

Maestría oficial en Big Data y Data Science

Actividad Guiada 3

Metodologías de gestión y diseño de proyectos Big Data

Alumno: Castillo Bastidas, José Ricardo

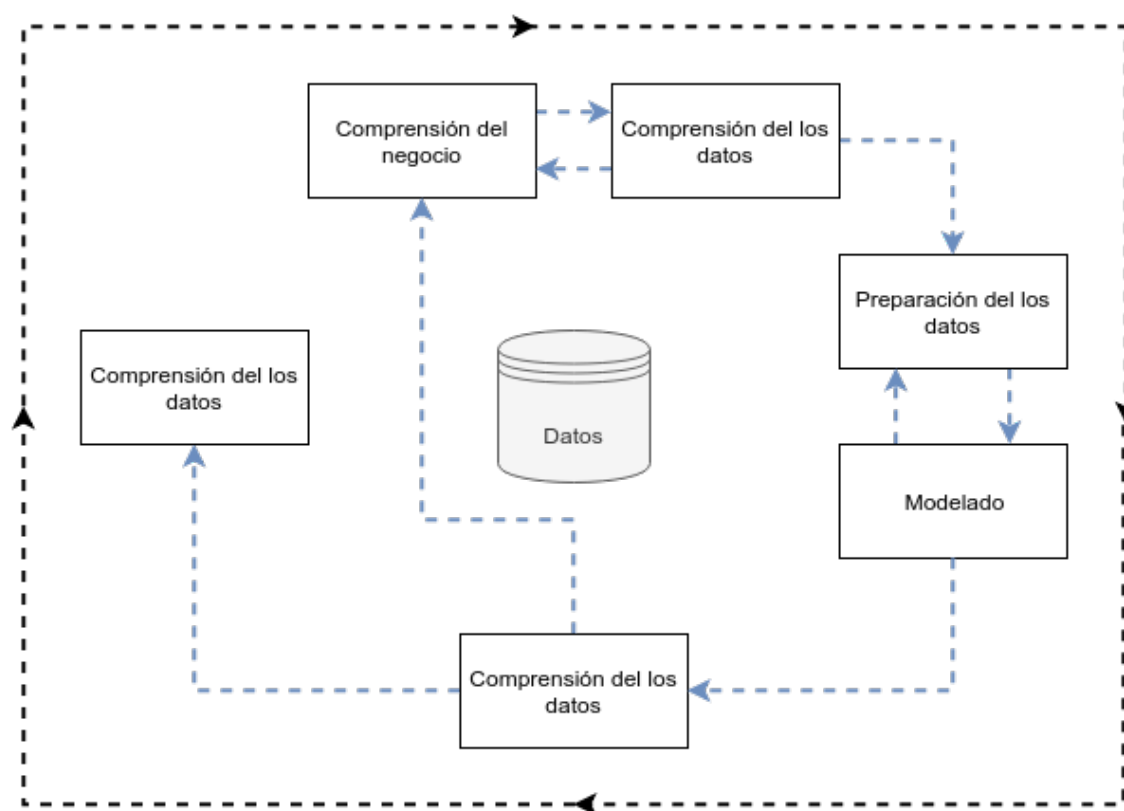
Fecha de entrega: 29/11/2021

Seminario II - Aplicando técnicas ágiles para la gestión de proyectos de ciencia de datos

El presente documento es una planilla que se utilizará para el desarrollo de la documentación correspondiente a las actividades del Seminario II y su correspondiente Actividad Guiada. El contenido será guiado según las fases y actividades de la metodología CRISP-DM.

Una vez completado con la información correspondiente al proyecto de ciencia de datos y complementado con los reportes de la ejecución de la libreta Jupyter desarrollada se podrán finalizar las tareas del proyecto.

La metodología CRISP-DM cuenta con 6 fases, ver figura 1, que forman un ciclo iterativo, con vistas a lo que se podrá considerar como un proceso iterativo-incremental de desarrollo de soluciones de ciencia de datos para un contexto en particular.



[A] Fase de comprensión del problema

- **Determinar los objetivos de la Organización**

Las autoridades de una Institución Universitaria desean obtener conocimiento a partir de los datos disponibles de los alumnos, principalmente en lo que respecta a su situación como estudiante **Activo** (continúa cursando la carrera) / **Pasivo** (ha abandonado los estudios o al menos no se ha reinscrito para continuar cursando en la actualidad). El objetivo final que se persigue es el de poder predecir con un margen de confianza considerable la situación de los nuevos alumnos inscritos para el periodo 2021.

- **Evaluación de la situación**

Se cuenta con los recursos para la ejecución del proyecto de ciencia de datos en cuestión.
Entre ellos:

- Los datos históricos y los datos de los estudiantes del ciclo lectivo 2021 que se integran a la Unidad Académica (Facultad – Escuela)
- Se cuenta con el personal para la ejecución de las tareas involucradas en el proyecto.
- Se cuenta con las herramientas software y plataformas de despliegue adecuadas para la presentación del producto a desarrollar.

- **Determinación de los objetivos del proyecto.**

Elaborar un **modelo de predicción** de la situación académica de los estudiantes de la Unidad Académica (Escuela – Facultad) en cuestión sobre la base de los datos históricos disponibles. La **efectividad** del modelo deberá ser del **80%**. El mismo modelo será aplicado para realizar la predicción del campo “**situación del estudiante**” con los datos de los que han ingresado en el presente periodo lectivo (**2021**).

- Definir plan del proyecto (tareas, recursos, etc)

El equipo va a estar conformado por:

- José Ricardo Castillo Bastidas

Duración de las iteraciones: una (1) semana

Velocidad del equipo: 40 SP

Hoja de ruta del proyecto (herramienta Jira):

Epic	OCT – DIC	
▼ MBID-1 Comprensión del negocio MBID-7 Determinar los objetivos FINALIZADA MBID-8 Evaluación de la situación FINALIZADA MBID-9 Determinación de los obj... FINALIZADA MBID-10 Definir el plan de proyecto EN CURSO	Sprin... 	
▼ MBID-2 Comprensión de los datos MBID-11 Recolección de los dato... TAREAS PO... MBID-12 Descripción de los datos TAREAS PO... MBID-13 Exploración de los datos TAREAS PO... MBID-14 Verificación de la calida... TAREAS PO...		
▼ MBID-3 Preparación de los datos MBID-15 Selección de los datos TAREAS PO... MBID-16 Limpieza de los datos TAREAS PO... MBID-17 Construcción de datos TAREAS PO... MBID-18 Integración de los datos TAREAS PO... MBID-19 Formateo de los datos TAREAS PO...		
MBID-4 Modelado		
MBID-5 Evaluación		
MBID-6 Despliegue		

Se van a desarrollar dos iteraciones a lo largo de las cuales se van a cubrir las 6 fases de la metodología CRISP-DM dando como resultado el desarrollo del MVP correspondiente al proyecto en ejecución. Sprint backlog:

Proyectos / MBID13

Backlog

¿Necesita más de Jira tu equipo? Consigue una versión de prueba gratuita de nuestro plan Standard.

trabajo ▾ Proyectos ▾ Filtros ▾ Paneles ▾ Personas ▾ Aplicaciones ▾ **Crear** ? ⚙️ JC

Proyectos / MBID13

Backlog

JC Epic ▾ Insights

▼ **Sprint 1** 6 nov. – 20 nov. (13 incidencias) 27 3 6 Completar sprint ...

Desarrollar las actividades correspondientes a la fases de: -Comprensión del negocio -Comprensión de los datos -Preparación de los datos Para el proyecto en curso

MBID-7	Determinar los objetivos de la organización	COMPENSIÓN DEL NEGOCIO	1	FINALIZADA
MBID-8	Evaluación de la situación	COMPENSIÓN DEL NEGOCIO	2	FINALIZADA
MBID-9	Determinación de los objetivos del proyecto	COMPENSIÓN DEL NEGOCIO	3	FINALIZADA
MBID-10	Definir el plan de proyecto	COMPENSIÓN DEL NEGOCIO	3	EN CURSO
MBID-11	Recolección de los datos iniciales	COMPENSIÓN DE LOS DATOS	2	TAREAS POR HACER
MBID-12	Descripción de los datos	COMPENSIÓN DE LOS DATOS	2	TAREAS POR HACER
MBID-13	Exploración de los datos	COMPENSIÓN DE LOS DATOS	3	TAREAS POR HACER
MBID-14	Verificación de la calidad de los datos	COMPENSIÓN DE LOS DATOS	5	TAREAS POR HACER
MBID-15	Selección de los datos	PREPARACIÓN DE LOS DATOS	2	TAREAS POR HACER
MBID-16	Limpieza de los datos	PREPARACIÓN DE LOS DATOS	3	TAREAS POR HACER
MBID-16	Limpieza de los datos	PREPARACIÓN DE LOS DATOS	3	TAREAS POR HACER
MBID-17	Construcción de datos	PREPARACIÓN DE LOS DATOS	5	TAREAS POR HACER
MBID-18	Integración de los datos	PREPARACIÓN DE LOS DATOS	2	TAREAS POR HACER
MBID-19	Formateo de los datos	PREPARACIÓN DE LOS DATOS	3	TAREAS POR HACER

+ Crear incidencia

▼ Backlog (0 incidencias) ...

[B] Fase de comprensión de los datos

- Recolección de datos iniciales

Los datos se agrupan en 3 dimensiones:

- **Datos Académicos:** referidos a la actividad académica del alumno y su situación (Activo, Pasivo).
- **Datos Censales:** referidos a su precedencia y algunos datos relativos al nivel de estudio alcanzado por sus padres.
- **Datos Personales:** referidos a sus estudios en el nivel medio, localidad desde la que proviene, entre otros.

- Descripción de los datos

Se detalla el contenido de cada uno de los datasets:

Dataset	Columnas	Observaciones
datos_academicos.csv	unidad_academica nro_inscripcion carrera regular cnt_readmisiones calidad fecha_ingreso_alumno anio_plan_estudios	Número de filas 2316 Filas con valores nulos: 45/2316 (1.94%)
datos_censales.csv	unidad_academica nro_inscripcion estado_civil sit_lab_alumno tipo_res_nuevo sit_lab_padres estudios_padres	Número de filas 2316 Filas con valores nulos: 0/2316 (0%)
datos_personas.csv	unidad_academica nro_inscripcion sexo nacionalidad fecha_nac_alumno fecha_egr_sec	Número de filas 2316 Filas con valores nulos: 0/2316 (0%)

- Exploración de datos

Datasets:

[a] *datos_academicos.csv*

[b] *datos_censales.csv*

[c] *datos_personas.csv*

Dataset	Atributo	Tipo de datos	Metadatos
[a]	nro_inscripcion	nominal	-- Valores presentes (10 primeros): ['FCEQN-877' 'FCEQN-1294' 'FCEQN-1351' 'FCEQN-1363' 'FCEQN-1367' '1527' 'FCEQN-1532' 'FCEQN-1618' 'FCEQN-1647' 'FCEQN-1665'] -- Cantidad de nulos: 0 = 0.00%
	carrera	nominal	-- Valores presentes (10 primeros): ['143' '102' '114' '106' '170' '104' '147' '172' '601' '108'] -- Cantidad de nulos: 2 = 0.09%
	regular	nominal	-- Valores presentes (10 primeros): ['S' 'N'] -- Cantidad de nulos: 0 = 0.00%
	cnt_readmisiones	numérico	min 0.000000 max 2.000000 mean 0.984456 std 0.718602 median 1.000000
	calidad	nominal	-- Valores presentes (10 primeros): ['A' 'P'] -- Cantidad de nulos: 0 = 0.00%
	anio_plan_estudios	numérico	min 2000.000000 max 2014.000000 mean 2007.449438

			std 3.893340 median 2008.000000
	fecha_ingreso_alumno	nominal	-- Valores presentes (10 primeros): ['04/27/2006' '06/13/2006' '03/01/2001' '05/02/2006' '05/15/2006' '12/18/2006' '12/21/2006' '12/26/2006' '12/04/2006' '12/06/2006'] -- Cantidad de nulos: 0 = 0.00%
	unidad_academica	nominal	-- Valores presentes (10 primeros): ['FCEQN'] -- Cantidad de nulos: 0 = 0.00%

Dataset	Atributo	Tipo de datos	Metadatos
[b]	ua	nominal	-- Valores presentes (10 primeros): ['FCEQN'] -- Cantidad de nulos: 0 = 0.00%
	insc	nominal	-- Valores presentes (10 primeros): ['FCEQN-3342' 'FCEQN-5396' 'FCEQN-3162' 'FCEQN-5739' '4683' 'FCEQN-5233' 'FCEQN-5226' 'FCEQN-5117' 'FCEQN-5630' 'FCEQN-5644'] -- Cantidad de nulos: 0 = 0.00%
	estado_civil	numérico	min 1.000000 max 6.000000 mean 1.041451 std 0.315216 median 1.000000
	sit_lab_alumno	nominal	-- Valores presentes (10 primeros): ['NC' 'No' 'Si' '1' 'S' 'NO']

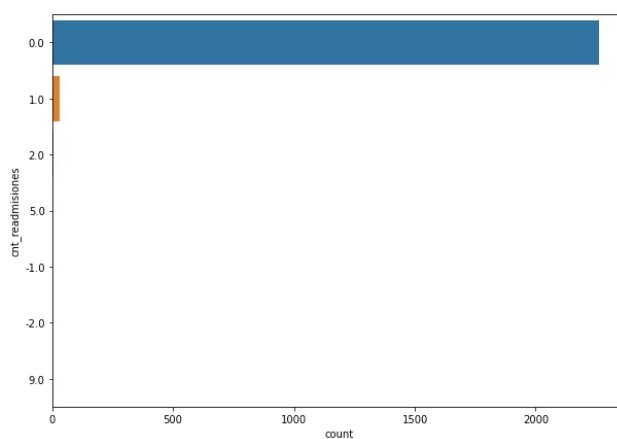
			-- Cantidad de nulos: 0 = 0.00%
	tipo_res_alumno	numérico	min -1.000000 max 4.000000 mean 0.050950 std 1.485711 median -1.000000
	sit_lab_padres	Numérico	min -1.000000 max 5.000000 mean 0.781088 std 2.085848 median -1.000000
	estudios_padres	Numérico	min -1.000000 max 7.000000 mean 1.986183 std 2.739800 median 1.000000

Dataset	Atributo	Tipo de datos	Metadatos
[c]	unidad_academica	nominal	Descripción de valores: -- Valores presentes (10 primeros): ['FCEQN'] -- Cantidad de nulos: 0 = 0.00%
	nro_inscripcion	nominal	-- Valores presentes (10 primeros): ['FCEQN-877' 'FCEQN-1294' 'FCEQN-1351' 'FCEQN-1363' 'FCEQN-1367' '1527' 'FCEQN-1532' 'FCEQN-1618' 'FCEQN-1647'

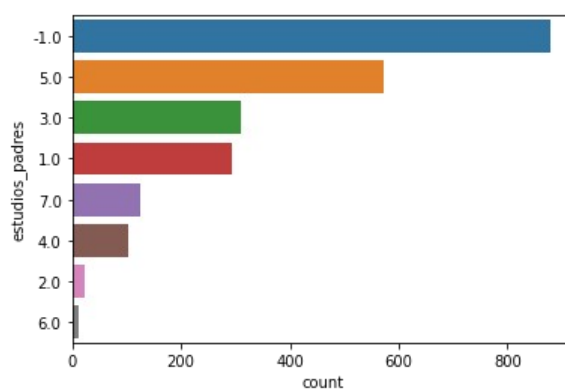
			'FCEQN-1665'] -- Cantidad de nulos: 0 = 0.00%
	Sexo	Numérico	min 1.000000 max 2.000000 mean 1.462435 std 0.498695 median 1.000000
	nacionalidad	Numérico	min -1.000000 max 4.000000 mean 1.032383 std 0.210493 median 1.000000
	fecha_nac_alumno	Numérico	min 181.000000 max 1996.000000 mean 1983.797496 std 99.685020 median 1993.000000
	fecha_egr_sec	Numérico	min 201.000000 max 2019.000000 mean 2009.807858 std 37.678975 median 2011.000000

Gráficos de distribución de valores por atributo:

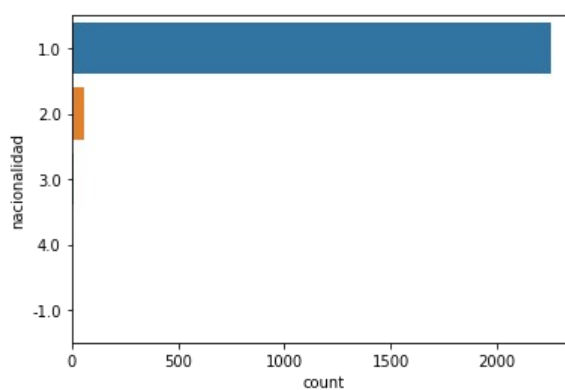
- Conteo de readmisiones - datos_academicos



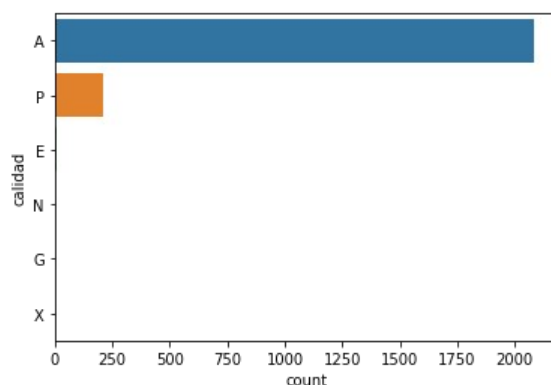
- Conteo de estudios en los padres – datos_censales



- Diferentes nacionalidades – datos_personas



- Conteo del atributo calidad – datos_academicos



- Verificación de la calidad de los datos

Dataset	Atributo	Descripción / Observaciones
[a, b, c]	nro_inscripcion	Se han detectado errores en el formateo del valor del atributo. La distribución por dataset es: * [a]: 71 filas / 2316 total 0.03% * [b]: 71 filas / 2316 total 0.03% * [c]: 71 filas / 2316 total 0.03%
[a]	carrera	Cantidad de filas con valores fuera de rango (incluye valores nulos): 8 filas (0.35 %)
[b]	sit_lab_alumno	Cantidad de filas con valores fuera de rango (incluye valores nulos): 121 filas (5.22%)
[c]	fecha_nac_alumno	Cantidad de filas con valores fuera de rango (incluye valores nulos): 23 filas (0.99 %)
[a]	calidad	Cantidad de filas con valores fuera de rango en atributo calidad: 24 Porcentaje de filas con errores de rango de valores (atributo calidad): 1.04 %

[c]	fecha_egr_sec	Cantidad de filas con valores fuera de rango (incluye valores nulos): 0 filas (0 %)
[a]	regular	Cantidad de filas con valores fuera de rango en atributo regular: 20 Porcentaje de filas con errores de rango de valores (atributo regular): 0.86 %

- Captura de la hoja de ruta con todo lo gestionado en la actualidad

Epic	OCT – DIC	
	Spri...	
▼ MBID-1 Comprensión del negocio		
MBID-7 Determinar los objetivos FINALIZADA		
MBID-8 Evaluación de la situación FINALIZADA		
MBID-9 Determinación de los obj... FINALIZADA		
MBID-10 Definir el plan de proyecto FINALIZADA		
▼ MBID-2 Comprensión de los datos		
MBID-11 Recolección de los datos... FINALIZADA		
MBID-12 Descripción de los datos FINALIZADA		
MBID-13 Exploración de los datos FINALIZADA		
MBID-14 Verificación de la calidad... FINALIZADA		
▼ MBID-3 Preparación de los datos		
MBID-15 Selección de los datos TAREAS PO...		
MBID-16 Limpieza de los datos TAREAS PO...		
MBID-17 Construcción de datos TAREAS PO...		
MBID-18 Integración de los datos TAREAS PO...		
MBID-19 Formateo de los datos TAREAS PO...		
MBID-4 Modelado		
MBID-5 Evaluación		
MBID-6 Despliegue		

- *Captura del backlog gestionado en la actualidad*

Visualización Historial Marcadores Ventana Ayuda

jrcastb.atlassian.net

jrcastb/visualizacion-ag1: Created with StackBlitz

MBID13: tablero ágil - Jira

abajo Proyectos Filtros Paneles Personas Aplicaciones Crear

Proyectos / MBID13

Backlog

Buscar

Insights

Sprint 1 6 nov. - 20 nov. (13 incidencias) 15 0 21 Completar sprint

Desarrollar las actividades correspondientes a la fases de: -Comprensión del negocio -Comprensión de los datos -Preparación de los datos Para el proyecto en curso

ID	Título	Fase	Asignado a	Estado
MBID-7	Determinar los objetivos de la organización	COMPRESIÓN DEL NEGOCIO	1	FINALIZADA
MBID-8	Evaluación de la situación	COMPRESIÓN DEL NEGOCIO	2	FINALIZADA
MBID-9	Determinación de los objetivos del proyecto	COMPRESIÓN DEL NEGOCIO	3	FINALIZADA
MBID-10	Definir el plan de proyecto	COMPRESIÓN DEL NEGOCIO	3	FINALIZADA
MBID-11	Recolección de los datos iniciales	COMPRESIÓN DE LOS DATOS	2	FINALIZADA
MBID-12	Descripción de los datos	COMPRESIÓN DE LOS DATOS	2	FINALIZADA
MBID-13	Exploración de los datos	COMPRESIÓN DE LOS DATOS	3	FINALIZADA
MBID-14	Verificación de la calidad de los datos	COMPRESIÓN DE LOS DATOS	5	FINALIZADA
MBID-15	Selección de los datos	PREPARACIÓN DE LOS DATOS	2	TAREAS POR HACER
MBID-16	Limpieza de los datos	PREPARACIÓN DE LOS DATOS	3	TAREAS POR HACER
MBID-17	Construcción de datos	PREPARACIÓN DE LOS DATOS	5	TAREAS POR HACER
MBID-18	Integración de los datos	PREPARACIÓN DE LOS DATOS	2	TAREAS POR HACER
MBID-19	Formateo de los datos	PREPARACIÓN DE LOS DATOS	3	TAREAS POR HACER

+ Crear incidencia

##INICIO AG3

[C] Fase de preparación de los datos

- Selección de datos

Los siguientes atributos se omitieron al inicio de la fase de preparación de los datos

- 'unidad_academica' >> Aplica para los tres datasets
- 'carrera' y 'anio_plan_estudios' >> Aplica solo al dataset de datos_academicos

Se deja constancia de que se considera adecuado eliminar la columna de 'nro_inscripcion' una vez que se hayan unificado los tres conjuntos de datos originales.

- Limpieza de los datos

Sobre los datasets disponibles se realizaron las siguientes acciones:

Dataset	Atributo	Descripción de lo realizado
[b]	sit_lab_alumno	<p>Cambios realizados:</p> <p>'No' >> 'N',</p> <p>'Si' >> 'S',</p> <p>'1' >> 'S',</p> <p>'NO' >> 'N'</p> <p>Los valores presentes inicialmente en el dataset (izq) han sido reemplazados por los que están definidos en el diccionario de datos (der). Esta operación ha sido realizada con el aval de los expertos en el dominio.</p>
[c]	fecha_nac_alumno	<p>Se aplicó un filtro para los valores del atributo. En conformidad con lo establecido por los expertos en el dominio el rango de valores posibles se fijó en: [1950 - 2006].</p> <p>Las filas que estuvieron fuera de rango se han filtrado.</p>
[c]	fecha_egr_sec	Se aplicó un filtro identificando y contando los valores que no cumplen las condiciones de fecha de

		<p>egreso [1965 – 2015] esto debido a la antigüedad de datos que estamos usando</p> <p>Las filas que estuvieron fuera de rango se han filtrado.</p>
[a]	regular	<p>Cambios realizados</p> <p>'D' » 'N', 'SI' » 'S', 'T' » 'S', 'X' » 'N'</p> <p>Los valores presentes inicialmente en el dataset (izq) han sido reemplazados por los que están definidos en el diccionario de datos (der). Esta operación ha sido realizada con el aval de los expertos en el dominio.</p>

- Construcción de datos

Las autoridades requieren que se generen los siguientes atributos a fin de caracterizar a los estudiantes en análisis:

- **Edad** (calculada en años al día de hoy)
- **Inactividad** (calculada en años entre el egreso del nivel medio y la inscripción a la carrera cuyos datos se están analizando)

- Integración de los datos

A fin de contar con un único dataset que contenga todos los datos del escenario de trabajo se integraron las fuentes inicialmente disponibles en una sola.

Posteriormente, se aplicaron filtros a nivel de columnas según el siguiente detalle:

- **'nro_inscripcion'** >> Se elimina dado que ya fue utilizado para generar el dataset unificado
- **'fecha_nac_alumno', 'fecha_egr_sec' y 'fecha_ingreso_alumno'** >> Se eliminan dado que han sido empleados para generar los atributos de **'edad'** e **'inactividad'**.

- Formateo de los datos



































En este caso se van a evaluar los nuevos atributos incorporados al dataset (edad e inactividad). En este proceso se detectan algunos valores anómalos en el atributo **'inactividad'** y se procede a filtrarlos (valores entre 0 y 38) según recomendación de los expertos en el dominio.

Descripción de metadatos en todo el dataset integrado

[d] dataset_completo.csv

Dataset	Atributo	Tipo de datos	Metadatos
[d]	sexo	numérico	Descripción de valores: min 1.000000 max 2.000000 mean 1.460797 std 0.498570 median 1.000000
	nacionalidad	numérico	min -1.000000 max 4.000000 mean 1.032413 std 0.211005 median 1.000000
	edad	numérico	Columna: edad Tipo de datos: numérico Descripción de valores: min 22.000 max 63.000 mean 27.899693 std 4.925476 median 27.000
	regular	nominal	-- Valores presentes (10 primeros): ['S' 'N'] -- Cantidad de nulos: 0 = 0.00%
	cnt_readmisiones	numérico	min 0.000 max 2.000 mean 0.985545 std 0.717419 median 1.000
	calidad	nominal	-- Valores presentes (10 primeros): ['A' 'P'] -- Cantidad de nulos: 0 = 0.00%
	inactividad	nominal	min 0.000 max 37.000 mean 1.072711 std 2.735806 median 0.000

	estado_civil	numérico	min 1.000000 max 6.000000 mean 1.041612 std 0.316814 median 1.000000
	sit_lab_alumno	nominal	-- Valores presentes (10 primeros): ['NC' 'N' 'S'] -- Cantidad de nulos: 0 = 0.00%
	tipo_res_alumno	numérico	min -1.000000 max 4.000000 mean 0.053438 std 1.489296 median -1.000000
	sit_lab_padres	numérico	min -1.000000 max 5.000000 mean 0.789750 std 2.088225 median -1.000000
	estudios_padres	numérico	min -1.000000 max 7.000000 mean 1.998686 std 2.743089 median 1.000000

Epic	OCT – DIC	
▼  MBID-1 Comprensión del negocio	Sprin...	
 MBID-7 Determinar los objetivos FINALIZADA		
 MBID-8 Evaluación de la situación FINALIZADA		
 MBID-9 Determinación de los obj... FINALIZADA		
 MBID-10 Definir el plan de proyecto FINALIZADA		
▼  MBID-2 Comprensión de los datos		
 MBID-11 Recolección de los datos... FINALIZADA		
 MBID-12 Descripción de los datos FINALIZADA		
 MBID-13 Exploración de los datos FINALIZADA		
 MBID-14 Verificación de la calidad... FINALIZADA		
▼  MBID-3 Preparación de los datos		
 MBID-15 Selección de los datos FINALIZADA		
 MBID-16 Limpieza de los datos FINALIZADA		
 MBID-17 Construcción de datos FINALIZADA		
 MBID-18 Integración de los datos FINALIZADA		
 MBID-19 Formateo de los datos FINALIZADA		
 MBID-4 Modelado		
 MBID-5 Evaluación		
 MBID-6 Despliegue		

[D] Fase de modelado

- Selección de la técnica de modelado

Con base en los objetivos del proyecto, las técnicas a aplicar van a pertenecer a la siguientes familias:

- Árboles de decisión
 - Métodos KNN
 - Regresión logística
 - Métodos de ensamblado de modelos
 - Entre otros
- Generación del plan de pruebas

En primer lugar, a nivel de distribución de filas del dataset resultante de la fase previa se va a trabajar con los estándares de la industria y la bibliografía del área:

- Datos para entrenamiento: 75%
- Datos para prueba: 25%

En segunda instancia, los lineamientos para ejecutar las pruebas de los modelos serán:

- Para cada modelo se deberán registrar sus parámetros de ejecución y la efectividad obtenida al probarse con los datos correspondientes.
 - Como mínimo se ejecutarán **tres instancias** de prueba a través de las cuales se irán seleccionando las técnicas con mejores resultados (efectividad) y seleccionar así la que será usada para la predicción de los datos nuevos (alumnos del ciclo 2021).
- Construcción del Modelo

En este punto se va a utilizar código en lenguaje Python con diferentes librerías para implementar los modelos requeridos. El código fuente se encuentra en la siguiente ubicación:

Repositorio - <https://github.com/jrcastb/13MBID-Metodologias>

Los resultados de esta actividad se registran en las siguientes tablas de la sección de evaluación del modelo.

Prevía a la ejecución de las técnicas seleccionadas se han ejecutado operaciones de transformación y/o adaptación de los datos que se describen a continuación:

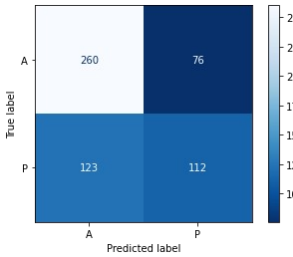
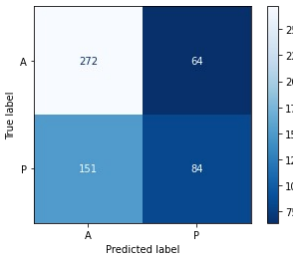
Atributo	Transformación aplicada (valor original => valor nuevo)
'sexo'	<ul style="list-style-type: none"> • 2 => 'F', • 1 => 'M'
'nacionalidad'	<ul style="list-style-type: none"> • 1 => 'AR', • 2 => 'PY', • 3 => 'BR', • 4 => 'OT', • -1 => 'NC' • Observaciones: "OT" = Otros países
'estado_civil'	<ul style="list-style-type: none"> • 1 => 'S', • 2 => 'P', • 3 => 'C', • 4 => 'D', • 5 => 'V', • 6 => 'NC'
'tipo_res_alumno'	<ul style="list-style-type: none"> • 1 : 'CP', • 2 : 'VA', • 3 : 'VP', • 4 : 'RU', • -1 : 'NC'
'sit_lab_padres'	<ul style="list-style-type: none"> • 1 : 'D', • 2 : 'BT', • 3 : 'TI', • 4 : 'TF',

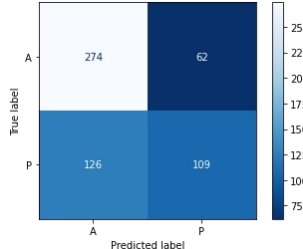
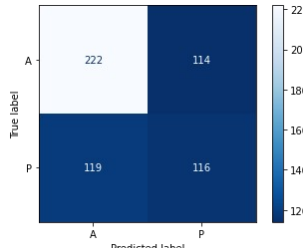
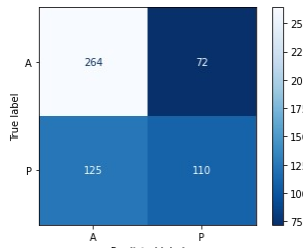
	<ul style="list-style-type: none"> • 5 : 'J', • -1 : 'NC'
'estudios_padres'	<ul style="list-style-type: none"> • 1 : 'SE', • 2 : 'PI', • 3 : 'PC', • 4 : 'SI', • 5 : 'SC', • 6 : 'UI', • 7 : 'UC', • -1 : 'NC'
'cnt_readmisiones'	<ul style="list-style-type: none"> • 1: 'SI', • 2: 'SI', • 0: 'NO'

Más allá de esto, se han binarizado los atributos resultantes para optimizar la ejecución de algunas técnicas de generación de modelos de clasificación / predicción.

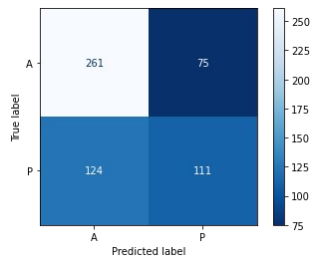
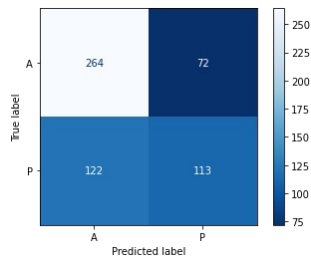
- Evaluación del modelo

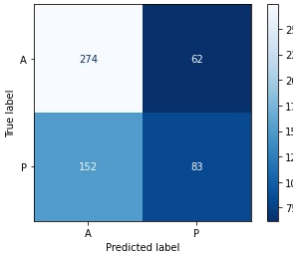
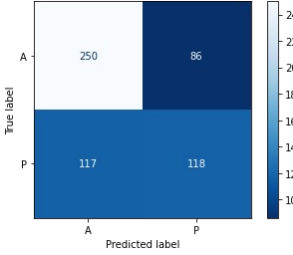
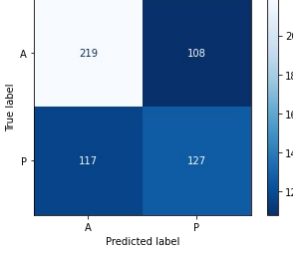
Prueba #1

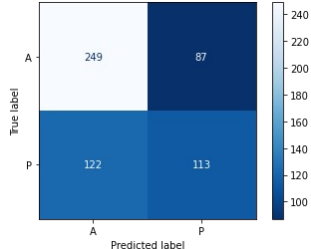
Técnica	Parametrización	Resultados obtenidos
LogisticRegression	LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True, intercept_scaling=1, l1_ratio=None, max_iter=100, multi_class='auto', n_jobs=None, penalty='l2', random_state=None, solver='liblinear', tol=0.0001, verbose=0, warm_start=False)	<p>Rendimiento obtenido: 0.6427320490367776</p> <p>Matriz de confusión:</p> 
KNeighbors	KNeighborsClassifier(algorithm='ball_tree', leaf_size=30, metric='minkowski', metric_params=None, n_jobs=None, n_neighbors=50, p=2, weights='uniform')	<p>Rendimiento obtenido: 0.6234676007005254</p> <p>Matriz de confusión:</p> 
DecisionTree	DecisionTreeClassifier(criterion='entropy', max_depth=6, min_samples_split=10)	<p>Rendimiento obtenido: 0.658493870402802</p> <p>Matriz de confusión:</p>

		
RandomForest	RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None, criterion='gini', max_depth=None, max_features='auto', max_leaf_nodes=None, max_samples=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=17, n_jobs=None, oob_score=False, random_state=None, verbose=0, warm_start=False)	Rendimiento obtenido: 0.5919439579684763 Matriz de confusión: 
GradientBoosting	GradientBoostingClassifier(ccp_alpha=0.0, criterion='friedman_mse', init=None, learning_rate=0.1, loss='deviance', max_depth=1, max_features=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=100, n_iter_no_change=None, presort='deprecated', random_state=0, subsample=0.1, tol=0.0001, validation_fraction=0.1, verbose=0, warm_start=False)	Rendimiento obtenido: 0.6549912434325744 Matriz confusión: 

Prueba #2

Técnica	Parametrización	Resultados obtenidos									
LogisticRegression	LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True, intercept_scaling=1, l1_ratio=None, max_iter=100, multi_class='auto', n_jobs=None, penalty='l2', random_state=None, solver='newton-cg', tol=0.0001, verbose=0, warm_start=False)	<p>Rendimiento obtenido: 0.6514886164623468</p> <p>Matriz de confusión:</p>  <table border="1"> <thead> <tr> <th></th> <th>Actual A</th> <th>Actual P</th> </tr> </thead> <tbody> <tr> <th>Predicted A</th> <td>261</td> <td>124</td> </tr> <tr> <th>Predicted P</th> <td>75</td> <td>111</td> </tr> </tbody> </table>		Actual A	Actual P	Predicted A	261	124	Predicted P	75	111
	Actual A	Actual P									
Predicted A	261	124									
Predicted P	75	111									
LogisticRegression	LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True, intercept_scaling=1, l1_ratio=None, max_iter=100, multi_class='auto', n_jobs=None, penalty='l2', random_state=None, solver='saga', tol=0.0001, verbose=0, warm_start=False)	<p>Rendimiento obtenido:0.660245183887916</p> <p>Matriz de confusión:</p>  <table border="1"> <thead> <tr> <th></th> <th>Actual A</th> <th>Actual P</th> </tr> </thead> <tbody> <tr> <th>Predicted A</th> <td>264</td> <td>122</td> </tr> <tr> <th>Predicted P</th> <td>72</td> <td>113</td> </tr> </tbody> </table>		Actual A	Actual P	Predicted A	264	122	Predicted P	72	113
	Actual A	Actual P									
Predicted A	264	122									
Predicted P	72	113									

KNeighbors	KNeighborsClassifier(algorithm='kd_tree', leaf_size=30, metric='minkowski', metric_params=None, n_jobs=None, n_neighbors=40, p=2, weights='uniform')	<p>Rendimiento obtenido: 0.6252189141856392</p> <p>Matriz de confusión:</p>  <table border="1"> <thead> <tr> <th></th> <th>A</th> <th>P</th> </tr> </thead> <tbody> <tr> <th>A</th> <td>274</td> <td>62</td> </tr> <tr> <th>P</th> <td>152</td> <td>83</td> </tr> </tbody> </table>		A	P	A	274	62	P	152	83
	A	P									
A	274	62									
P	152	83									
DecisionTree	DecisionTreeClassifier(criterion='gini', max_depth=5, min_samples_split=10)	<p>Rendimiento obtenido: 0.6444833625218914</p> <p>Matriz de confusión:</p>  <table border="1"> <thead> <tr> <th></th> <th>A</th> <th>P</th> </tr> </thead> <tbody> <tr> <th>A</th> <td>250</td> <td>86</td> </tr> <tr> <th>P</th> <td>117</td> <td>118</td> </tr> </tbody> </table>		A	P	A	250	86	P	117	118
	A	P									
A	250	86									
P	117	118									
RandomForest	RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None, criterion='gini', max_depth=None, max_features='auto', max_leaf_nodes=None, max_samples=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=10, n_jobs=None, oob_score=False, random_state=None, verbose=0, warm_start=False)	<p>Rendimiento obtenido: 0.6059544658493871</p> <p>Matriz de confusión:</p>  <table border="1"> <thead> <tr> <th></th> <th>A</th> <th>P</th> </tr> </thead> <tbody> <tr> <th>A</th> <td>219</td> <td>108</td> </tr> <tr> <th>P</th> <td>117</td> <td>127</td> </tr> </tbody> </table>		A	P	A	219	108	P	117	127
	A	P									
A	219	108									
P	117	127									
GradientBoosting	GradientBoostingClassifier(ccp_alpha=0.0, criterion='friedman_mse', init=None, learning_rate=1.0, loss='deviance', max_depth=1, max_features=None, max_leaf_nodes=None,	<p>Rendimiento obtenido: 0.6339754816112084</p> <p>Matriz de confusión:</p>									

	<pre> min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=100, n_iter_no_change=None, presort='deprecated', random_state=0, subsample=0.5, tol=0.0001, validation_fraction=0.1, verbose=0, warm_start=False) </pre>	 <table border="1"> <thead> <tr> <th></th> <th>Predicted label: A</th> <th>Predicted label: P</th> </tr> </thead> <tbody> <tr> <th>True label: A</th> <td>249</td> <td>87</td> </tr> <tr> <th>True label: P</th> <td>122</td> <td>113</td> </tr> </tbody> </table>		Predicted label: A	Predicted label: P	True label: A	249	87	True label: P	122	113
	Predicted label: A	Predicted label: P									
True label: A	249	87									
True label: P	122	113									

[E] Fase de evaluación

- Evaluación de los resultados

A partir de la experimentación realizada la/s técnicas a emplear sobre los datos nuevos del ciclo 2021 son:

Técnica utilizada	Mejor rendimiento obtenido
Decision Tree	0.658493870402802
Logistic Regression	0.6427320490367776
Random Forest	0.6059544658493871
Gradient Boosting	0.6549912434325744
KNeighbors	0.6252189141856392

- **Proceso de revisión**

En función de los resultados obtenidos se ha determinado que la técnica a emplear para proseguir con el proyecto ha sido la de: **Decision Tree**.

En este sentido, los resultados de la ejecución de la predicción con el modelo de la técnica mencionada sobre los datos del ciclo 2021 se encuentran en la sección **Informe Final** del presente documento.

- **Determinación de futuras tareas**

Como tareas a ejecutar que permitirían un mejor rendimiento del modelo de predicción generado se pueden mencionar:

- *A fin de contar con una mayor información sobre los estudiantes se podrían incorporar atributos que registren la distancia entre su lugar de residencia y el lugar de cursado de las asignaturas de la carrera seleccionada.*
- *Se podría indagar acerca del monto salarial promedio de los padres y los estudiantes para determinar si influye en la deserción y mortandad académica de carreras y asignaturas.*
- *Sería de ayuda obtener información acerca de que estudiantes cuentan con hijos, con el fin de determinar si esto influye en su situación académica.*

[F] Fase de implementación

- **Plan de implementación**

Las autoridades han dispuesto la replicación de las acciones ejecutadas en el presente proyecto al inicio de cada ciclo lectivo. De esta manera se garantiza la disponibilidad de los datos y de los recursos para realizar el análisis aquí descrito.

Ante cada ejecución del proyecto se deberán adaptar las tareas involucradas en el mismo a raíz de los avances que se pudieran haber logrado previamente.

- **Supervisión y Mantenimiento**

Una vez que el producto desarrollado se encuentre en funcionamiento, el monitoreo a realizar consistirá en:

- Monitoreo de consultas diarias
- Comparación de efectividad contra la realidad de los estudiantes al final del ciclo lectivo
- Chequear nuevas entradas de datos en los datasets originales y probar nuevamente los modelos
- Visualizar los datos para tener un tracking mas entendible de los mismos

- **Informe Final**

Se presentan los resultados obtenidos preliminarmente para los estudiantes del ciclo 2021 en función de la predicción realizada con base en los datos históricos disponibles:

	Cantidad	Porcentaje
Activos	29	82.8%
Pasivos	6	17.2%
Totales	35	100%

- Revisión del proyecto

Incorporar cuestiones a revisar y/o mejorar referidas al proceso mediante el cual se gestionó y ejecutó el proyecto.

- Se puede aplicar un PCA o un análisis de correlaciones para llegar a determinar los atributos o variables que influyen en mayor medida, con el fin de optimizar el modelo planteado y requerir menor cantidad de recursos a la hora de entrenarlo.
- Calidad de los datos (previa al inicio del proyecto) En cuanto a la calidad de los datos, es requerido tal vez un par de atributos propuestos durante el proceso de revisión. Esto con el fin de analizar su posible influencia en la predicción del modelo
- Se recomienda la implementación de un repositorio TDSP estandarizado con el fin de mejorar la eficiencia en la ejecución del proyecto
- Son requeridas diariamente reuniones de aproximadamente 12 a 15min para realizar una revisión de los avances del proyecto, inconvenientes y posibles nuevas solicitudes sobre el mismo.

Visualización Historial Marcadores Ventana Ayuda

Upload files · jrcastb/13MBID-Metodologias MBID13: tablero ágil - Jira

trabajo Proyectos Filtros Paneles Personas Aplicaciones Crear

Proyectos / MBID13


































Backlog

Buscar JC Epic Insights

▼ Sprint 2 20 nov. - 27 nov. (11 incidencias) 0 0 40 Completar sprint

Cerrar la generación del MVP del proyecto

ID	Título	Etiqueta	Asignado a	Estado	Usuario
MBID-44	Selección la(s) técnicas del modelado	MODELADO	4	FINALIZADA	JC
MBID-45	Generación de un plan de pruebas	MODELADO	2	FINALIZADA	JC
MBID-46	Construcción del modelo	MODELADO	8	FINALIZADA	JC
MBID-47	Evaluación del modelo	MODELADO	5	FINALIZADA	JC
MBID-48	Evaluación de los resultados	EVALUACIÓN	3	FINALIZADA	JC
MBID-49	Proceso de revisión	EVALUACIÓN	3	FINALIZADA	JC
MBID-50	Determinación de futuras líneas de trabajo	EVALUACIÓN	5	FINALIZADA	JC
MBID-51	Plan de implantación	DESPLIEGUE	2	FINALIZADA	JC
MBID-52	Supervisión y mantenimiento	DESPLIEGUE	1	FINALIZADA	JC
MBID-53	Redacción de informe final	DESPLIEGUE	5	FINALIZADA	JC
MBID-54	Revisión del proyecto	DESPLIEGUE	2	FINALIZADA	JC

Epic	OCT – DIC	
>  <u>MBID-1 Comprensión del negocio</u>		
>  <u>MBID-2 Comprensión de los datos</u>		
>  <u>MBID-3 Preparación de los datos</u>		
▼  <u>MBID-4 Modelado</u>		
 MBID-44 Selección l... FINALIZADA JOSE R C...		
 MBID-45 Generación... FINALIZADA JOSE R C...		
 MBID-46 Construcci... FINALIZADA JOSE R C...		
 MBID-47 Evaluación... FINALIZADA JOSE R C...		
▼  <u>MBID-5 Evaluación</u>		
 MBID-48 Evaluación... FINALIZADA JOSE R C...		
 MBID-49 Proceso de... FINALIZADA JOSE R C...		
 MBID-50 Determinac... FINALIZADA JOSE R C...		
▼  <u>MBID-6 Despliegue</u>		
 MBID-51 Plan de imp... FINALIZADA JOSE R C...		
 MBID-52 Supervisión... FINALIZADA JOSE R C...		
 MBID-53 Redacción... FINALIZADA JOSE R C...		
 MBID-54 Revisión de... FINALIZADA JOSE R C...	