Análisis de datos educativos usando aprendizaje no supervisado, para determinar factores que influyen en la deserción de estudiantes en la facultad de ingeniería de la UdeA.

Eder Samir Correa Acosta

Grupo de Investigación en Telecomunicaciones Aplicadas, GITA

30 de mayo de 2020

Índice

Introducción

Problema

Estado del Arte

- Objetivo:
  - Desarrollar un modelo predictivo usando el aprendizaje no supervisado, que evidencie los factores determinantes en la deserción estudiantil, en los programas presenciales de la sede Medellin de la facultad de ingeniería de la UdeA.
- Útil:
  - Para la creación de un sistema de alertas para prevenir la deserción.
  - Para construir perfiles de los estudiantes.

Base de datos

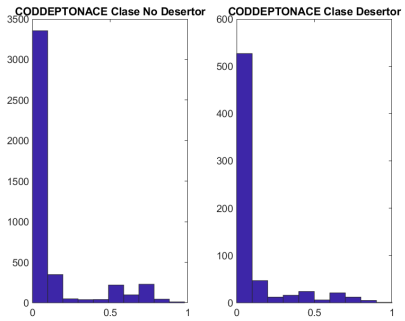
- Base de datos de la Facultad de ingeniería del pregrado (UdeA) 2019-2.

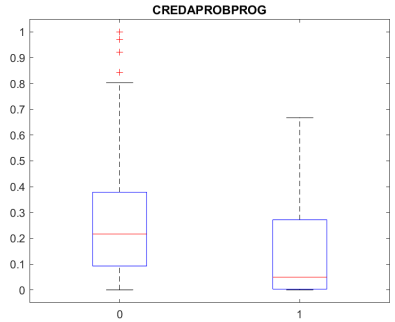
N° Registros Inicial	Enfoque	Descartados	N° Registros final	Distribución de los registros
Registros:8289. Caracterizticas: 39	Sede Medellín y Colombianos	Promedios 9,99 .	Total= 7293	Entrenamiento= 5105
			No desertores= 6347	Prueba = 1458
			desertores = 946	Validación = 729

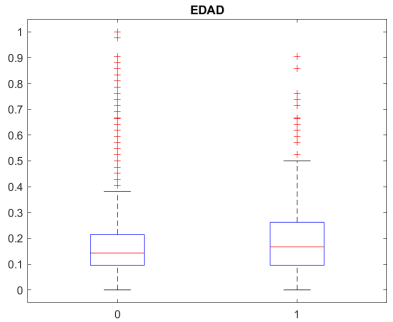
CARACTERÍSTICAS			
CATEGÓRICAS		NUMÉRICAS	
TIPO DOC	NATURALEZA COLE	PROMEDIO SEMESTRE	SEMESTRE INICIA PROGRAMA
SEXO	PROGRAMA	PROMEDIO PROGRAMA	ULTIMO SEMESTRE TERMINADO
TIPO PROGRAMA	DEPARTAMENTO NACE	ESTRATO	CREDITOS VALIDOS UDEA
SEDE	COD DEPTO NACE	PERIODOS PRUEBA PROGRAMA	CREDITOS APROBADOS UDEA
COD PROGRAMA	COD MUNI NACE	RANGO	CRED CURS PROG
FECH NACE	COD DEPTO VIVE	EDAD	CRED APROB PROG
COD PAIS	COD MUNI VIVE	VERSION	TERCIO
PAIS NACE	CREDITOS ULTIM SEMEST MATRIC		NIVEL PREGRADO
ULTIMO SEMESTRE MATRIC	TIPO ACEPTACIÓN		
NOMBRE MUNI NACE	COD PAIS VIVE	ETIQUETAS A PREDECIR (ESTADO)	
PAIS VIVE	NOMBRE MUNI VIVE	NO DESERTOR	0
DEPARTAMENTO VIVE		DESERTOR	1

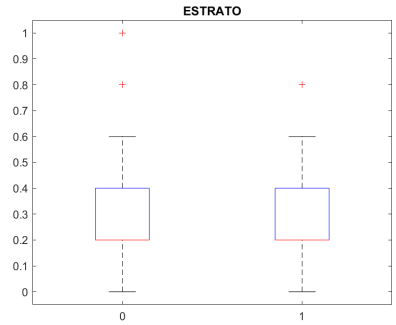
Descartadas inicialmente	
Descartadas: BoxPlot, Histogramas	

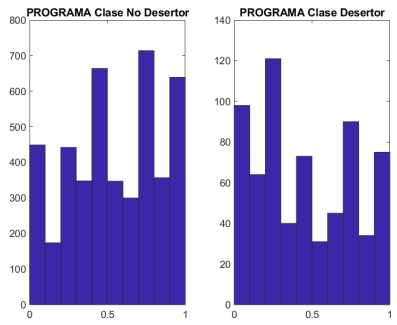
Selección de características

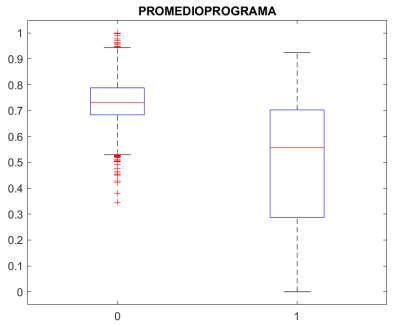


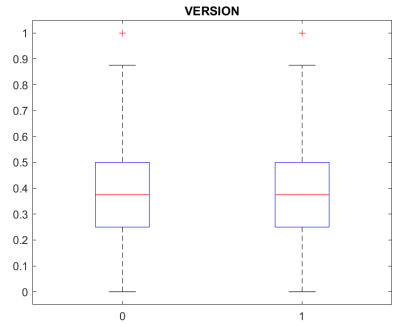


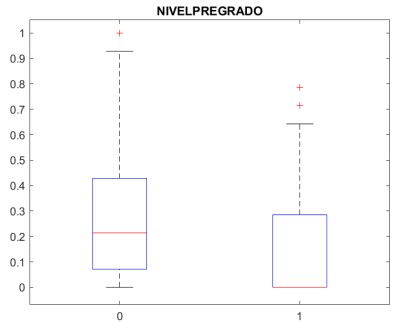


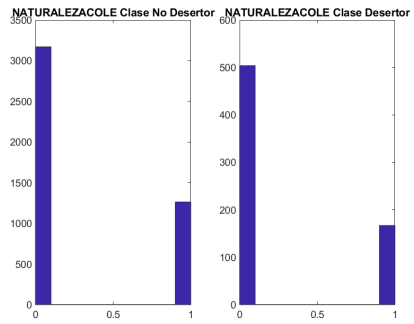






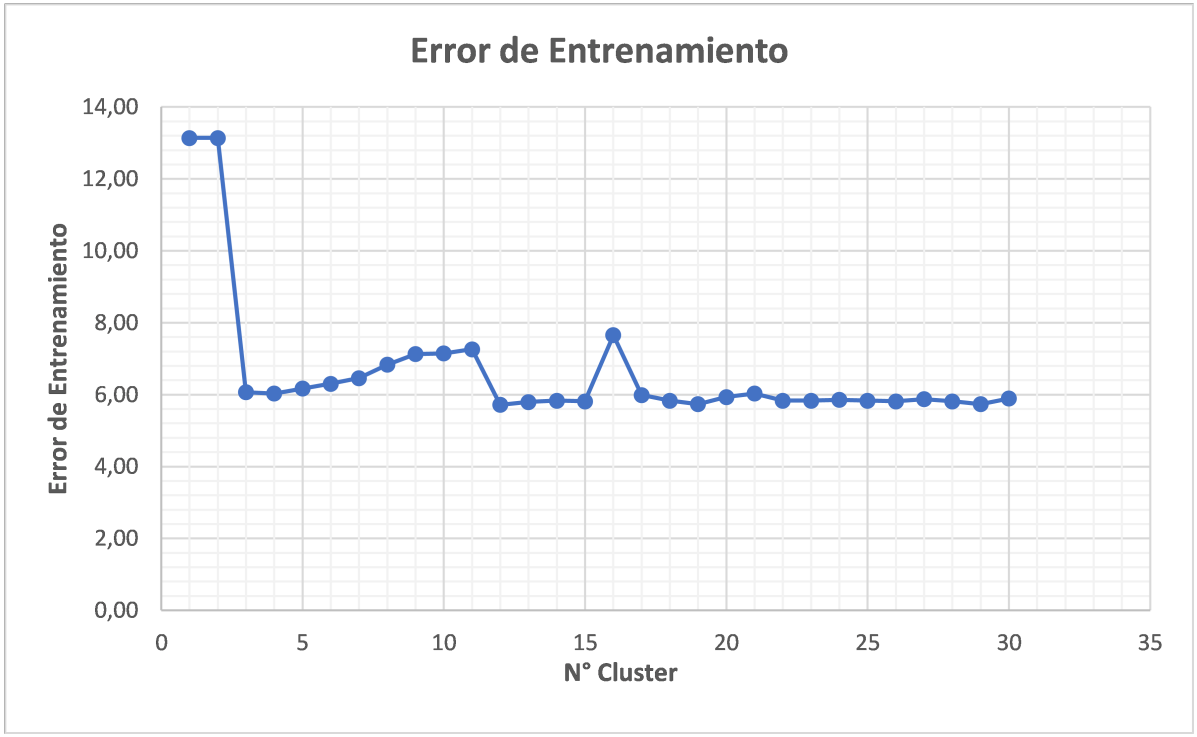






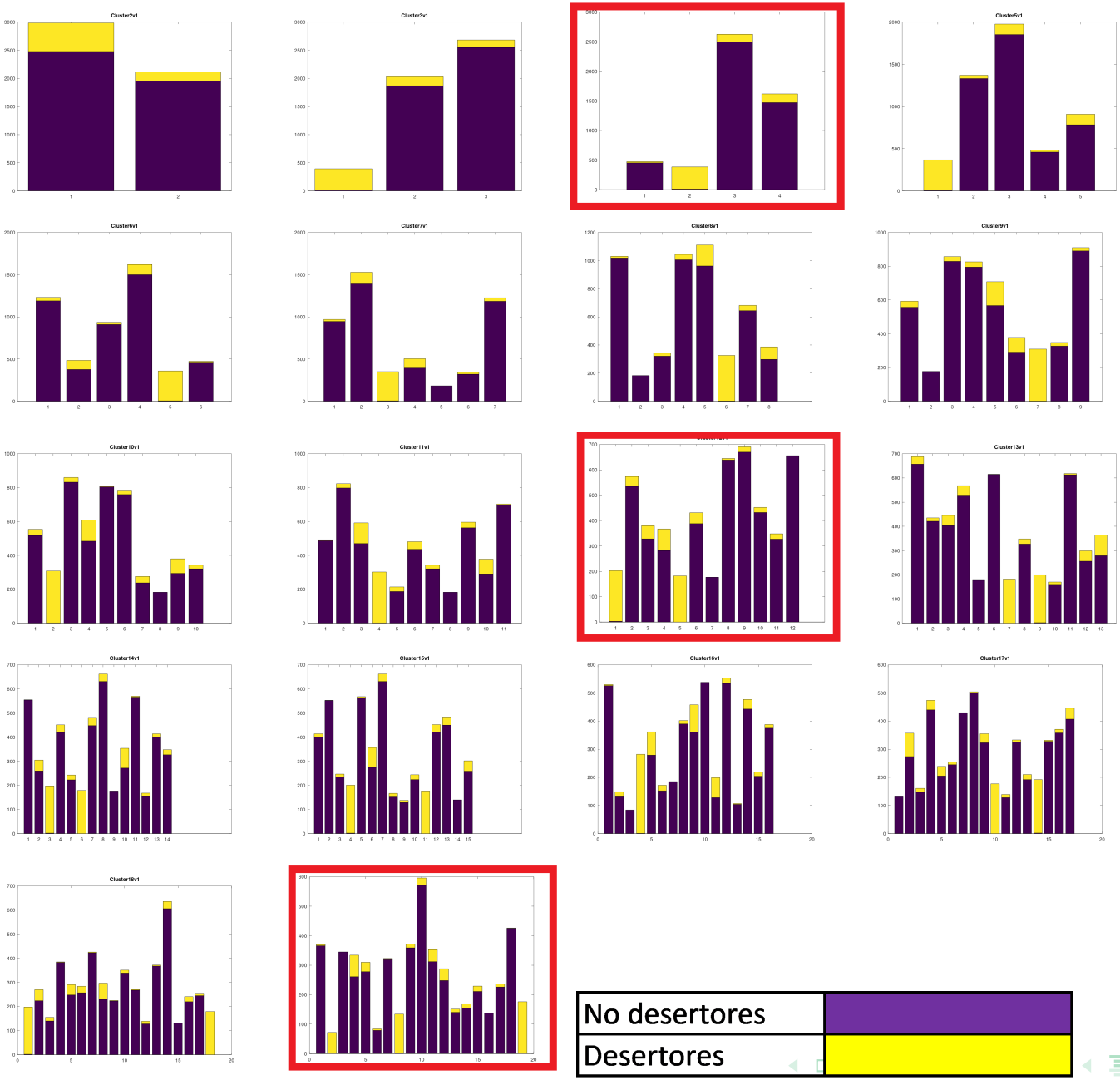
Entrenamiento y selección de clusters

- K-Prototypes (K-Means + K-Modes)[1]
- 5105 instancias (70 % Dataset).
- 13 características (11 numéricas y 2 categóricas)



Cluster	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Error	13,14	13,14	6,07	6,03	6,17	6,31	6,46	6,84	7,13	7,15	7,27	5,72	5,80	5,84	5,82	7,66	5,99	5,84	5,74	5,93

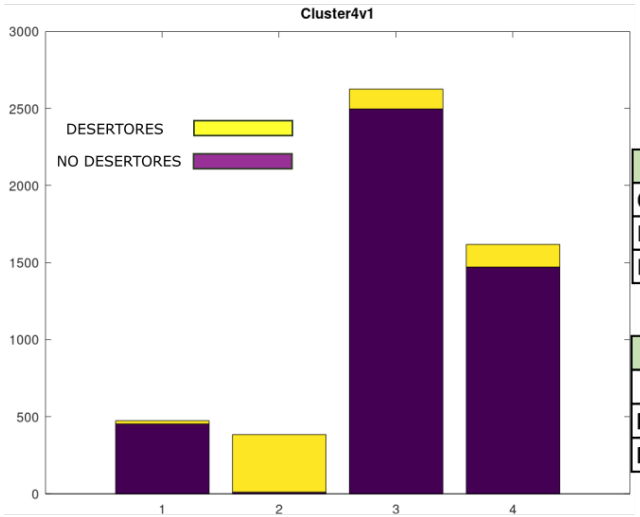
Clusters


Cluster2v1
Cluster3v1
Cluster5v1
Cluster6v1
Cluster7v1
Cluster8v1
Cluster9v1
Cluster10v1
Cluster11v1
Cluster13v1
Cluster14v1
Cluster15v1
Cluster16v1
Cluster17v1
Cluster18v1
No desertores
Desertores

Cluster 4

Cluster	SEXO	PROGRAMA	CREDITOS VALIDOS UDEA	CREDITOS APROBADOS UDEA	CRED CURS PROG	CRED APROBPROG	PROMEDIO SEMESTRE	PROMEDIO PROGRAMA	RANGO	CREDITOS ULTIMSEMES MATRIC	SEMESTRE INICIA PROGRAMA	EDAD	NIVEL PREGRADO
1	1,00	11,00	0,40	0,26	0,45	0,40	0,86	0,82	1,00	0,58	0,89	0,16	0,43
2	1,00	1,00	0,03	0,03	0,07	0,03	0,18	0,31	0,21	0,33	0,97	0,20	0,02
3	1,00	11,00	0,12	0,09	0,15	0,12	0,67	0,71	0,43	0,48	0,95	0,13	0,09
4	1,00	13,00	0,42	0,28	0,52	0,42	0,74	0,75	0,40	0,49	0,84	0,22	0,47

Figura: Centroides para 4 clusters



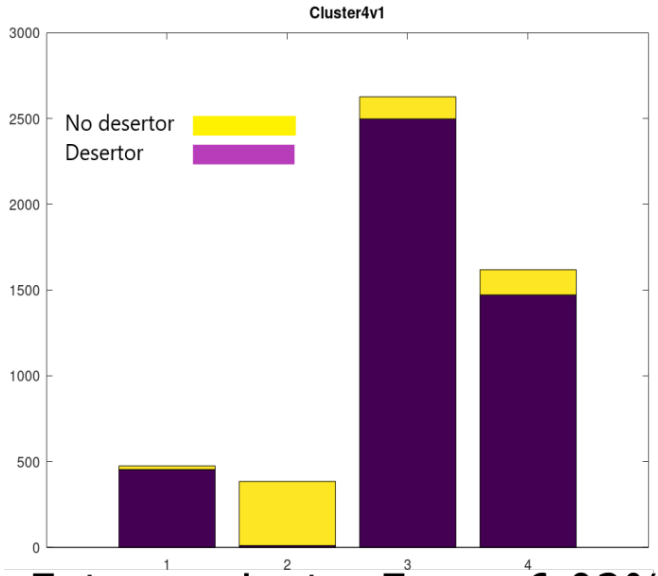
SEXO - Distribución Por cada Cluster				
Cluster	1	2	3	4
Femenino %	43,8	22,3	33,7	30,3
Masculino %	56,2	77,7	66,3	69,7

Distribución genero entre los clusters				
	1	2	3	4
Femenino %	12,5	5,1	53,1	29,3
Masculino %	7,8	8,7	50,7	32,9

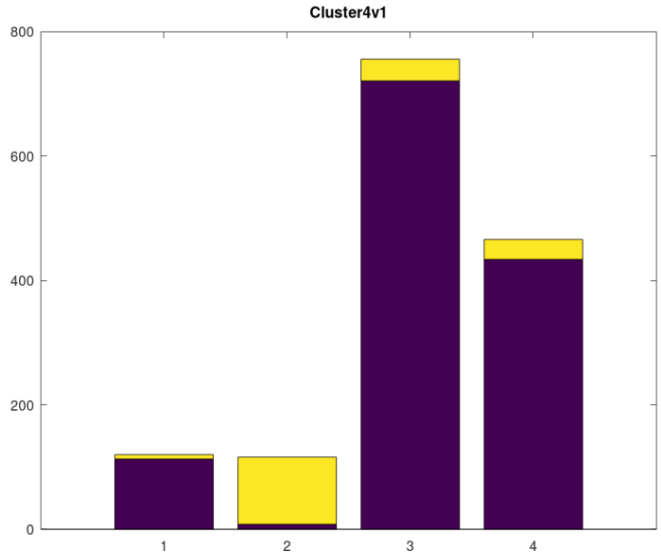
Programa				
Cluster	1	2	3	4
BIOINGENIERÍA %	6	4	6	9
ING AMBIENTAL 1%	1	15	3	2
ING DE SISTEMAS %	1	14	4	4
ING INDUSTRIAL %	3	12	6	2
ING TELECOMUNICACIONES %	2	12	7	5
INGENIERÍA AMBIENTAL 2%	14	3	8	7
INGENIERÍA CIVIL %	13	3	8	10
INGENIERA DE MATERIALES %	5	4	5	7
INGENIERÍA DE SISTEMAS %	5	4	7	10
INGENIERÍA ELÉCTRICA %	5	6	7	6
INGENIERÍA ELECTRÓNICA %	8	7	6	9
INGENIERÍA INDUSTRIAL %	19	3	9	6
INGENIERÍA MECÁNICA %	4	5	8	9
INGENIERÍA QUÍMICA %	6	3	7	11
INGENIERÍA SANITARIA %	9	5	7	4

Datos de prueba (4 clusters)

- 1458 instancias (20 % Dataset)



Entrenamiento. Error=6.03%



Prueba. Error=5.62%

Bibliografía I

[1] Z. Huang, "Clustering large data sets with mixed numeric and categorical values," proceedings of 1st pacific-asia conference on knowledge discouvery and data mining," 1997.