Webinar: Machine Learning (ML) aplicado a la deserción estudiantil

Johann A. Ospina

Universidad Autónoma de Occidente

Facultad de Ciencias Básicas

Departamento de Matemáticas y Estadística

Contenido

1. Entendiendo el ML
2. Aplicaciones del ML
3. Programas para trabajar ML
4. Tipos de algoritmos de ML
5. Caso de estudio

Entendiendo el ML

Entendiendo el ML

- Las organizaciones buscan extraer conocimiento de las enormes cantidades de datos que se almacenan y procesan diariamente.

- El apasionante deseo de predecir el futuro impulsa el trabajo de las empresas analistas y científicos de datos en campos que van desde el mercadeo hasta la atención médica [Nwanganga & Chapple, 2020].

Entendiendo el ML

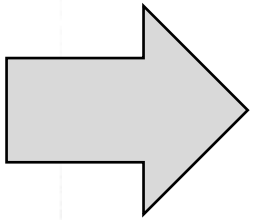
Datos

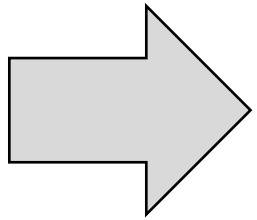
Máquina

?

¿Cómo una máquina puede identificar si un estudiante piensa cancelar?

Entendiendo el ML





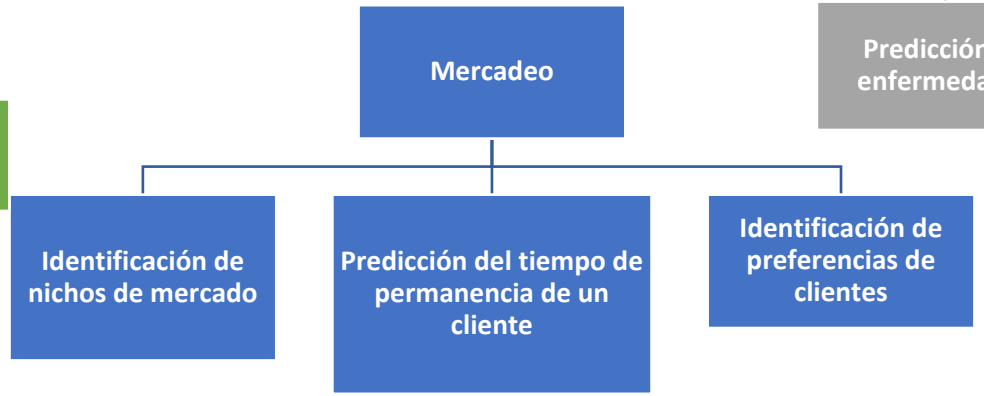
Datos

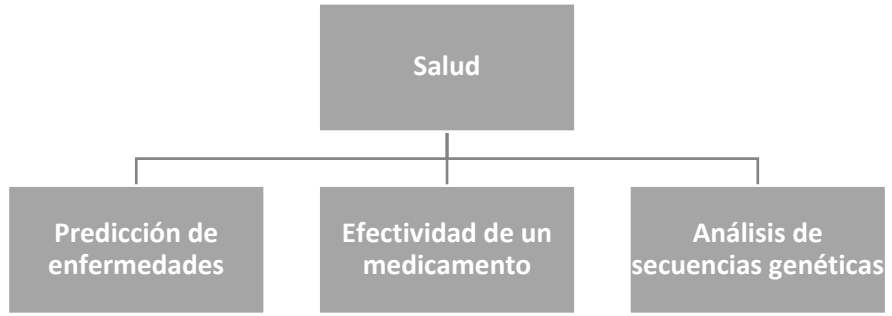
Máquina

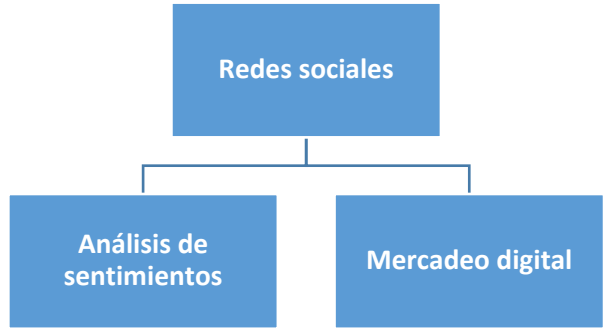
Resultado

Aplicaciones del ML

Aplicaciones del ML



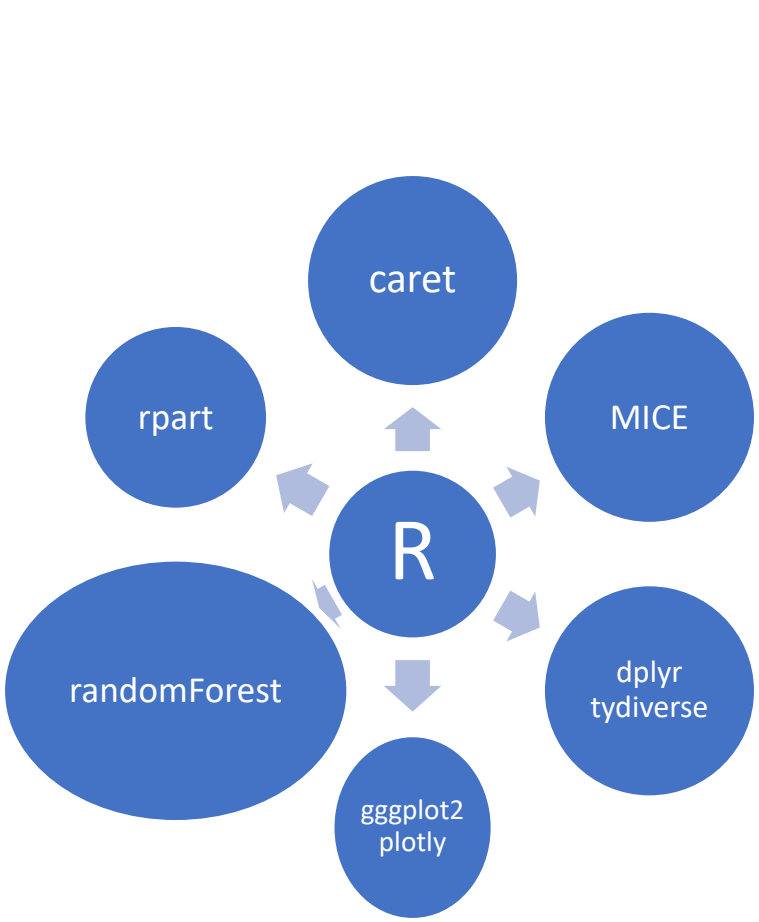


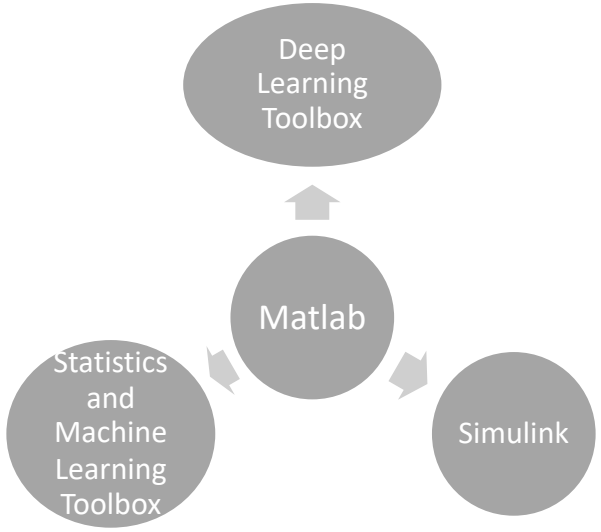


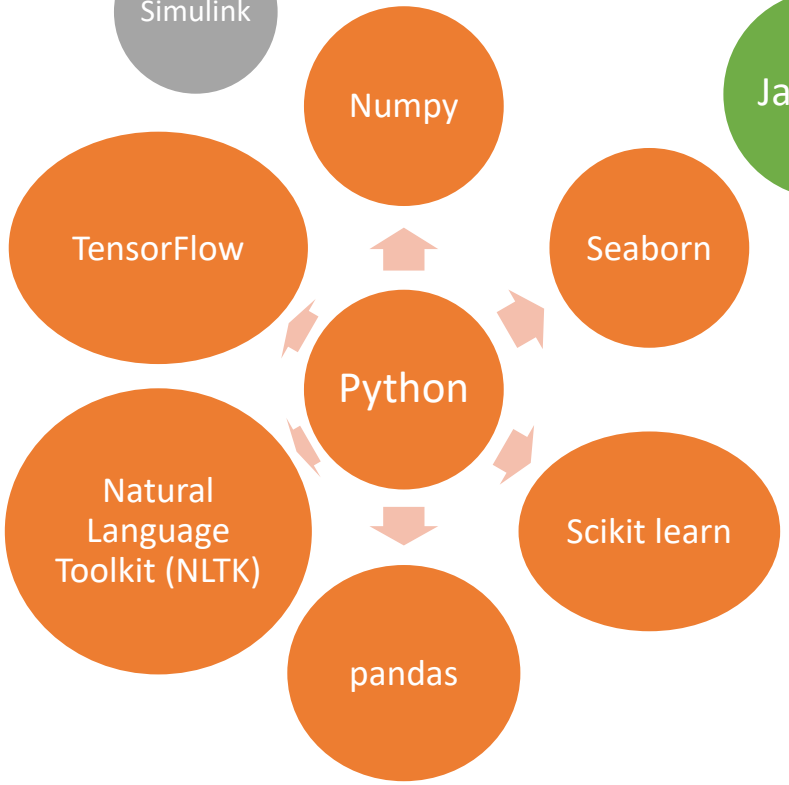
Programas para trabajar ML

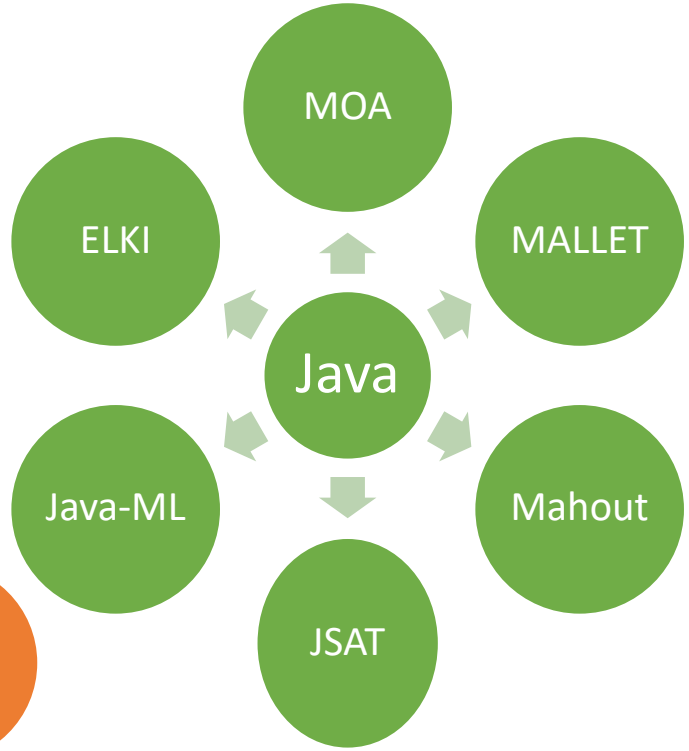
Programas para trabajar ML

Programas para trabajar ML









¿Por qué R?

Código abierto

Orientada a objetos

Aplicativos web

Constante desarrollo

Etapas del Machine Learning

Paso 1

Gestión de datos

Procesamiento de datos

Paso 2

Paso 3

Algoritmo de entrenamiento

Algoritmo de prueba

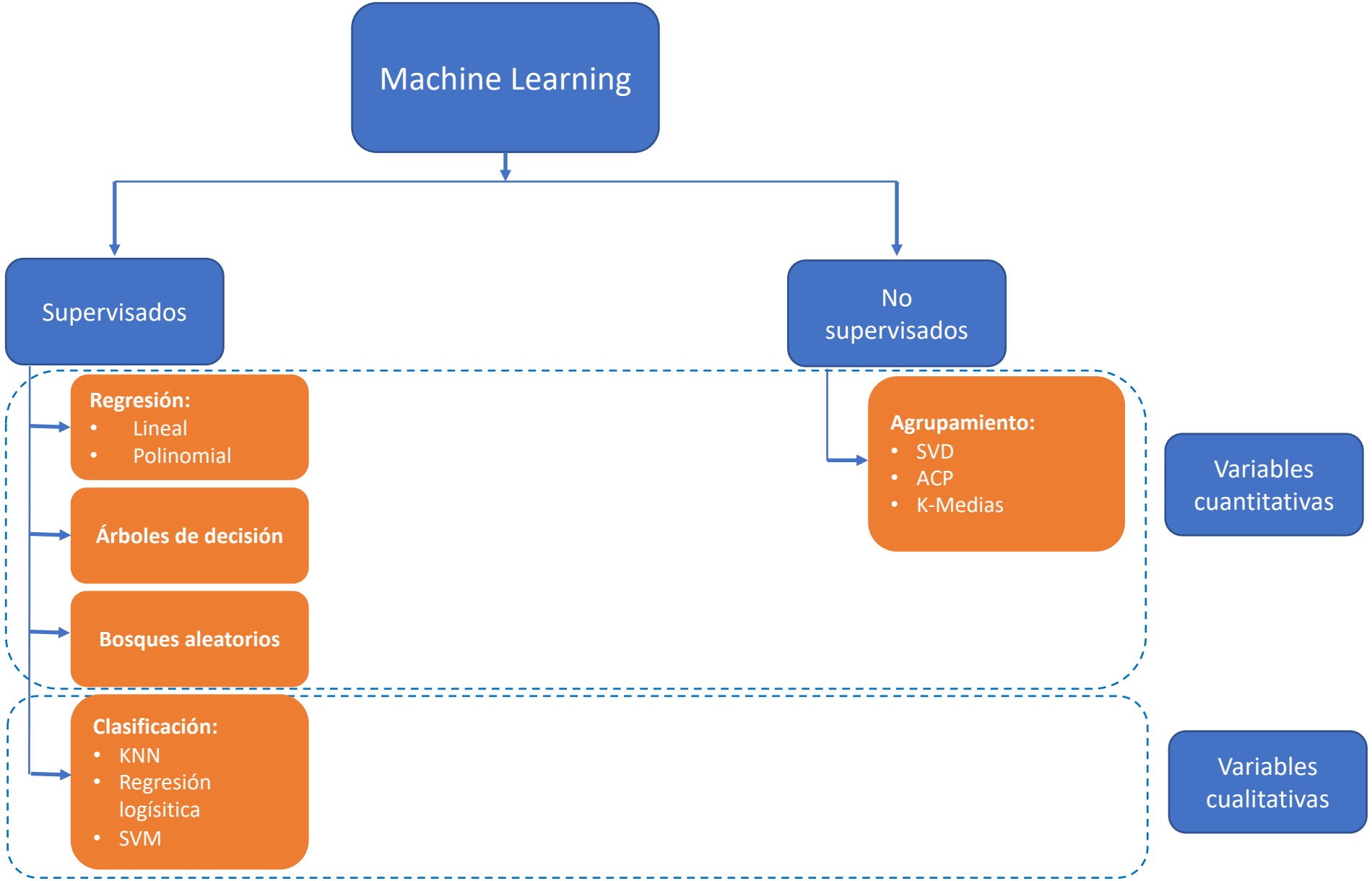
Paso 4

Paso 5

Evaluación de resultados

Tipos de algoritmos de Machine Learning

Tipos de algoritmos de Machine Learning



Aplicación del Machine Learning a la deserción estudiantil

Datos de estudio

Datos de rendimiento de 649 estudiantes de bachillerato. Las variables incluyen calificaciones de los estudiantes, características sociodemográficas y variables relacionadas con el desempeño educativo [Cortez & Silva, 2008].

Diccionario de datos

Attribute	Description (Domain)
sex	student's sex (binary: female or male)
age	student's age (numeric: from 15 to 22)
school	student's school (binary: *Gabriel Pereira* or *Mousinho da Silveira*)
address	student's home address type (binary: urban or rural)
Pstatus	parent's cohabitation status (binary: living together or apart)
Medu	mother's education (numeric: from 0 to 4[a])
Mjob	mother's job (nominal[b])
Fedu	father's education (numeric: from 0 to 4[a])
Fjob	father's job (nominal[b])
guardian	student's guardian (nominal: mother, father or other)
famsize	family size (binary: ≤ 3 or > 3)
famrel	quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
reason	reason to choose this school (nominal: close to home, school reputation, course preference or other)
traveltime	home to school travel time (numeric: 1 – $<$ 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour or 4 – $>$ 1 hour).
studytime	weekly study time (numeric: 1 – $<$ 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – $>$ 10 hours)
failures	number of past class failures (numeric: n if $1 \leq n < 3$, else 4)
schoolsup	extra educational school support (binary: yes or no)
famsup	family educational support (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
paidclass	extra paid classes (binary: yes or no)
internet	Internet access at home (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
freetime	free time after school (numeric: from 1 – very low to 5 – very high)
goout	going out with friends (numeric: from 1 – very low to 5 – very high)
Walc	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
Dalc	workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
health	current health status (numeric: from 1 – very bad to 5 – very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)

a 0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education.
b teacher, health care related, civil services (e.g. administrative or police), at home or other.

Fuente: Cortez & Silva, 2008

En esta parte del webinar se realizará la aplicación de Machine Learning usando R y Rstudio

Observaciones

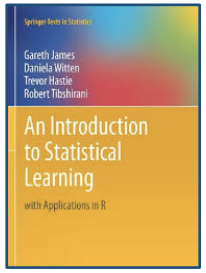
- Antes de aplicar los métodos de ML es importante conocer los datos (tipos de variables, identificación de datos faltantes, datos atípicos, etc).

- Estudiar muy bien la teoría de estadística que hay detrás de las metodologias de ML que se vayan a implementar.

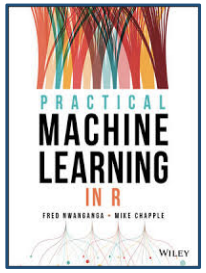
- En el caso de la regresión logística se debe tener cuidado cuando predomina uno de los valores de la variable respuesta, puesto que los enlaces simétricos son inadecuados, por lo tanto, es importante considerar enlaces asimétricos (Chen et al, 1999).

Referencias

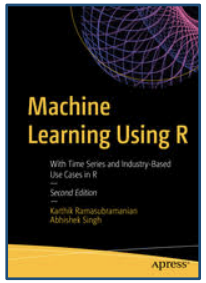
Cortez, P & Silva, A. *Using Data Mining to Predict Secondary School Student Performance*. In A. Brito and J. Teixeira Eds., Proceedings of 5th Future Business Technology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.



James, G., Witten, D., Hastie, T., & Tibshirani. *An introduction to statistical learning* R. New York: springer. 2013.



Nwanganga, F. & Chapple, M. *Practical machine learning in R*. Wiley, 2020.



Ramasubramanian, K.; Singh, A. *Machine learning using R*. New Delhi, India: Apress, 2017.

Referencias

Fellman, D. *Predicting dropout rate in e-learning* (2019). Enlace https://www.rpubs.com/dfellman/elearningdropout

Vilas-Boas, L. *Crafting a Machine Learning Model to Predict Student Retention Using R* (2020). Enlace: https://towardsdatascience.com/crafting-a-machine-learning-model-to-predict-student-retention-using-r-5eb009dcb1ec

CHEN, Ming-Hui; DEY, Dipak K.; SHAO, Qi-Man. *A new skewed link model for dichotomous quantal response data*. Journal of the American Statistical Association, 1999, vol. 94, no 448, p. 1172-1186.

Próximos webinars del departamento de Matemáticas y Estadística

- 12 de febrero (4 a 5 pm). Método de clasificación supervisada y su aplicación en datos de salud.
  Andrés F. Ochoa

- 26 de febrero (4 a 5 pm). Estimación del riesgo de incumplimiento de las empresas de un Banco, con técnicas Machine Learning.
  Diego A. Castro

- 12 de marzo (4 a 5 pm). Creación de dasboard para la generación de reportes dinámicos utilizando la librería shinydashboard de R.
  Johann A. Ospina

GRACIAS