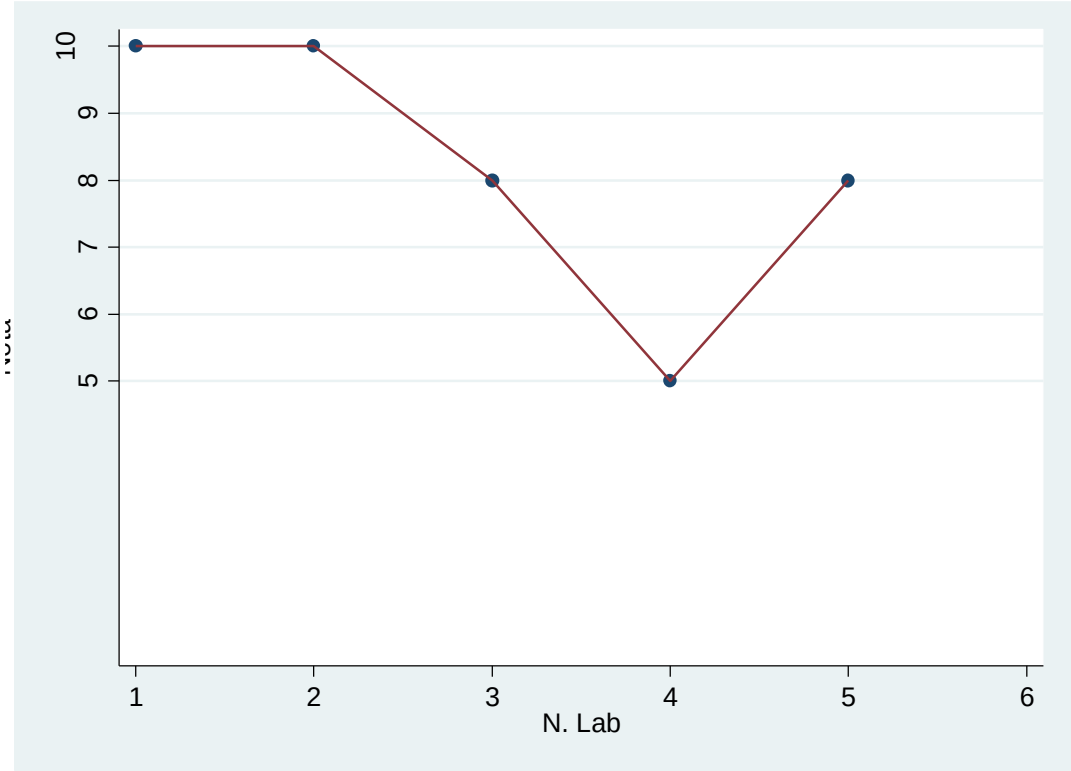
Clase 23

Machine Learning

ECON. Y CIENCIA DE DATOS - EAE 253B

I SEM 2019

Supervised Learning



Econometría vs ML

Econometría	ML
Busca explicar el pasado y obtener conclusiones para el futuro	Busca predecir el resultado (“Y”) de experiencias futuras (“X”)
Modelo específico es muy relevante (motivación teórica)	Modelo específico es irrelevante
Importancia en coeficientes (dy/dx) y significancia	Valor de coeficientes es irrelevante

Econometría vs ML

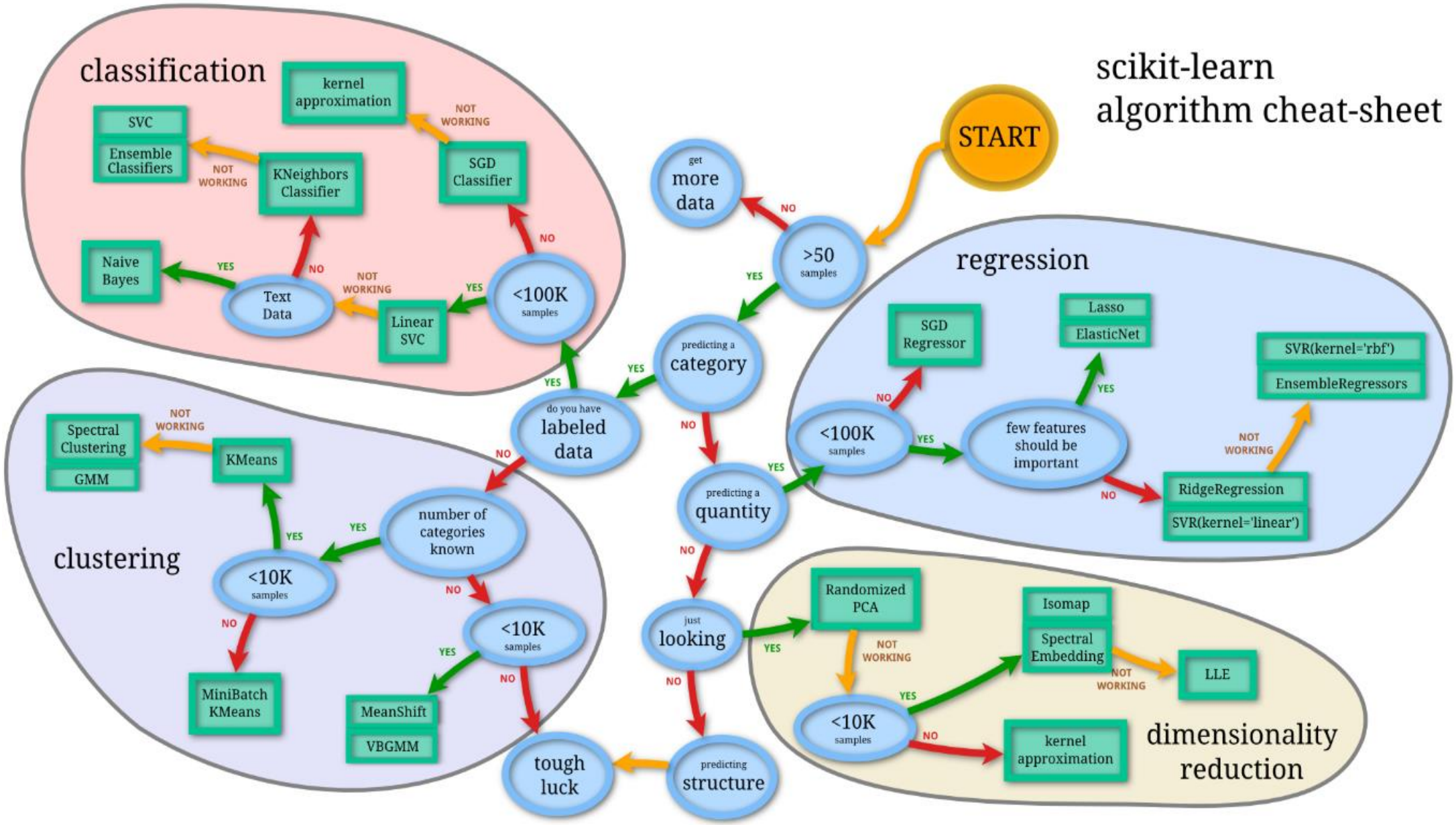
Econometría	ML
“Performance” no es tan relevante (R^2)	Performance es muy relevante; métrica difiere según tipo de problema
Evaluación dentro de muestra	Evaluación fuera de muestra
Mucho énfasis en poder argumentar causalidad	Lo único que interesa es el poder predictivo

Modelos de Predicción / Clasificación

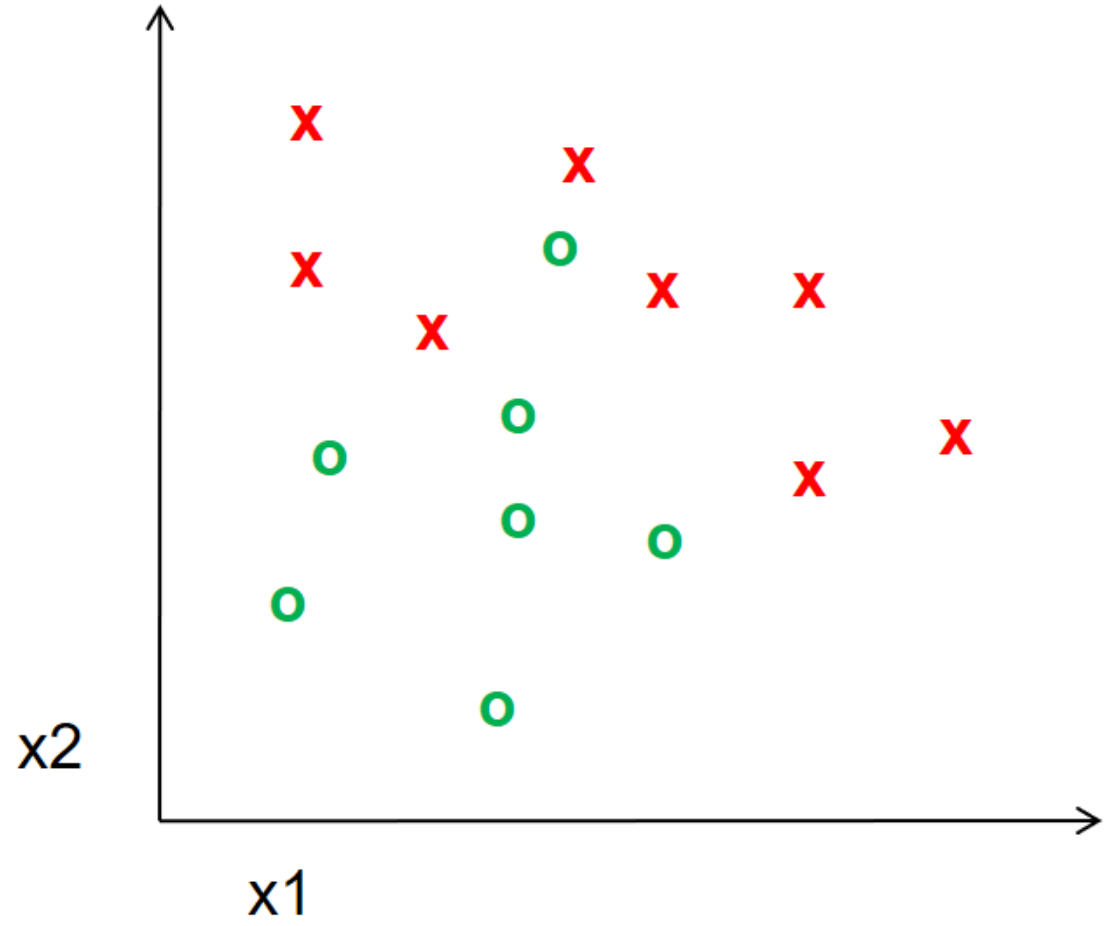
- Regresión lineal
  - OLS / 2SLS
  - Lasso
  - Ridge
  - Panel (FE / RE) o Series de tiempo
- Modelos estadísticos
  - Logistic (Logit) / Probit
  - Poisson
  - Duración

Modelos de Predicción / Clasificación

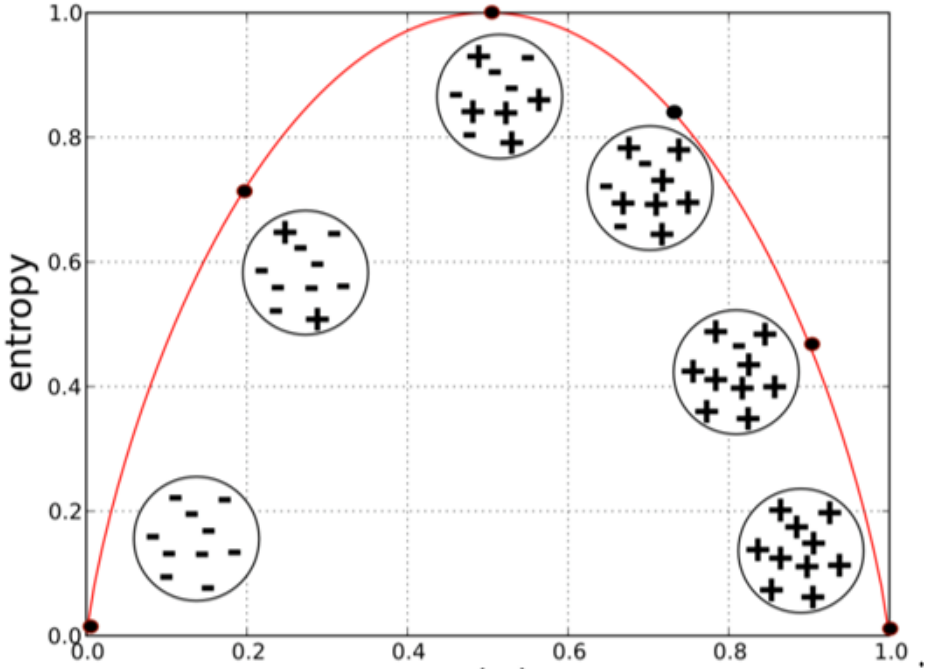
- K-vecinos más cercanos (K-NN)
- Decision Trees
- Support Vector Machines
- Bayes Classifier
- Ensembles
  - Bagging
  - Boosting
  - Random Forests
- Neural Networks / Deep learning


scikit-learn
algorithm cheat-sheet
START
>50 samples
NO
get more data
YES
predicting a category
YES
do you have labeled data
NO
predicting a quantity
NO
just looking
NO
predicting structure
tough luck
YES
<100K samples
NO
Linear SVC
NOT WORKING
Text Data
YES
Naive Bayes
NO
KNeighbors Classifier
NOT WORKING
Ensemble Classifiers
SVC
NOT WORKING
SGD Classifier
NO
kernel approximation
classification
YES
<100K samples
NO
SGD Regressor
NOT WORKING
Lasso
ElasticNet
YES
few features should be important
NO
RidgeRegression
SVR(kernel='linear')
NOT WORKING
SVR(kernel='rbf')
EnsembleRegressors
regression
YES
number of categories known
NO
<10K samples
NO
MiniBatch KMeans
YES
KMeans
NOT WORKING
Spectral Clustering
GMM
YES
<10K samples
NO
MeanShift
VBGMM
clustering
YES
Randomized PCA
NOT WORKING
Isomap
Spectral Embedding
NOT WORKING
LLE
YES
<10K samples
NO
kernel approximation
dimensionality reduction

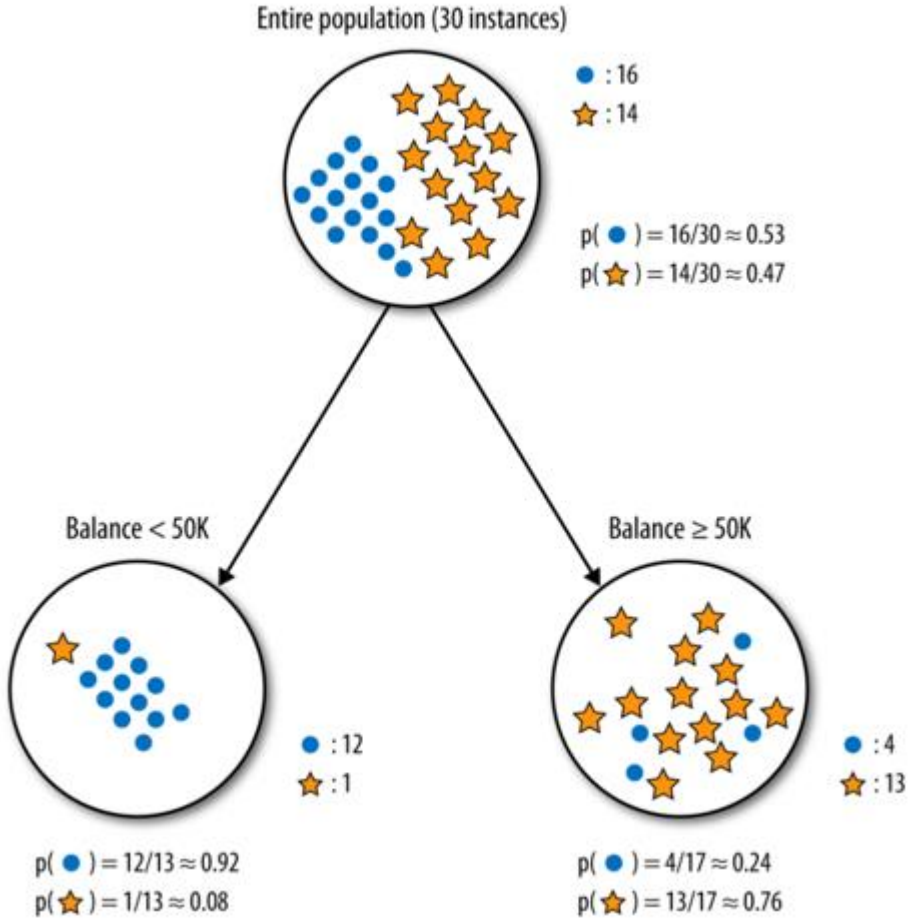
“Modelo” más simple: K-NN



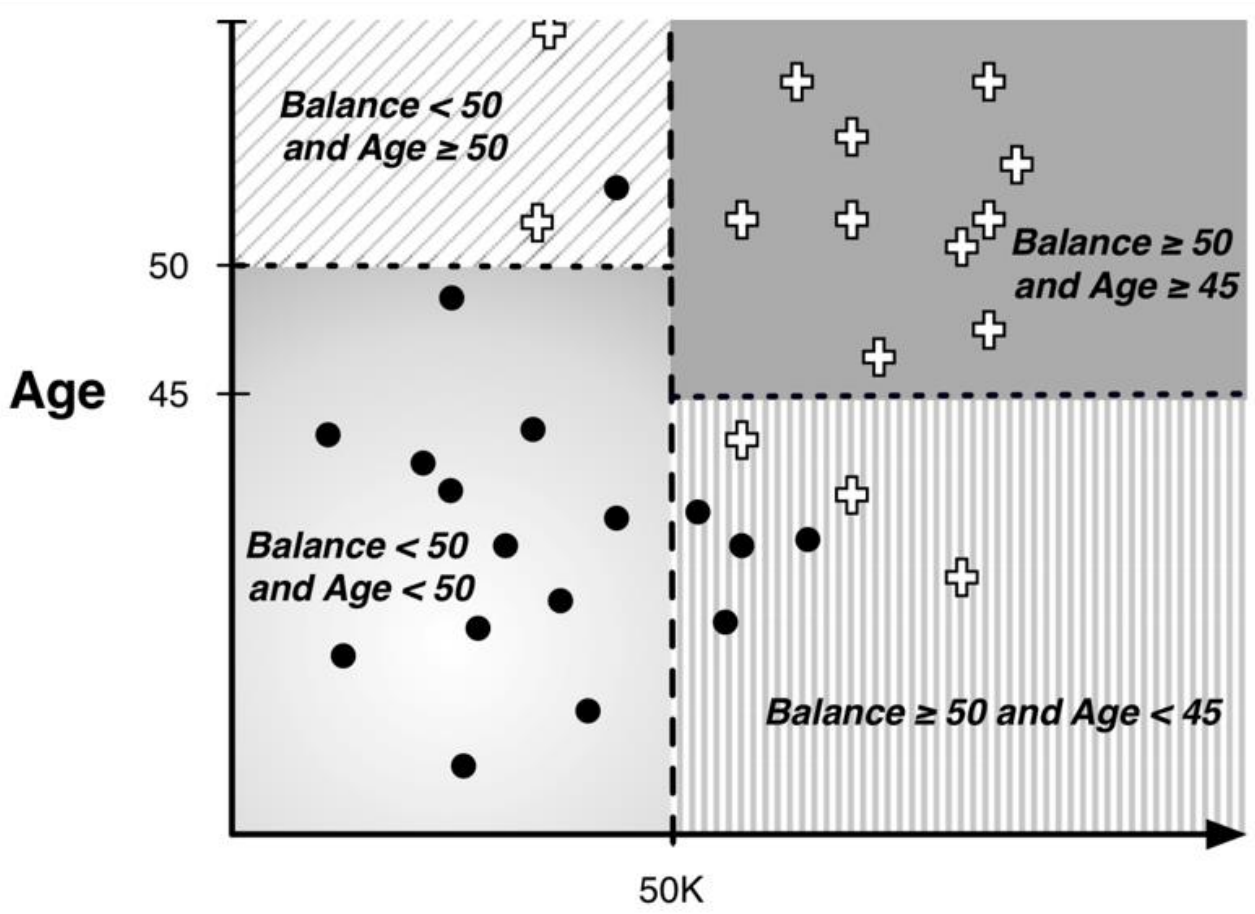
Decision Trees



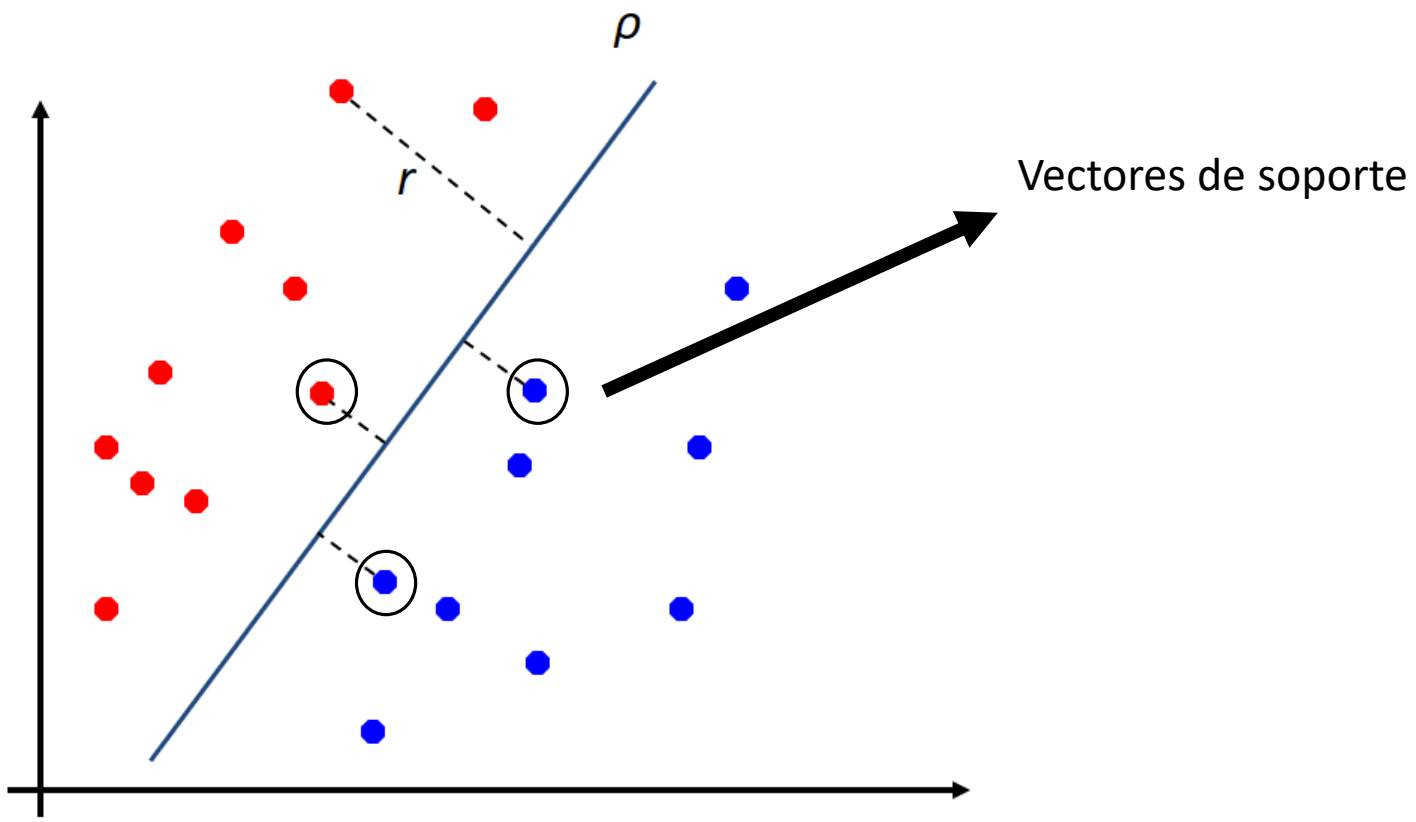
$$entropy = -p_1 \log_2(p_1) - p_2 \log_2(p_2) - ..$$



Decision Trees



Support Vector Machines (SVM)



“Ensembles” o ensambles

- Combina las predicciones de muchos modelos
- Predice usando una predicción “promedio”
- Random Forests es de los más conocidos

Neural Networks (Deep Learning)

Métricas de evaluación

Precision

Recall

F-1

Accuracy

AUC

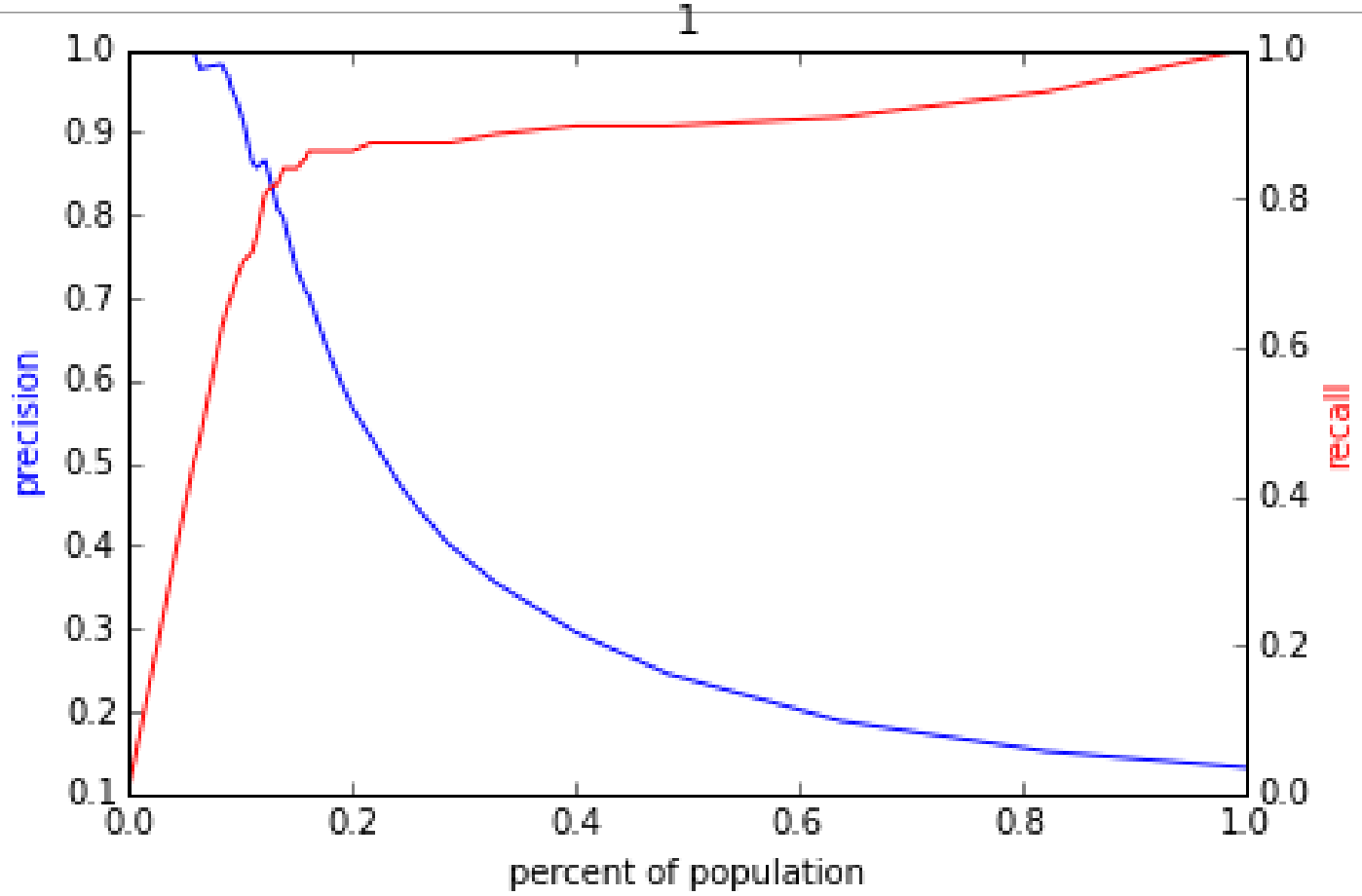
P-R curves

ROC curves

		Predicted Class: Yes	Predicted Class: No
Actual Class	Yes	TP	FN
	No	FP	TN

Measure	Formula
precision	$tp/(tp + fp)$
recall	$tp/(tp + fn)$
f-score	$2pr * re/(pr + re)$
accuracy	$(tp + tn)/(tp + tn + fp + fn)$

Métricas de evaluación



Loop típico de ML

for **train-test in muestras**:

- for **subsets in set_variables:** (demografica (i), geografica (c), temporales (t), comportamiento (it), relacionales (ij), etc)
  - for **classifier in models:**
    - for **parameter in parameters:**
      - **Fit (train)**
      - **Predict (test)**
      - **Store Metrics**

Max_metric, Best_Model = max(

- **[(metric, model) for (metric, model) in results])**

Análisis más profundo de los resultados…

Classifier	Param 1	Param 2	Test Split	Feature Subsets used	Accuracy	Prec @ 1%	AUC

Puesta en marcha de modelo ML

new_X_data = read_data()

prediction = Best_model.predict(new_X_data)

If prediction …:

Cabeats de Machine Learning

Operativos

- Problema con los datos (*garbage in – garbage out*)
- Costo computacional (trade-off precisión versus costo)
- Problemas con el modelamiento (trade-off bias/variance)
- Deterioro del modelo
- Mal mapeo problema / solución
- Consideraciones éticas

Conceptuales