# classification algorithm Gaussian Bayes Classifier (non Naive)

presentation by
Anyinssan Nava Sanchez



# Business problem

**Optical Character Recognition** 



upward trend



#### Save man hours



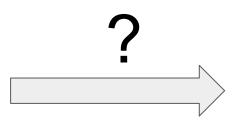


# Challenges in Handwriting Recognition

- 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 raciedad de la nada la glenitud de minguna cora lo que me llevará al micidio er un inpulso como el que mor lleva a acortarnol pronto"

(Barón de Teive, o mejor dicho, l'erroa)



Decar	W-9 December 2014) trace of the Treasury of Revenue Service	Request for Taxpa Identification Number an	yer d Certific	ation	Give Form requester, send to the	Don			
Print ortiges See Specific leatherform on page 2.	1. Name (as shown or your income bulletury). Name is required on this line, so not leave this line blank.								
	2 Sucress rane Sangarded entity rame, if offerent from above								
	3 Check appropriate local for depart lac creations of check plot one of the following even local:    Interchality express or   Interchality   Companies   Anterior   Interchality express or   Interchality express or   Interchality   Companies   Interchality   Inter								
	il Dig, state, and ZP code								
	7 Listacquie rumawijo	are (optional)				_			
Par	Taxpayer In	entification Number (TIN)				_			
Enter	your TIN in the appropri	ate box. The TIN provided must match the name given on line 1 to due to this is generally your social security number (SSN). However,		starty name		_			
reside entité	ent alien, sole proprietor,	or disregarded entity, see the Part I instructions on page 3. For or intification number (EIN): If you do not have a number, see How or	her	∐-L	∐-∐	Т			
	If the account is in more	er dert ficasio	rificación number						
gude	lines on whose numbers	o enter.		-					
Par	till Certification					-			
Unde	r penalties of pegury, I o	HSFy that:							
1. TR	te number shown on this	form is my correct taxpayer identification number (or I am waiting	for a number to be	issued to me	ric and				
86		witholding because, (d) I am exempt from backup withholding, o ject to backup withholding as a result of a failure to eport all inter- withholding; and							
	en a II G. ridnen nomber	U.S. person (defined below); and							

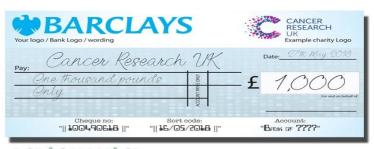
FORMULIR P	ERMOHONAN KARTU TAN	IDA PENDUDUK (KTP) WARGA NEGARA IND	F-1.21 ONESIA
	dan menggunakan linta hitam mberi tanda silang (X) pada kotak pilih tandatangani, harap diserahkan kemba		1
PEMERINTAH PROPINSI PEMERINTAH KABUPATENIKO KECAMATAN KELURAHANIDESA	OTA.	SULAWESI SELATAN TAKALAR POLONGBANEKENE UTARA MALEWANE	
PERMOHOMAN KTP	A Baru R F	experijongen C. Penggurtien	ROI
3.NK 73	016068486 050444029	20001	
Per Print (2+3) Cap Jampol	O O Z RW O		ESEMBER 20
(.)	Ket: Cap Jempol/Tanda Tangan	ROI 4 ROI 5	nu Ranjai
****	Const	Mengetahul. Kepala Desat, oran	
	-	l Name	

mi a 3 di 21 e - 11 22 e - 121 - ex d - 1

## Use case: Banking

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

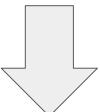
In many developing countries, the present cheque processing procedure requires a bank employee to read and manually enter the information present on a cheque and also verify the entries like signature and date





### 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















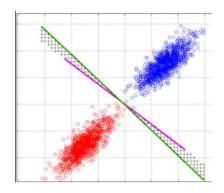






#### Model selection

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	S
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