# classification algorithm

## Gaussian Bayes Classifier (non Naive)

presentation by
Anyinssan Nava Sanchez

# Business problem

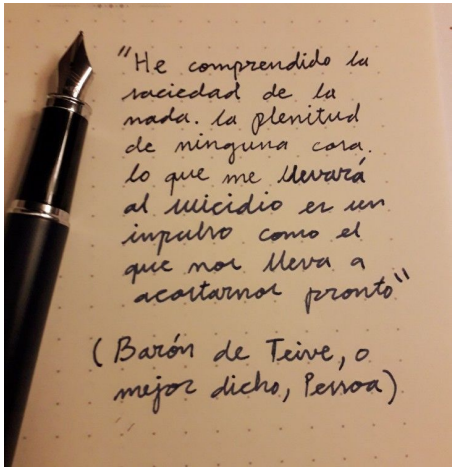
Optical Character Recognition


upward trend

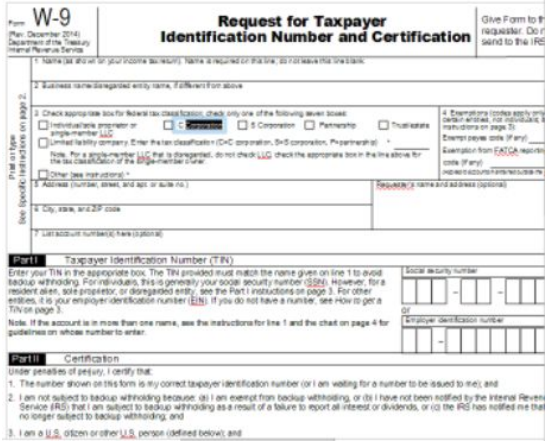

Save man hours

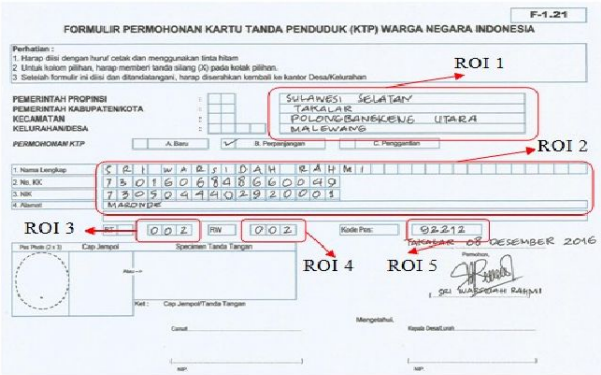
# Challenges in Handwriting Recognition

- Huge variability and ambiguity of strokes from person to person
- Handwriting style of an individual person also varies time to time and is inconsistent
- Poor quality of the source document/image due to degradation over time
- Cursive handwriting makes separation and recognition of characters challenging


"He comprendido la sociedad de la nada. La plenitud de ninguna cara. Lo que me llevará al suicidio es un impulso como el que nos lleva a acortarnos pronto"
(Barón de Teive, o mejor dicho, Pessoa)

?

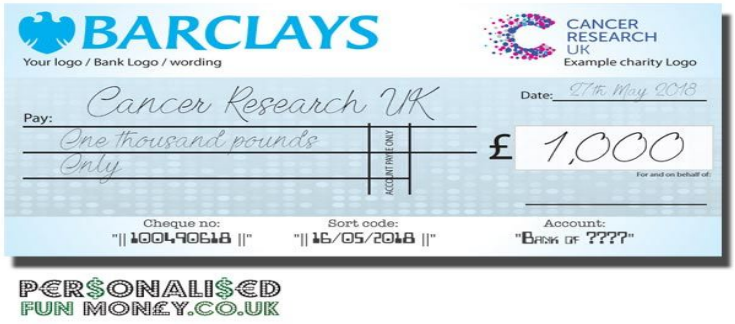

W-9
Request for Taxpayer Identification Number and Certification
Part I Taxpayer Identification Number (TIN)
Part II Certification

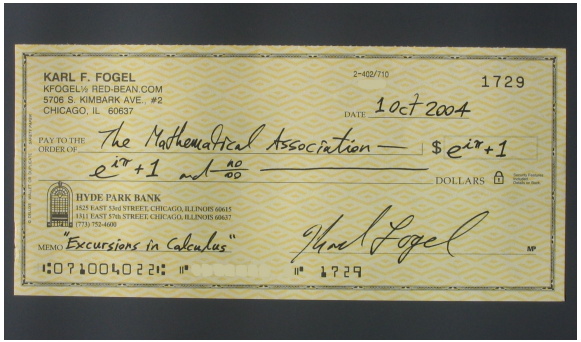

F-1.21
FORMULIR PERMOHONAN KARTU TANDA PENDUDUK (KTP) WARGA NEGARA INDONESIA
ROI 1
ROI 2
ROI 3
ROI 4
ROI 5
SULAWESI SELATAN
TAKALAR
POLONGBANGKENG UTARA
MALEWANG
92212
TAKALAR 08 DESEMBER 2016

# Use case : Banking

People write cheques on a regular basis and cheques still play a major role in most non-cash transactions.

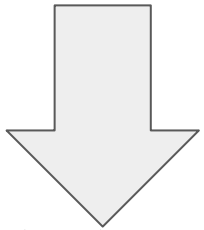
In many developing countries, the present cheque processing procedure requires a bank employee to read and manually enter the information present on a cheque and also verify the entries like signature and date



# ML problem

From a business perspective, you have already identified the type of outcome you need to produce.

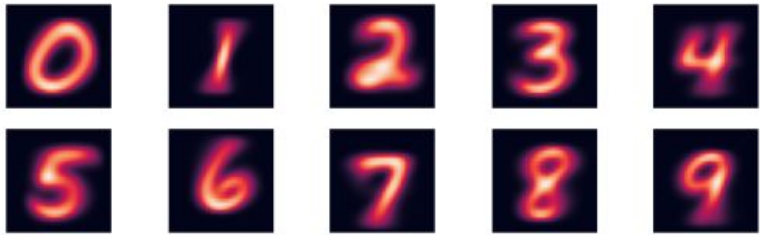


Now consider how to express that outcome in terms of various outcomes a machine learning model might produce. For example, regression, classification, clustering, and so forth.

# Data

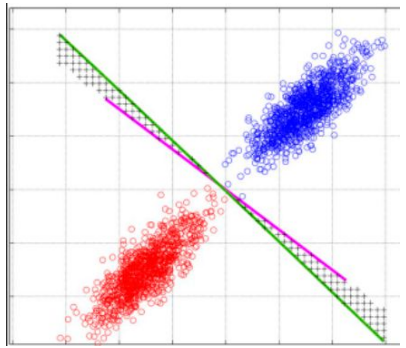
We have a set of data to train and test. The labels are the numbers from 0 to 9 and based on the characteristics of the data X, the labels will be classified.

How the labels look



# Model selection

The Gaussian Bayes Classifier (non Naive) model was selected to solve this problem.



# Metrics

accuracy for test data : 0.95805

accuracy for training data: 0.9511

Confusion matrix

| predictions / actual | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 968 | 0 | 1 | 1 | 0 | 0 | 3 | 1 | 6 | 0 |
| 1 | 0 | 1120 | 6 | 1 | 0 | 0 | 5 | 0 | 3 | 0 |
| 2 | 3 | 3 | 973 | 9 | 3 | 0 | 1 | 6 | 33 | 1 |
| 3 | 5 | 0 | 4 | 943 | 0 | 12 | 0 | 5 | 31 | 10 |
| 4 | 0 | 2 | 3 | 0 | 938 | 0 | 4 | 2 | 3 | 30 |
| 5 | 2 | 0 | 2 | 20 | 0 | 820 | 12 | 3 | 25 | 8 |
| 6 | 6 | 3 | 1 | 0 | 3 | 9 | 928 | 0 | 8 | 0 |
| 7 | 0 | 11 | 13 | 2 | 10 | 1 | 0 | 944 | 5 | 42 |
| 8 | 8 | 5 | 7 | 14 | 3 | 4 | 2 | 4 | 918 | 9 |
| 9 | 5 | 6 | 4 | 8 | 6 | 1 | 0 | 6 | 14 | 959 |

# Conclusions

- An accuracy was obtained in the test data very similar to that of the training data, which means that the model has been well trained and there are no problems of overfitting or predictive deficiency.

- With the implementation of the algorithm, costs and hours of work can be reduced to make work more efficient.

- The solution can be expanded to other similar problems

# Thanks for your attention