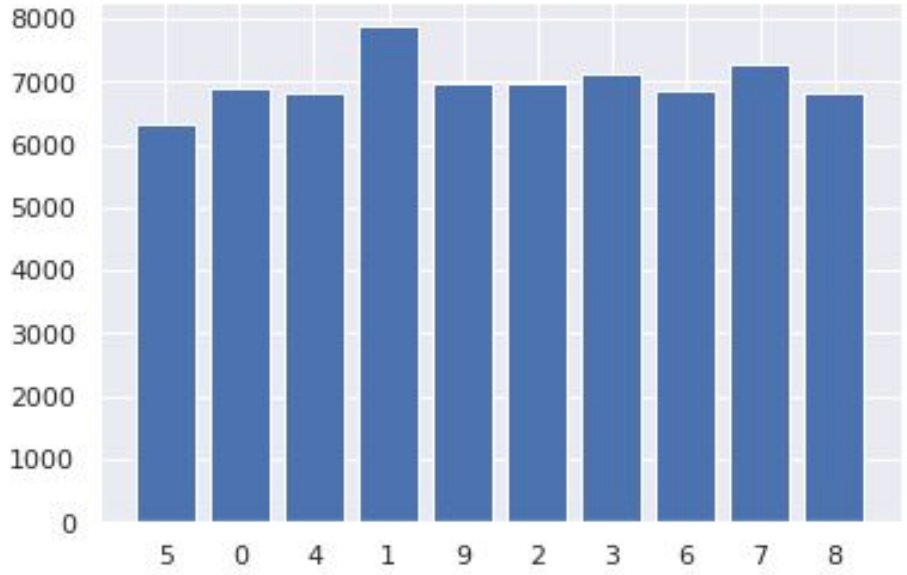
- We flatten image data
- It lead us to a classical regression problem

We splitted the Train dataset into

Train	Test
80%	20%

Handling the problem - Distribution of categories

It seems like data is balanced:

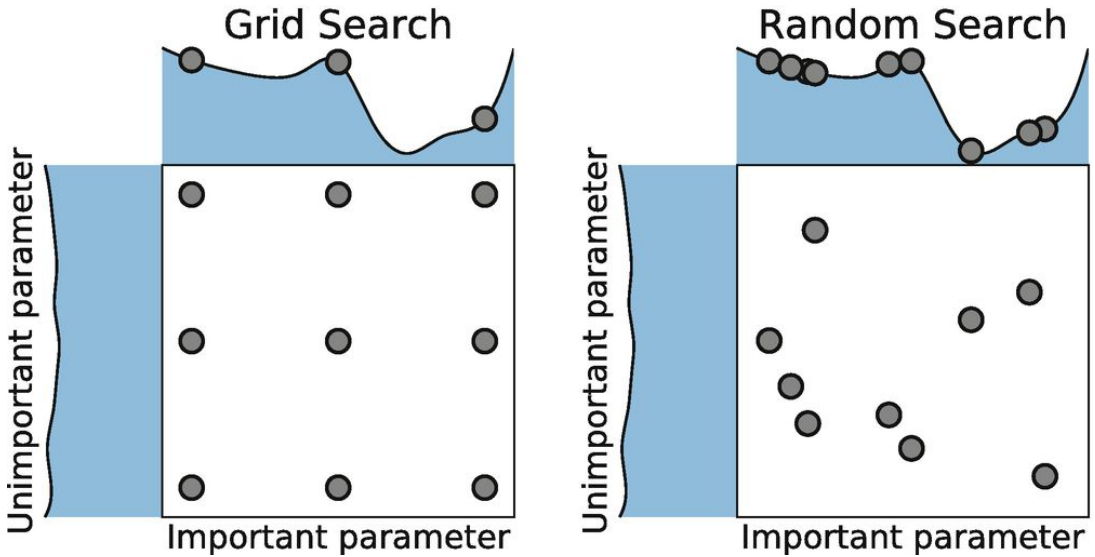


Handling the problem - Models

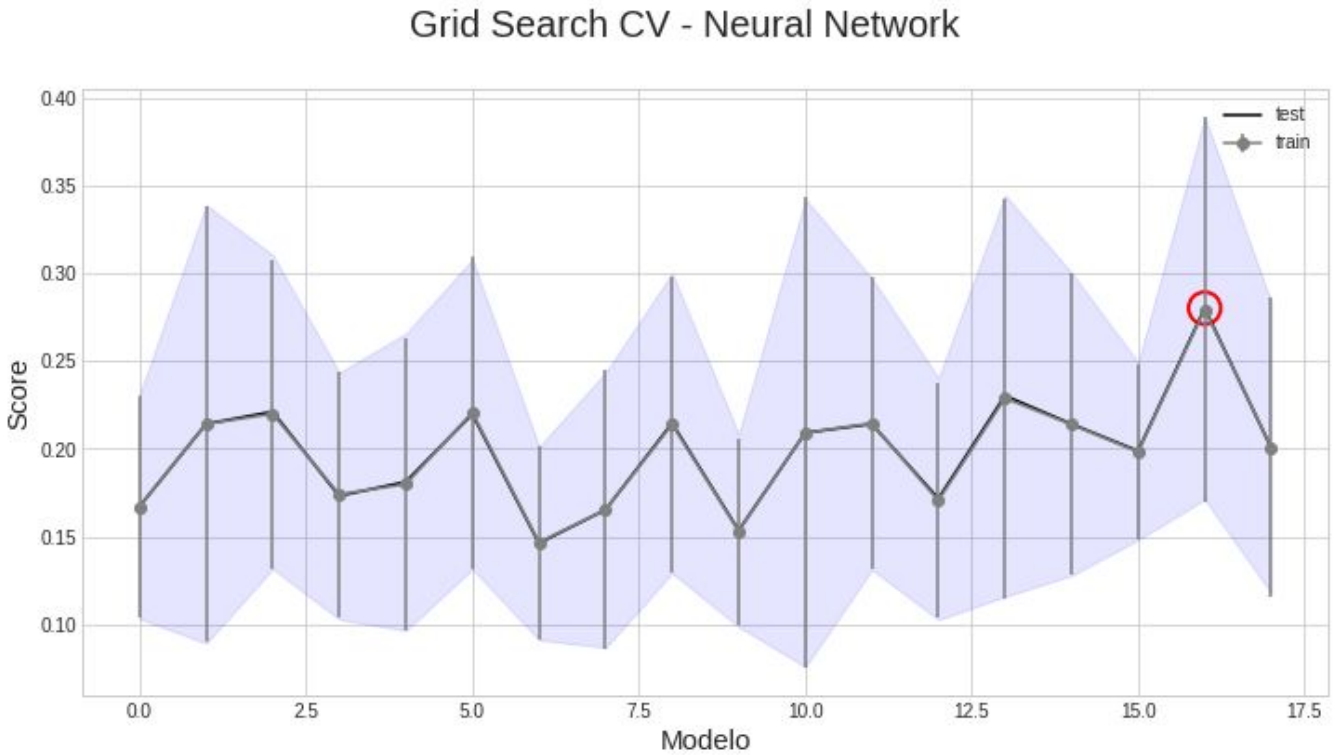
1. Mult. Regression
2. Linear Discriminant Analysis
3. Logistic Regression
4. Neural Network
5. Tree
6. Quadratic Discrimination Analisys
7. SVM
8. Boosting

Handling the problem - Model Selection

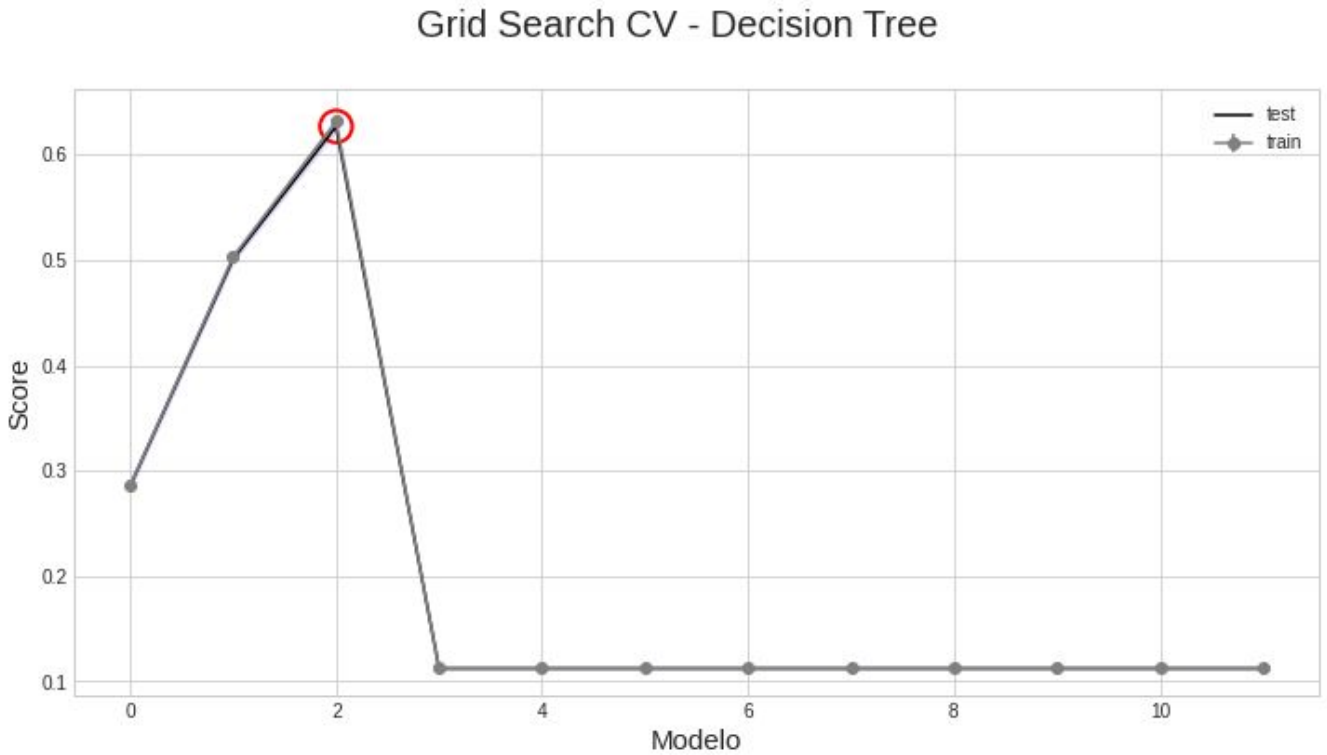
We used Grid Search to select models



Model Selection – Results

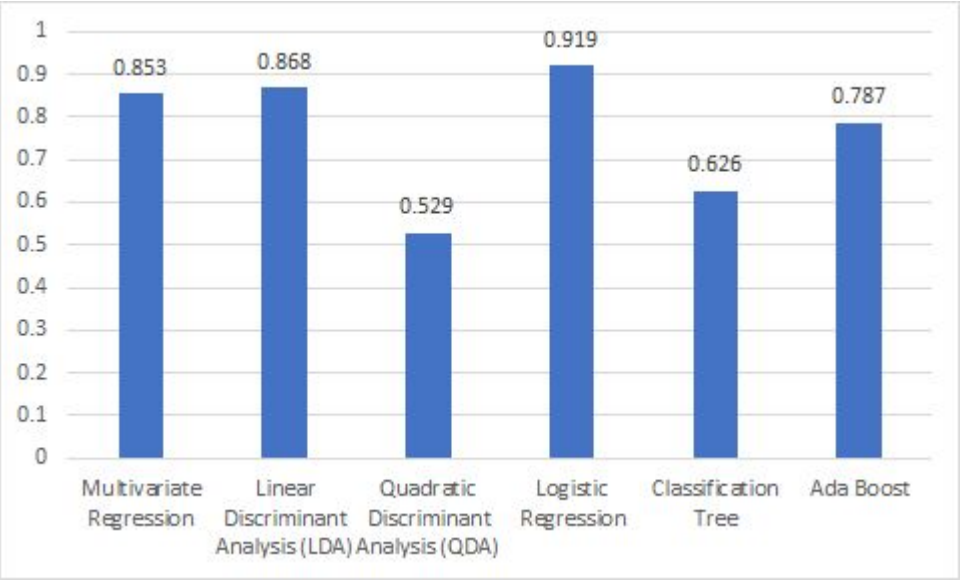


Model Selection – Results



Model Selection – Results

Model Selection – General Performance



Plot twist: Challenge's Data

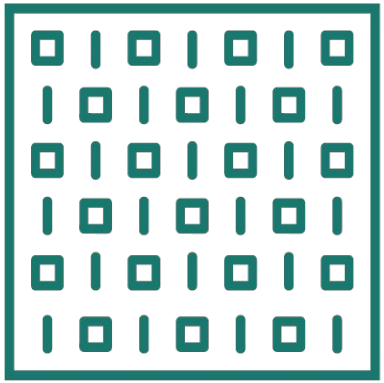
mnist sklearn dataset

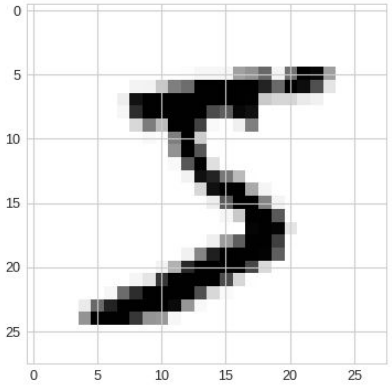
mnist bhchallenge dataset

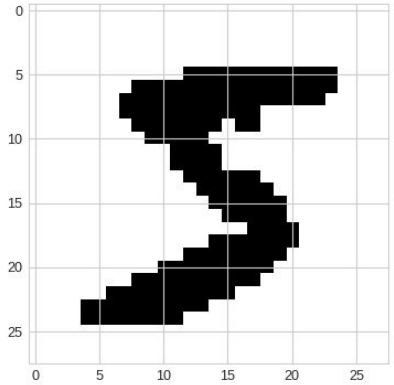
"Everything went wrong when we changed dataset"

One way to handle challenge data

- We noticed that min(train_image) = 22, while min(test_image) = 0
- Also noticed that max(train_image) = 220, while max(test_image) = 255
- Standarize with Boolean Masks!







To improve for the next Challenge

- Display the data before the challenge dataset
- Convolutional neural networks
- Take naps