

viu  
.es



# ACTIVIDAD GUIADA 3

**Máster en Big Data y Data Science**

**Metodologías de gestión y diseño de proyectos Big Data**

**Nombre: Carlos Alberto Mejía Rodríguez**

**Fecha: mayo 2023**

**viu**

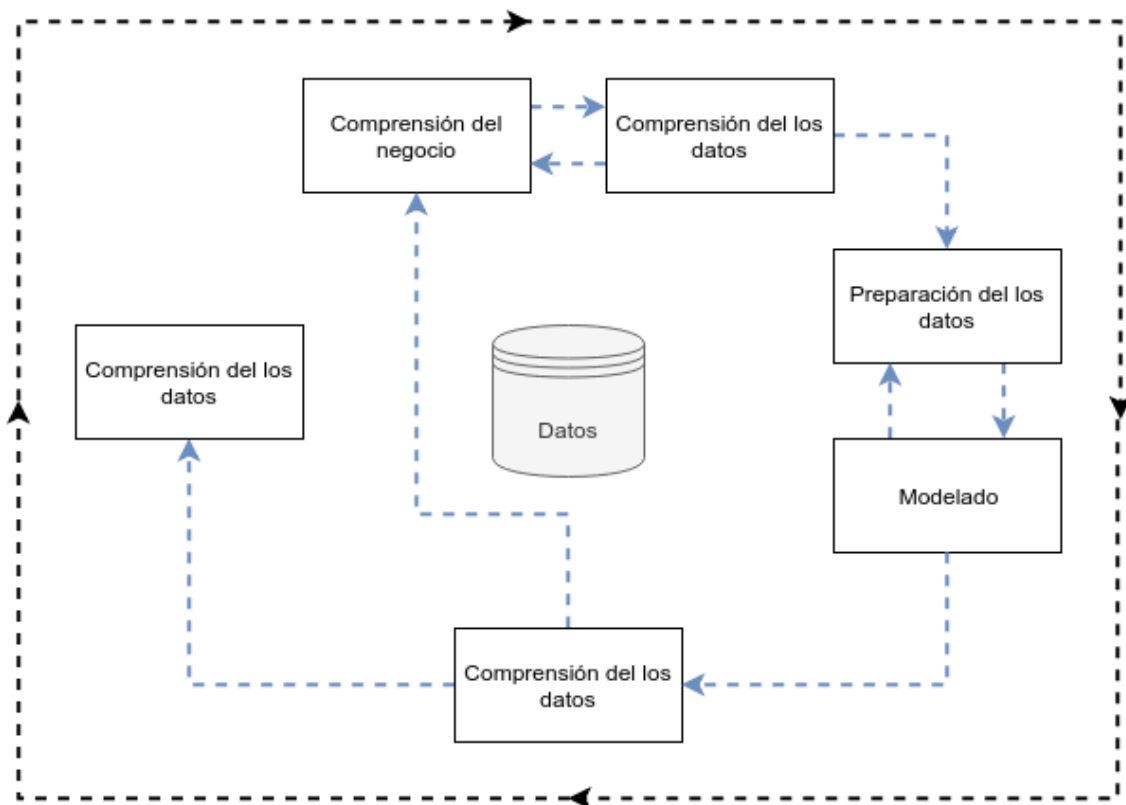
**Universidad  
Internacional  
de Valencia**

## 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 2022.

- **Evaluación de la situación**

Se cuenta con los siguientes recursos para la ejecución del proyecto:

- Los datos históricos de los estudiantes distribuidos en las tres dimensiones de análisis (inscripción, cursado del 1er. semestre y datos académicos al final el año)
- Se cuenta con el personal para la realización de las tareas del proyecto.
- Se cuenta con las herramientas y equipos necesarios para la experimentación e implementación requerida.

- **Determinación de los objetivos del proyecto**

Elaborar un **modelo de predicción** de la situación académica de los alumnos de la Unidad Académica (Facultad / Escuela) en cuestión en base a los datos históricos de la misma.

La efectividad mínima del modelo de predicción deberá ser del **80%**.

- **Definir plan del proyecto**

- **Comprensión del negocio:** Es necesario comprender el sector o la industria en la que opera la organización, en este caso es el sector Educativo, su modelo de negocio es la formación en educación superior, sus procesos son académicos y de gestión relacionada con los mismos, como todo ente educador uno de sus principales desafíos es la deserción de sus estudiantes y en cuanto a las oportunidades con que cuenta, podría resaltar su gran motivación por la aplicación de la innovación.
- **Especificar recursos disponibles:** Se cuenta con los dataset clasificados para el desarrollo del proyecto de big data, el personal técnico calificado y los equipos de computo necesarios para el procesamiento.
- **Especificar tareas:** Se agrupa y organiza el trabajo que debe realizarse para el desarrollo del proyecto en épicas, cada épica será una fase o tarea de gran tamaño que a su vez contiene tareas más desglosadas.

A continuación, se presentan capturas de pantalla que evidencia las tareas y épicas asociadas que permitirán el desarrollo integro del proyecto.

Figura 1. Product backlog

carlosmejiaRodriguez / 13MBID\_2023\_AG\_2\_3 / Boards / Backlogs

13MBID\_2023\_AG\_2\_3 Team

Backlog Analytics + New Work Item View as Board Column Options ...

Order	Work Item Type	Title	State	Effort	Busin...	Value Area	Tags
1	Epic	Comprensión del negocio	New			Business	
	Product Backl...	Determinar los objetivos de la organización	Done	1		Business	
	Product Backl...	Evaluación de la situación	Done	2		Business	
	Product Backl...	Determinar los objetivos del Proyecto	Done	3		Business	
	Product Backl...	Definir plan del proyecto	Done	8		Business	
2	Epic	Comprensión de los Datos	New			Business	
	Product Backl...	Recolección de datos iniciales	Approved	2		Business	
	Product Backl...	Descripción de los datos	Approved	2		Business	
	Product Backl...	Exploración de los datos	Approved	3		Business	
	Product Backl...	Verificación de la calidad de los datos	Approved	8		Business	
3	Epic	Preparación de los datos	New			Business	
	Product Backl...	Selección de los datos	New	3		Business	
	Product Backl...	Limpieza de los datos	New	3		Business	
	Product Backl...	Construcción de datos	New	5		Business	
	Product Backl...	Integración de los datos	New	2		Business	
	Product Backl...	Formateo de datos	New	3		Business	
4	Epic	Modelado	New	3		Business	

Figura 2. Sprint backlog

carlosmejiaRodriguez / 13MBID\_2023\_AG\_2\_3 / Boards / Sprints

13MBID\_2023\_AG\_2\_3 Team

Taskboard **Backlog** Capacity Analytics | + New Work Item Column Options ...

Order	Title	State	Assigned To	Rema...
1	<ul style="list-style-type: none"> <li>Determinar los objetivos de la organización</li> <li>Relevar los objetivos de negocio</li> </ul>	Done	Carlos Mejia Rodriguez	
2	<ul style="list-style-type: none"> <li>Evaluación de la situación</li> <li>Reconocer recursos de información disponibles</li> </ul>	Done	Carlos Mejia Rodriguez	
3	<ul style="list-style-type: none"> <li>Determinar los objetivos del Proyecto</li> <li>especificar los objetivos de efectividad del modelo a generar</li> <li>Especificar los resultados del desarrollo</li> </ul>	Done	Carlos Mejia Rodriguez	
4	<ul style="list-style-type: none"> <li>Definir plan del proyecto</li> <li>Especificar tareas</li> <li>Especificar tiempos</li> <li>Especificar recursos disponibles</li> </ul>	Done	Carlos Mejia Rodriguez	
5	<ul style="list-style-type: none"> <li>Recolección de datos iniciales</li> <li>Recuperar los archivos de datos</li> <li>Verificar la integridad de los datos</li> </ul>	App... In Pr... In Pr...	Carlos Mejia Rodriguez	
6	Descripción de los datos	App...	Carlos Mejia Rodriguez	
7	Exploración de los datos	App...	Carlos Mejia Rodriguez	
8	Verificación de la calidad de los datos	App...	Carlos Mejia Rodriguez	

### ● Personas y Roles

En cuanto al **Especialista en el dominio** se tomarán como miembros a los docentes de la asignatura quienes han proveído de gran información mediante las guías y videoconferencias.

Como **equipo de minería de datos** se adelantará el desarrollo del trabajo de forma individual, apoyado en los materiales suministrados por los docentes de la asignatura.

## [B] Fase de comprensión de los datos

- **Recolección de datos iniciales**

Se han recuperado los datos históricos agrupados en 3 dimensiones / archivos:

- **Datos de inscripciones:** referidos a la inscripción de cada estudiante a alguna de las carreras de la unidad académica.
- **Datos del 1er cuatrimestre:** referidos al rendimiento académico de cada estudiante en su primer cuatrimestre de cursado y el primer periodo de exámenes.
- **Datos académicos:** referidos al avance en la carrera de cada estudiante y su situación académica actualizada

Los datos han sido verificados contra los orígenes y se encuentran aptos para su uso.

- **Descripción de los datos**

Se describen las características principales de cada dataset:

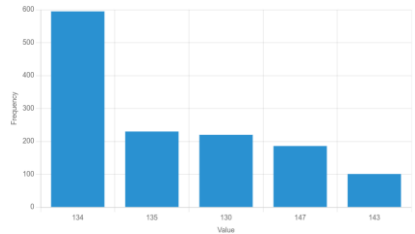
Dataset	Columnas / Atributos	Cantidad de filas
datos_inscripciones	<b>id_estudiante</b> propuesta estado_inscripcion plan_estudios version_plan modalidad fecha_ingreso fecha_inscripcion	Cantidad de filas: 1332
datos_cursado	<b>id_estudiante</b> estado_inscripcion ingreso_aprobadas ingreso_libres ingreso_totales cursadas_aprobadas cursadas_regulares cursadas_libres cursadas_totales inscripciones_exámenes exámenes_aprobados	Cantidad de filas: 1332
datos_academicos	<b>id_estudiante</b> plan anio_ingreso fecha_ingreso fecha_ultimo_examen	Cantidad de filas: 815

	anio_ultima_reinscripcion promedio_sin_aplazos promedio_con_aplazos actividades_aprobadas total_actividades regular calidad segundo_anio	
--	---------------------------------------------------------------------------------------------------------------------------------------------------------------	--

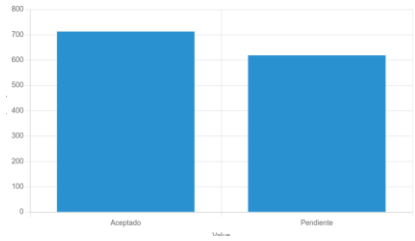
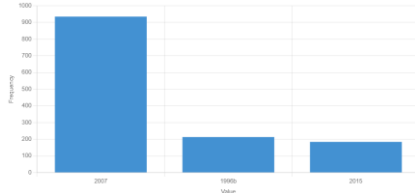
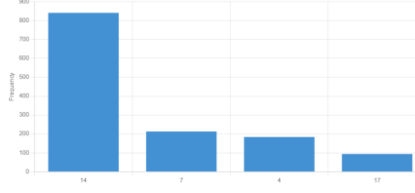
- Exploración de datos

Se describen a continuación los meta-datos de cada dataset:

- [a] datos\_inscripciones

Dataset	Columna / Atributo	Tipo de datos	Observaciones
[a]	id_estudiante	String	<p>Atributo con un formato especial: <b>##-#####-#</b></p> <p>Valores de ejemplo: CA-000281-5, CA-003269-2, CA-004968-4, CA-005300-2, CA-006018-1</p> <p>Hace referencia al número de matrícula de cada estudiante.</p> <p>Cantidad de nulos = 0</p>
	propuesta	Integer	<p>Descripción y distribución de valores (cantidad):</p> <p><b>134</b> (595)  <b>135</b> (230)  <b>130</b> (220)  <b>147</b> (186)  <b>143</b> (101)</p> <p>Cantidad de nulos = 0</p> 

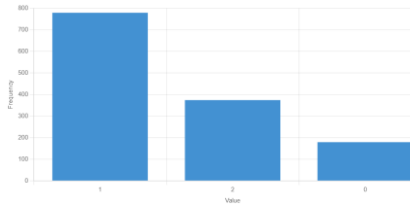
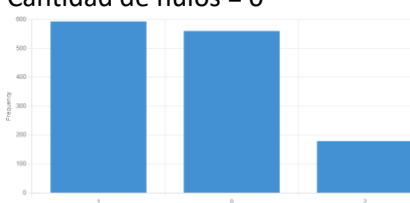
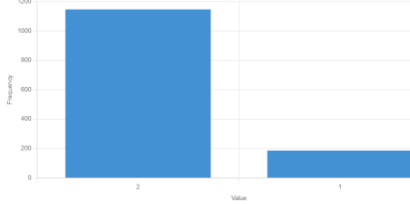


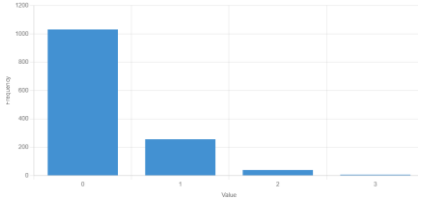
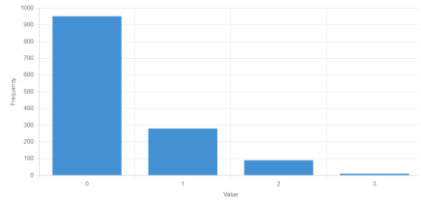
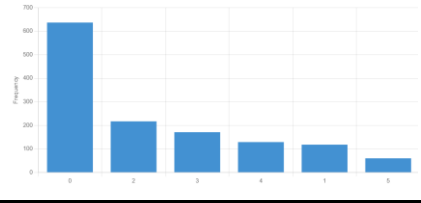
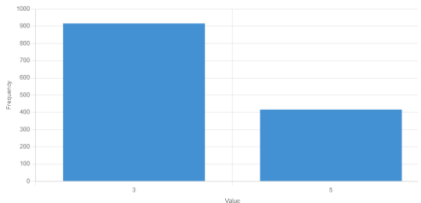
	estado_inscripcion	String	<p>Descripción y distribución de valores (cantidad):</p> <p><b>Aceptado</b> (713)</p> <p><b>Pendiente</b> (619)</p> <p>Cantidad de nulos = 0</p> 
	plan_estudios	String	<p>Descripción y distribución de valores (cantidad):</p> <p><b>2007</b> (935)</p> <p><b>1996b</b> (213)</p> <p><b>2015</b> (184)</p> <p>Cantidad de nulos = 0</p> 
	version_plan	Integer	<p>Descripción y distribución de valores (cantidad):</p> <p><b>14</b> (841)</p> <p><b>7</b> (213)</p> <p><b>4</b> (184)</p> <p><b>17</b> (94)</p> <p>Cantidad de nulos = 0</p> 
	modalidad	String	<p>Descripción y distribución de valores (cantidad):</p> <p><b>Presencial</b> (1332)</p> <p>Cantidad de nulos = 0</p> <p>La cantidad de valores diferentes es:</p> <p>1</p> <p>Valor único, atributo sin variación.</p>

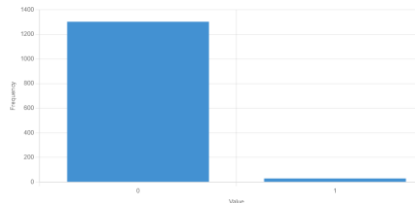
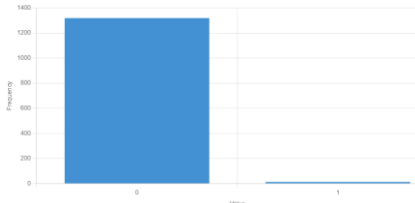
	fecha_ingreso	String	<p>Formato de fecha: <i>dd/mm/aaaa</i></p> <p>La cantidad de valores diferentes es: 1</p> <p>El valor sin variación (cantidad): <b>01/04/2020</b> (1332)</p>
	fecha_inscripcion	String	<p>Formato de fecha: <i>dd/mm/aaaa</i></p> <p>La cantidad de valores diferentes es: 77</p> <p>Los más frecuentes son: <b>04/02/2020</b> (73) <b>14/02/2020</b> (71) <b>18/12/2019</b> (68) <b>17/12/2019</b> (65)</p> <p>Cantidad de nulos = 0</p>

- [b] datos\_cursado

Dataset	Columna / Atributo	Tipo de datos	Observaciones
[b]	id_estudiante	String	<p>Atributo con un formato especial: <i>##-#####-#</i></p> <p>Valores de ejemplo: CA-000281-5, CA-003269-2, CA-004968-4, CA-005300-2, CA-006018-1</p> <p>Hace referencia al número de matrícula de cada estudiante.</p> <p>Cantidad de nulos = 0</p>
	estado_inscripción	String	<p>Descripción y distribución de valores (cantidad): <b>Aceptado</b> (713) <b>Pendiente</b> (619)</p> <p>Cantidad de nulos = 0</p>

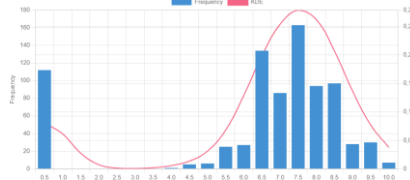
			
	ingreso_aprobadas	integer	<p>Descripción y distribución de valores (cantidad):</p> <p><b>1</b> (779)</p> <p><b>2</b> (374)</p> <p><b>0</b> (179)</p> <p>Cantidad de nulos = 0</p> 
	ingreso_libres	integer	<p>Descripción y distribución de valores (cantidad):</p> <p><b>1</b> (593)</p> <p><b>0</b> (560)</p> <p><b>2</b> (179)</p> <p>Cantidad de nulos = 0</p> 
	ingreso_totales	integer	<p>Descripción y distribución de valores (cantidad):</p> <p><b>2</b> (1146)</p> <p><b>1</b> (186)</p> <p>Cantidad de nulos = 0</p> 
	cursadas_aprobadas	integer	<p>Descripción y distribución de valores (cantidad):</p> <p><b>0</b> (1031)</p> <p><b>1</b> (256)</p> <p><b>2</b> (39)</p>

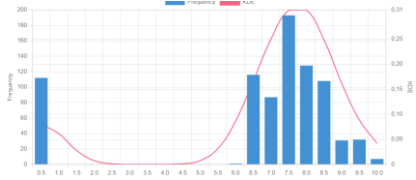
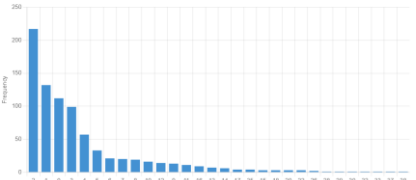
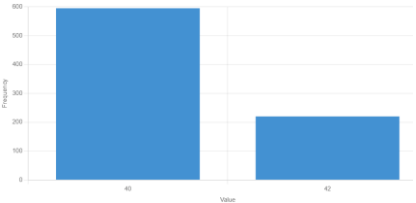
			<b>3 (6)</b> Cantidad de nulos = 0 
	cursadas_regulares	integer	Descripción y distribución de valores (cantidad): <b>0 (952)</b> <b>1 (280)</b> <b>2 (90)</b> <b>3 (10)</b> Cantidad de nulos = 0 
	cursadas_libres	integer	Descripción y distribución de valores (cantidad): <b>0 (637)</b> <b>2 (217)</b> <b>3 (171)</b> <b>4 (129)</b> <b>1 (118)</b> <b>5 (60)</b> Cantidad de nulos = 0 
	cursadas_totales	integer	Descripción y distribución de valores (cantidad): <b>3 (916)</b> <b>5 (416)</b> Cantidad de nulos = 0 

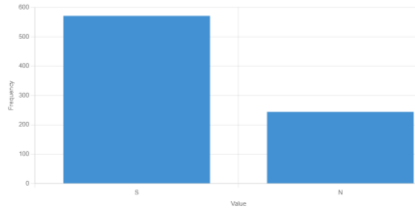
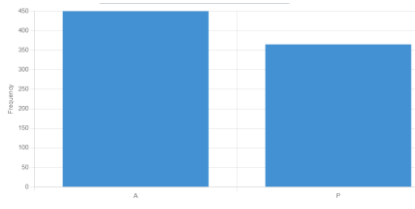
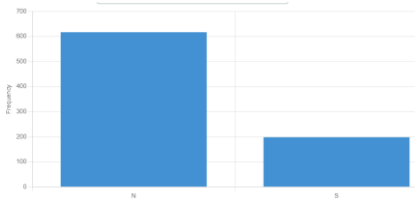
	inscripciones_exámenes	integer	<p>Descripción y distribución de valores (cantidad):</p> <p><b>0</b> (1303)</p> <p><b>1</b> (29)</p> <p>Cantidad de nulos = 0</p> 
	exámenes_aprobados	integer	<p>Descripción y distribución de valores (cantidad):</p> <p><b>0</b> (1319)</p> <p><b>1</b> (13)</p> <p>Cantidad de nulos = 0</p> 

- [c] datos\_academicos

Dataset	Columna / Atributo	Tipo de datos	Observaciones
[c]	id_estudiante	String	<p>Atributo con un formato especial: <b>##-#####-#</b></p> <p>Valores de ejemplo: CA-000281-5, CA-003269-2, CA-004968-4, CA-005300-2, CA-006018-1</p> <p>Hace referencia al número de matrícula de cada estudiante.</p> <p>Cantidad de nulos = 0</p>
	plan	integer	<p>Descripción y distribución de valores (cantidad):</p> <p><b>2007</b> (815)</p> <p>Valores únicos: 1</p> <p>Cantidad de nulos = 0</p>

	anio_ingreso	integer	<p>Descripción y distribución de valores (cantidad):  <b>2020</b> (815)  Valores únicos: 1  Cantidad de nulos = 0</p>
	fecha_ingreso	string	<p>Formato de fecha:  dd/mm/aaaa</p> <p>Valores únicos: 1  valor (cantidad)  <b>01/04/2020</b> (815)</p> <p>Cantidad de nulos = 0</p>
	fecha_ultimo_examen	string	<p>Formato de fecha:  dd/mm/aaaa</p> <p>Valores únicos: 24  Los más frecuentes son:  <b>18/08/2021</b> (48)  <b>20/08/2021</b> (40)  <b>23/08/2021</b> (36)  <b>19/08/2021</b> (16)  ...  Cantidad de nulos = 596 (73.13%)</p>
	anio_ultima_reinscripcion	integer	<p>Descripción y distribución de valores (cantidad):  <b>2021</b> (450)  Valores únicos: 1  Cantidad de nulos = 365 (44.79%)</p>
	promedio_sin_aplazos	float	<p>Descripción y distribución de valores (cantidad) con más frecuencias:  <b>0</b> (112), <b>7</b> (103), <b>6</b> (100), <b>8</b> (84),  <b>7.5</b> (48), <b>6.5</b> (30), <b>9</b> (28), <b>7.33</b> (23),  <b>8.5</b> (23), <b>7.67</b> (19) ...  Cantidad de nulos = 0</p>  <p>The figure is a histogram showing the frequency distribution of the 'promedio_sin_aplazos' variable. The x-axis represents the average score, ranging from 0.5 to 10.0. The y-axis represents the frequency, ranging from 0 to 160. The histogram bars are blue. A red line represents the normal distribution curve (KDE). The distribution is roughly bell-shaped, centered around 7.5.</p>

	promedio_con_aplazos	float	<p>Descripción y distribución de valores (cantidad) con más frecuencias:</p> <p><b>7 (114), 0 (112), 6 (96), 8 (90), 7.5 (56), 6.5 (31), 9 (30), 7.33 (27), 7.67 (24), 8.5 (24)</b></p> <p>Cantidad de nulos = 0</p> 
	actividades_aprobadas	integer	<p>Descripción y distribución de valores (cantidad) con más frecuencias:</p> <p><b>2 (217), 1 (132), 0 (112), 3 (99), 4 (57), 5 (33), 6 (21), 7 (20), 8 (19), 10 (16), 12 (14)</b></p> <p>Cantidad de nulos = 0</p> 
	total_actividades	integer	<p>Descripción y distribución de valores (cantidad):</p> <p><b>40 (595)</b> <b>42 (220)</b></p> <p>Cantidad de nulos = 0</p> 
	regular	string	<p>Descripción y distribución de valores (cantidad):</p> <p><b>S (571)</b> <b>N (244)</b></p> <p>Cantidad de nulos = 0</p>

			
	calidad	string	<p>Descripción y distribución de valores (cantidad):  <b>A</b> (450)  <b>P</b> (365)            Cantidad de nulos = 0</p> 
	segundo_anio	string	<p>Descripción y distribución de valores (cantidad):  <b>N</b> (617)  <b>S</b> (198)            Cantidad de nulos = 0</p> 



## Gráficos de interés

Figura 3. Pearson Correlation Matrix dataset: df\_inscripciones

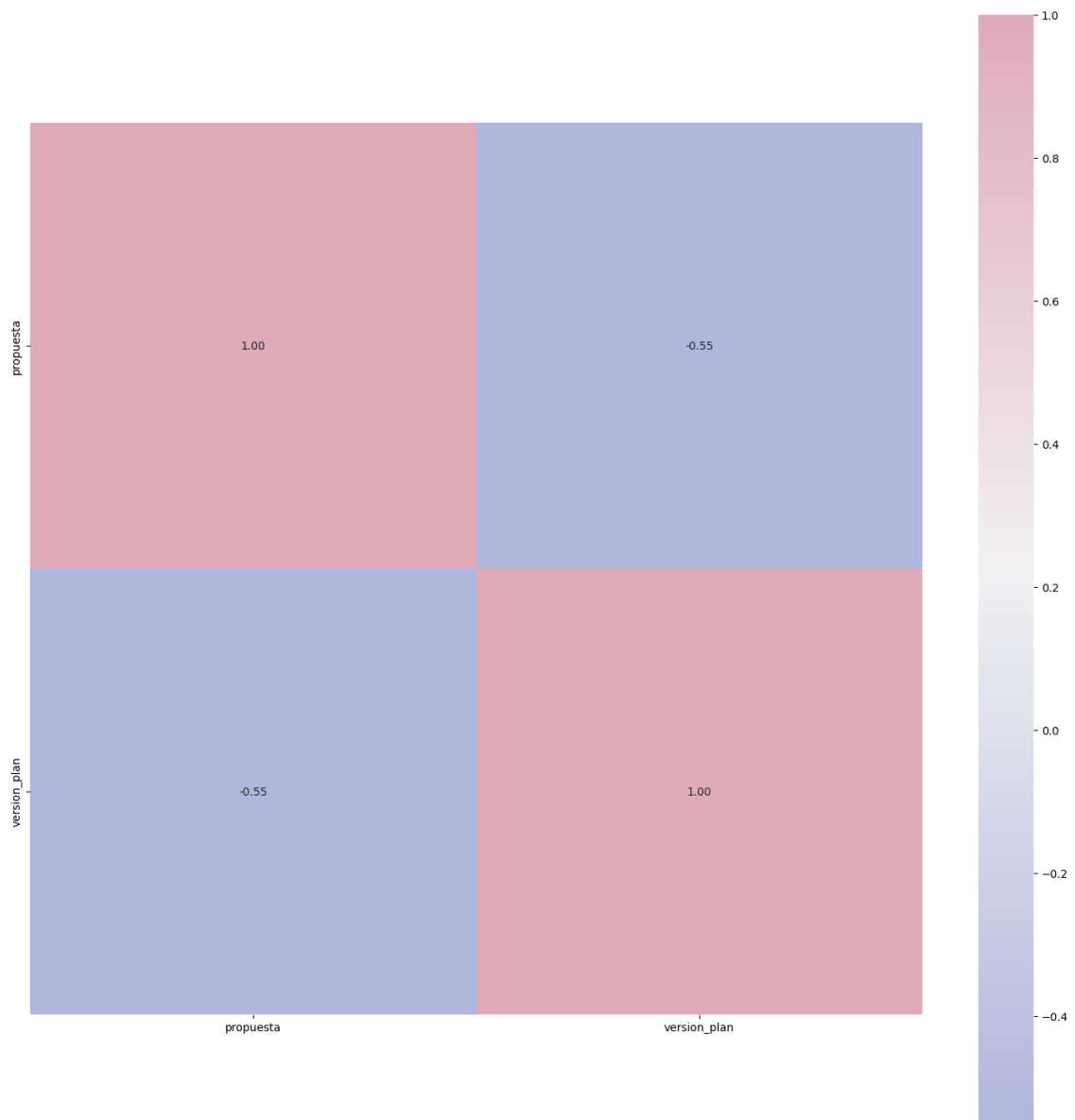


Figura 4. Pearson Correlation Matrix dataset: df\_cursado

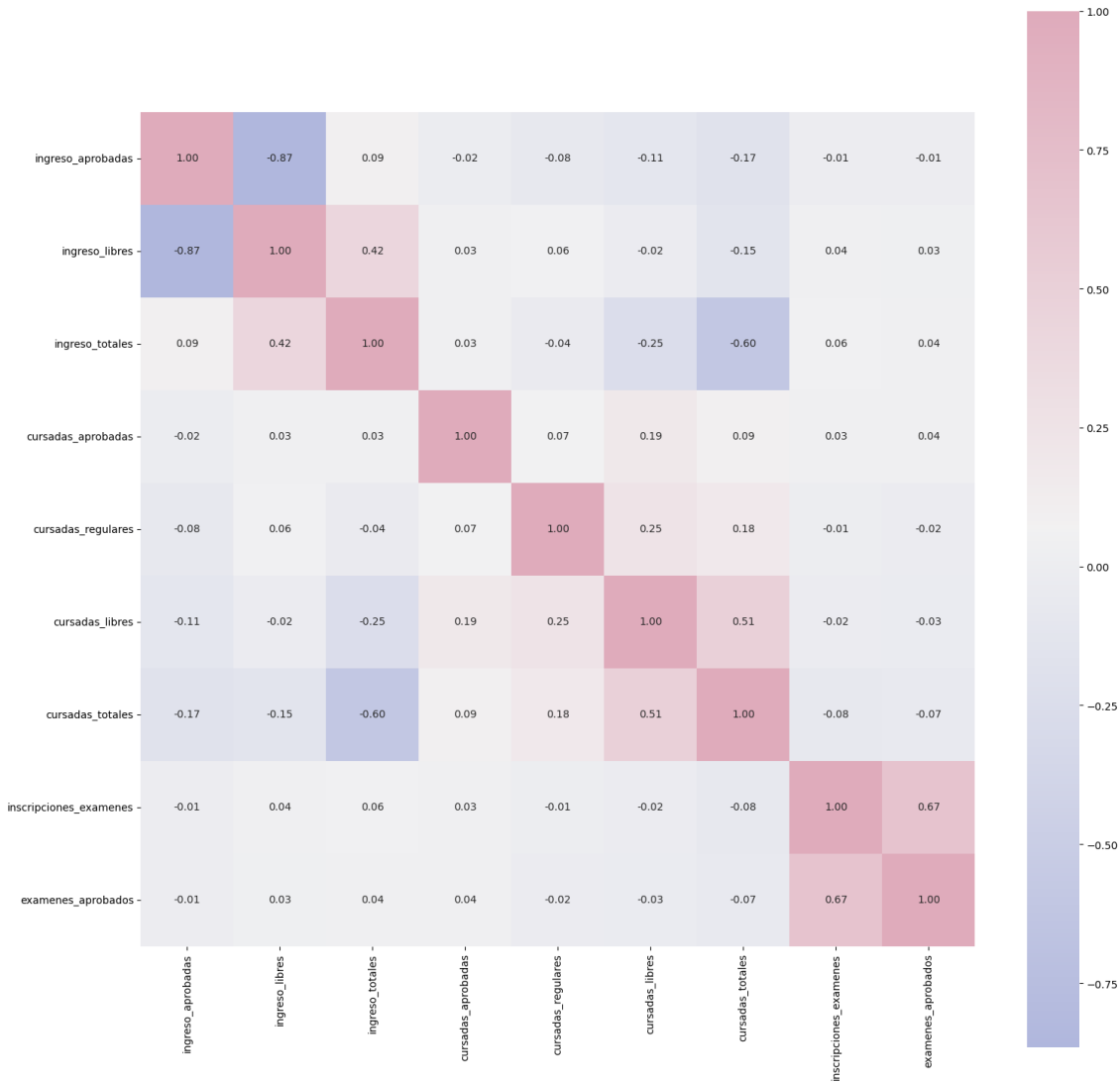
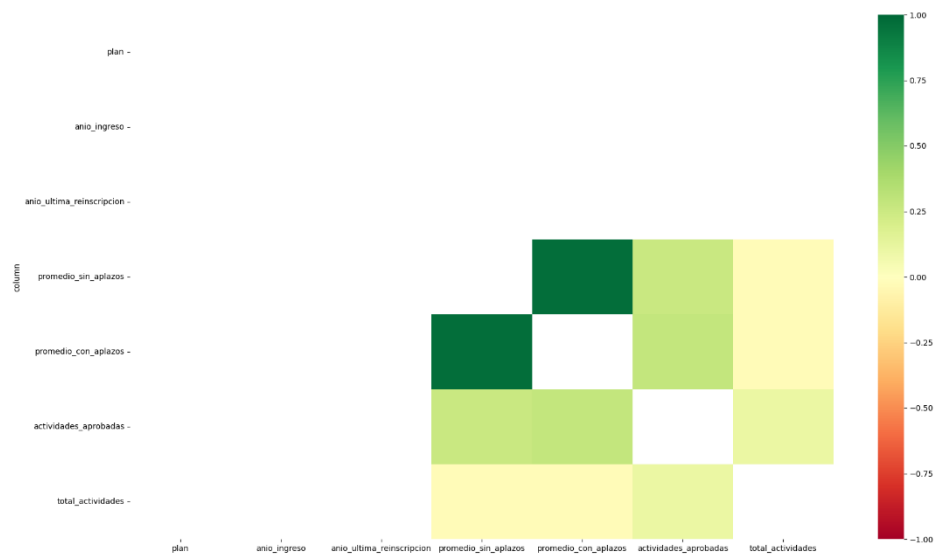


Figura 5. Pearson Correlation Matrix dataset: df\_academicos



- Verificación de la calidad de los datos

Se procede a verificar la calidad de los datos de los tres datasets. Los resultados se describen a continuación:

- [a] datos\_inscripcion
- [b] datos\_cursado
- [c] datos\_academicos

Dataset	Columna / Atributo / Item	Descripción / Observaciones
[a]	id_estudiante	<p>Se verifica el cumplimiento del formato establecido y comunicado por los expertos del negocio. Resultados:</p> <p>Conteo de errores en el formato del atributo id_estudiante:</p> <ul style="list-style-type: none"><li>- Dataset: datos_inscripciones.csv =&gt; <b>Cantidad: 0   0.0%</b></li><li>- Dataset: datos_cursado.csv =&gt; <b>Cantidad: 0   0.0%</b></li><li>- Dataset: datos_academicos.csv =&gt; <b>Cantidad: 0   0.0%</b></li></ul>

[a]	fecha_ingreso fecha_inscripcion	<p>Se verifica el cumplimiento de la regla del negocio que establece que la <b>fecha_inscripcion</b> no puede ser mayor a la <b>fecha_ingreso</b>.</p> <p>Resultados:</p> <p>Casos de problemas detectados: <b>46 filas</b>  Porcentaje de filas con problemas de valores fuera de reglas de negocio: <b>3.45 %</b></p>
[b]	estado_inscripcion cursadas_aprobadas cursadas_regulares cursadas_libres inscripciones_exámenes exámenes_aprobados	<p>Se verifica el cumplimiento de la regla del negocio que establece que un estudiante con inscripción en estado “<b>Pendiente</b>” no puede registrar actividad académica (<i>esto es: cursadas o exámenes</i>)</p> <p>Resultados:</p> <p>Casos de problemas de valores por reglas: <b>12 filas</b>  Porcentaje de filas con problemas de valores fuera de reglas de negocio: <b>0.9 %</b></p>
[c]	calidad regular	<p>Se verifica la regla del negocio que establece que no pueden darse las siguientes combinaciones de valores para estos atributos:</p> <ul style="list-style-type: none"> <li>• <math>regular = N \mid calidad = A</math></li> <li>• <math>regular = S \mid calidad = P</math></li> </ul> <p>Resultados:</p> <p>Casos de problemas de valores por reglas: <b>197 filas</b>  Porcentaje de filas con problemas de valores fuera de reglas de negocio: <b>14.79 %</b></p>
[c]	Identificación de outliers en atributos numéricos	<p>Para garantizar la calidad de los datos se verifican e identifican los siguientes atributos con outliers.</p> <p>Los atributos con outliers y los umbrales:</p> <p><b>promedio_sin_aplazos</b>  umbral: (<math>(\text{promedio\_sin\_aplazos} &lt; 3.64)</math> or <math>(\text{promedio\_sin\_aplazos} &gt; 9.93)</math>)</p>

		<p><b>promedio_con_aplazos</b> umbral: `promedio_con_aplazos` &lt; 3.44</p> <p><b>actividades_aprobadas</b> umbral: `actividades_aprobadas` &gt; 8.5</p>
--	--	------------------------------------------------------------------------------------------------------------------------------------------------------------------

## [C] Fase de preparación de los datos

- **Selección de datos**

Se describe a continuación la selección de columnas realizada inicialmente en los datasets:

- Dataset **datos\_inscripciones**, columnas eliminadas:
  - 'plan\_estudios'
  - 'version\_plan'
  - 'modalidad'
  - 'err\_formato\_matricula' (\*)
  - 'regla\_fechas\_ingreso' (\*)
- Dataset **datos\_academicos**, columnas eliminadas:
  - 'plan'
  - 'fecha\_ingreso'
  - 'err\_formato\_matricula'
  - 'regla\_verificacion\_calidad' (\*)
- Dataset **datos\_cursado**, columnas eliminadas
  - 'err\_formato\_matricula' (\*)
  - 'regla\_estado\_inscripcion' (\*)

*(\*) Se trata de columnas agregadas como parte del proceso de verificación de calidad de los datos. No se corresponden a datos del dominio.*

Finaliza el desarrollo de la AG2.

- Estado del proyecto en Azure Boards con el desarrollo de los puntos de la AG2.

13MBID\_2023\_AG\_2\_3 Team

Backlog Analytics + New Work Item View as Board Column Options

Order	Work Item Type	Title	State	Effort	Busin...	Value Area
1	Epic	Comprensión del negocio	New			Business
	Product Backl...	Determinar los objetivos de la organización	Done	1		Business
	Product Backl...	Evaluación de la situación	Done	2		Business
	Product Backl...	Determinar los objetivos del Proyecto	Done	3		Business
	Product Backl...	Definir plan del proyecto	Done	8		Business
2	Epic	Comprensión de los Datos	New			Business
	Product Backl...	Recolección de datos iniciales	Done	2		Business
	Product Backl...	Descripción de los datos	Done	2		Business
	Product Backl...	Exploración de los datos	Done	3		Business
	Product Backl...	Verificación de la calidad de los datos	Done	8		Business
3	Epic	Preparación de los datos	New			Business
	Product Backl...	Selección de los datos	New	3		Business
	Product Backl...	Limpieza de los datos	New	3		Business
	Product Backl...	Construcción de datos	New	5		Business
	Product Backl...	Integración de los datos	New	2		Business

13MBID\_2023\_AG\_2\_3 Team

Taskboard Backlog Capacity Analytics + New Work Item Column Options

Order	Title	State	Assigned To	Rema...
4	Definir plan del proyecto	Done	Carlos Mejia Rodriguez	
	✓ Especificar tareas	Done	Carlos Mejia Rodriguez	
	✓ Especificar tiempos	Done	Carlos Mejia Rodriguez	
	✓ Especificar recursos disponibles	Done	Carlos Mejia Rodriguez	
5	Recolección de datos iniciales	Done	Carlos Mejia Rodriguez	
	✓ Recuperar los archivos de datos	Done	Carlos Mejia Rodriguez	
	✓ Verificar la integridad de los datos	Done	Carlos Mejia Rodriguez	
6	Descripción de los datos	Done	Carlos Mejia Rodriguez	
	✓ Detallar composición de los atributos del dataset	Done		
7	Exploración de los datos	Done	Carlos Mejia Rodriguez	
	✓ Describir metadatos por dataset	Done		
8	Verificación de la calidad de los datos	Done	Carlos Mejia Rodriguez	
	✓ verificación de datos nulos o faltantes	Done		
	✓ verificación de cumplimiento de reglas de formateo	Done		
	✓ verificación de rango de valores	Done		

Vues

## En este punto inicia el desarrollo de la AG3.

- **Verificación de valores nulos**

En los datasets datos\_inscripcion y datos\_cursado no se han encontrado filas nulas.

En el dataset datos\_academicos los siguientes atributos presentan valores nulos:

- ✓ fecha\_ultimo\_examen - **406** filas nulas
- ✓ anio\_ultima\_reinscripcion - **206** filas nulas

En este caso se ha definido con los expertos en el dominio no proceder con la eliminación de estas filas dado que son indicadores de no haber rendido un examen final y / o no haber manifestado la voluntad de continuar cursando por parte del estudiante.

- **Limpieza de los datos**

Se han realizado las siguientes operaciones de eliminación de filas:

- ✓ [a] datos\_inscripcion
- ✓ [b] datos\_cursado
- ✓ [c] datos\_academicos

Dataset	Atributo / Columna	Observaciones / Resultados
[a]	Se eliminan las filas que no cumplen con las condiciones de la regla de verificación de fechas de ingreso / inscripción.  Atributo utilizado: regla_fechas_ingreso Valor de filtro: 'err'	El dataset queda conformado por: 1286 filas  Total original: 1332  Diferencia: <b>-46 (3,45%)</b>
[b]	Se eliminan las filas que no cumplen con las condiciones de la regla de verificación del estado de inscripción de la persona y su condición como estudiante.  Atributo utilizado: regla_estado_inscripcion Valor del filtro: 'err'	El dataset queda conformado por: 1320 filas  Total original: 1332  Diferencia: <b>-12 (0,9%)</b>
[c]	Se eliminan las filas que no cumplen con las condiciones de la regla de verificación de la calidad de los estudiantes con respecto al resto de los ítems de análisis.  Atributo utilizado: regla_verificacion_calidad Valor del filtro: 'err'	El dataset queda conformado por: 618 filas  Total original: 815  Diferencia: <b>-197 (24,17%)</b>



- **Construcción de datos**

Se ha generado el siguiente atributo:

- ✓ **“pct\_avance\_carrera”** [datos\_academicos]: se calcula con la siguiente fórmula:
  - actividades\_aprobadas / total\_actividades
- ✓ **“anios\_transcurridos”** [datos\_academicos]: se crea una columna que represente la diferencia en años entre la fecha de ingreso y la última reinscripción.

- **Integración de los datos**

Se comienza por la integración de los tres datasets originales, con los cambios que se hayan registrado hasta el momento.

Resultados obtenidos:

- ✓ Datos de inscripciones: **1286** - Datos de cursado: **1320** - Coincidencias entre inscripciones y cursado: **1275**
- ✓ Datos completos: **582** integrando **todas las fuentes**

Se observa que se han perdido en la unión una gran cantidad de filas al realizar la integración con el dataset de datos\_academicos. Se recomienda elevar estos resultados y consultar sobre la situación a los expertos en la gestión de los sistemas base.

El dataset final tiene las siguientes dimensiones: **24 columnas y 582 filas**

Registro de **meta-datos** del dataset completo:

Dataset	Columna / Atributo	Tipo de datos	Observaciones											
completo	id_estudiante	string	Código único, ejemplos del formato: CB-015876-0											
	propuesta	int	<table><tr><th colspan="3">TOP CATEGORIES</th></tr><tr><td>134</td><td>426</td><td>73%</td></tr><tr><td>130</td><td>156</td><td>27%</td></tr></table>	TOP CATEGORIES			134	426	73%	130	156	27%		
	TOP CATEGORIES													
	134	426	73%											
	130	156	27%											
estado_inscripcion	string	<table><tr><th colspan="3">TOP CATEGORIES</th></tr><tr><td>Pendiente</td><td>331</td><td>57%</td></tr><tr><td>Aceptado</td><td>251</td><td>43%</td></tr></table>	TOP CATEGORIES			Pendiente	331	57%	Aceptado	251	43%			
TOP CATEGORIES														
Pendiente	331	57%												
Aceptado	251	43%												
fecha_ingreso	string	<table><tr><th colspan="3">TOP CATEGORIES</th></tr><tr><td>01/04/2020</td><td>582</td><td>100%</td></tr></table>	TOP CATEGORIES			01/04/2020	582	100%						
TOP CATEGORIES														
01/04/2020	582	100%												
fecha_inscripcion	string	<p>Fecha en formato: 4/12/2019</p> <table><thead><tr><th>Fecha</th><th>Porcentaje</th></tr></thead><tbody><tr><td>04/02/2020</td><td>~10%</td></tr><tr><td>18/12/2019</td><td>~10%</td></tr><tr><td>11/12/2019</td><td>~5%</td></tr><tr><td>17/12/2019</td><td>~5%</td></tr><tr><td>(Other)</td><td>~65%</td></tr></tbody></table>	Fecha	Porcentaje	04/02/2020	~10%	18/12/2019	~10%	11/12/2019	~5%	17/12/2019	~5%	(Other)	~65%
Fecha	Porcentaje													
04/02/2020	~10%													
18/12/2019	~10%													
11/12/2019	~5%													
17/12/2019	~5%													
(Other)	~65%													

	ingreso_aprobadas	int	<table><tr><th colspan="3">TOP CATEGORIES</th></tr><tr><td>1</td><td>296</td><td>51%</td></tr><tr><td>2</td><td>204</td><td>35%</td></tr><tr><td>0</td><td>82</td><td>14%</td></tr></table>	TOP CATEGORIES			1	296	51%	2	204	35%	0	82	14%																				
TOP CATEGORIES																																			
1	296	51%																																	
2	204	35%																																	
0	82	14%																																	
	ingreso_libres	int	<table><tr><th colspan="3">TOP CATEGORIES</th></tr><tr><td>1</td><td>296</td><td>51%</td></tr><tr><td>0</td><td>204</td><td>35%</td></tr><tr><td>2</td><td>82</td><td>14%</td></tr></table>	TOP CATEGORIES			1	296	51%	0	204	35%	2	82	14%																				
TOP CATEGORIES																																			
1	296	51%																																	
0	204	35%																																	
2	82	14%																																	
	ingreso_totales	int	<table><tr><th colspan="3">TOP CATEGORIES</th></tr><tr><td>2</td><td>582</td><td>100%</td></tr></table>	TOP CATEGORIES			2	582	100%																										
TOP CATEGORIES																																			
2	582	100%																																	
	cursadas_aprobadas	int	<table><tr><th colspan="3">TOP CATEGORIES</th></tr><tr><td>0</td><td>480</td><td>82%</td></tr><tr><td>1</td><td>86</td><td>15%</td></tr><tr><td>2</td><td>13</td><td>2%</td></tr><tr><td>3</td><td>3</td><td>&lt;1%</td></tr></table>	TOP CATEGORIES			0	480	82%	1	86	15%	2	13	2%	3	3	<1%																	
TOP CATEGORIES																																			
0	480	82%																																	
1	86	15%																																	
2	13	2%																																	
3	3	<1%																																	
	cursadas_regulares	int	<table><tr><th colspan="3">TOP CATEGORIES</th></tr><tr><td>0</td><td>454</td><td>78%</td></tr><tr><td>1</td><td>96</td><td>16%</td></tr><tr><td>2</td><td>31</td><td>5%</td></tr><tr><td>3</td><td>1</td><td>&lt;1%</td></tr></table>	TOP CATEGORIES			0	454	78%	1	96	16%	2	31	5%	3	1	<1%																	
TOP CATEGORIES																																			
0	454	78%																																	
1	96	16%																																	
2	31	5%																																	
3	1	<1%																																	
	cursadas_libres	int	<table><tr><th colspan="3">TOP CATEGORIES</th></tr><tr><td>0</td><td>342</td><td>59%</td></tr><tr><td>2</td><td>111</td><td>19%</td></tr><tr><td>1</td><td>69</td><td>12%</td></tr><tr><td>3</td><td>60</td><td>10%</td></tr></table>	TOP CATEGORIES			0	342	59%	2	111	19%	1	69	12%	3	60	10%																	
TOP CATEGORIES																																			
0	342	59%																																	
2	111	19%																																	
1	69	12%																																	
3	60	10%																																	
	cursadas_totales	int	<table><tr><th colspan="3">TOP CATEGORIES</th></tr><tr><td>3</td><td>582</td><td>100%</td></tr></table>	TOP CATEGORIES			3	582	100%																										
TOP CATEGORIES																																			
3	582	100%																																	
	inscripciones_examenes	int	<table><tr><th colspan="3">TOP CATEGORIES</th></tr><tr><td>0</td><td>571</td><td>98%</td></tr><tr><td>1</td><td>11</td><td>2%</td></tr></table>	TOP CATEGORIES			0	571	98%	1	11	2%																							
TOP CATEGORIES																																			
0	571	98%																																	
1	11	2%																																	
	examenes_aprobados	int	<table><tr><th colspan="3">TOP CATEGORIES</th></tr><tr><td>0</td><td>577</td><td>&gt;99%</td></tr><tr><td>1</td><td>5</td><td>&lt;1%</td></tr></table>	TOP CATEGORIES			0	577	>99%	1	5	<1%																							
TOP CATEGORIES																																			
0	577	>99%																																	
1	5	<1%																																	
	anio_ingreso	int	<table><tr><th colspan="3">TOP CATEGORIES</th></tr><tr><td>2020</td><td>582</td><td>100%</td></tr></table>	TOP CATEGORIES			2020	582	100%																										
TOP CATEGORIES																																			
2020	582	100%																																	
	fecha_ultimo_examen	string	<table><tr><td>18/08/2021</td><td>20/08/2021</td><td>23/08/2021</td><td>19/08/2021</td><td>(Other)</td></tr></table>	18/08/2021	20/08/2021	23/08/2021	19/08/2021	(Other)																											
18/08/2021	20/08/2021	23/08/2021	19/08/2021	(Other)																															
	anio_ultima_reinscripcion	float	<table><tr><th colspan="3">TOP CATEGORIES</th></tr><tr><td>2021.0</td><td>378</td><td>100%</td></tr></table>	TOP CATEGORIES			2021.0	378	100%																										
TOP CATEGORIES																																			
2021.0	378	100%																																	
	promedio_sin_aplazos	float	<table><tr><td>MAX</td><td>10.0</td><td>RANGE</td><td>10.0</td></tr><tr><td>95%</td><td>8.5</td><td>IQR</td><td>2.00</td></tr><tr><td>Q3</td><td>7.5</td><td>STD</td><td>2.86</td></tr><tr><td>MEDIAN</td><td>6.8</td><td>VAR</td><td>8.17</td></tr><tr><td>AVG</td><td>5.7</td><td></td><td></td></tr><tr><td>Q1</td><td>5.5</td><td>KURT.</td><td>0.135</td></tr><tr><td>5%</td><td>0.0</td><td>SKEW</td><td>-1.25</td></tr><tr><td>MIN</td><td>0.0</td><td>SUM</td><td>3,326</td></tr></table>	MAX	10.0	RANGE	10.0	95%	8.5	IQR	2.00	Q3	7.5	STD	2.86	MEDIAN	6.8	VAR	8.17	AVG	5.7			Q1	5.5	KURT.	0.135	5%	0.0	SKEW	-1.25	MIN	0.0	SUM	3,326
MAX	10.0	RANGE	10.0																																
95%	8.5	IQR	2.00																																
Q3	7.5	STD	2.86																																
MEDIAN	6.8	VAR	8.17																																
AVG	5.7																																		
Q1	5.5	KURT.	0.135																																
5%	0.0	SKEW	-1.25																																
MIN	0.0	SUM	3,326																																

	promedio_con_aplazos	float	<table><tr><td>MAX</td><td>10.0</td><td></td></tr><tr><td>95%</td><td>8.5</td><td></td></tr><tr><td>Q3</td><td>7.7</td><td></td></tr><tr><td>MEDIAN</td><td>7.0</td><td></td></tr><tr><td>AVG</td><td>6.0</td><td></td></tr><tr><td>Q1</td><td>6.0</td><td></td></tr><tr><td>5%</td><td>0.0</td><td></td></tr><tr><td>MIN</td><td>0.0</td><td></td></tr></table> <table><tr><td>RANGE</td><td>10.0</td></tr><tr><td>IQR</td><td>1.67</td></tr><tr><td>STD</td><td>2.92</td></tr><tr><td>VAR</td><td>8.52</td></tr><tr><td>KURT.</td><td>0.403</td></tr><tr><td>SKEW</td><td>-1.42</td></tr><tr><td>SUM</td><td>3,488</td></tr></table>	MAX	10.0		95%	8.5		Q3	7.7		MEDIAN	7.0		AVG	6.0		Q1	6.0		5%	0.0		MIN	0.0		RANGE	10.0	IQR	1.67	STD	2.92	VAR	8.52	KURT.	0.403	SKEW	-1.42	SUM	3,488
MAX	10.0																																								
95%	8.5																																								
Q3	7.7																																								
MEDIAN	7.0																																								
AVG	6.0																																								
Q1	6.0																																								
5%	0.0																																								
MIN	0.0																																								
RANGE	10.0																																								
IQR	1.67																																								
STD	2.92																																								
VAR	8.52																																								
KURT.	0.403																																								
SKEW	-1.42																																								
SUM	3,488																																								
	actividades_aprobadas	int	<table><tr><td>MAX</td><td>38.0</td><td></td></tr><tr><td>95%</td><td>14.0</td><td></td></tr><tr><td>Q3</td><td>5.0</td><td></td></tr><tr><td>AVG</td><td>4.3</td><td></td></tr><tr><td>MEDIAN</td><td>3.0</td><td></td></tr><tr><td>Q1</td><td>1.0</td><td></td></tr><tr><td>5%</td><td>0.0</td><td></td></tr><tr><td>MIN</td><td>0.0</td><td></td></tr></table> <table><tr><td>RANGE</td><td>38.0</td></tr><tr><td>IQR</td><td>4.00</td></tr><tr><td>STD</td><td>5.56</td></tr><tr><td>VAR</td><td>31.0</td></tr><tr><td>KURT.</td><td>9.90</td></tr><tr><td>SKEW</td><td>2.74</td></tr><tr><td>SUM</td><td>2,515</td></tr></table>	MAX	38.0		95%	14.0		Q3	5.0		AVG	4.3		MEDIAN	3.0		Q1	1.0		5%	0.0		MIN	0.0		RANGE	38.0	IQR	4.00	STD	5.56	VAR	31.0	KURT.	9.90	SKEW	2.74	SUM	2,515
MAX	38.0																																								
95%	14.0																																								
Q3	5.0																																								
AVG	4.3																																								
MEDIAN	3.0																																								
Q1	1.0																																								
5%	0.0																																								
MIN	0.0																																								
RANGE	38.0																																								
IQR	4.00																																								
STD	5.56																																								
VAR	31.0																																								
KURT.	9.90																																								
SKEW	2.74																																								
SUM	2,515																																								
	total_actividades	int	<table><tr><th colspan="3">TOP CATEGORIES</th></tr><tr><td>40</td><td>426</td><td>73%</td></tr><tr><td>42</td><td>156</td><td>27%</td></tr></table>	TOP CATEGORIES			40	426	73%	42	156	27%																													
TOP CATEGORIES																																									
40	426	73%																																							
42	156	27%																																							
	regular	string	<table><tr><th colspan="3">TOP CATEGORIES</th></tr><tr><td>S</td><td>378</td><td>65%</td></tr><tr><td>N</td><td>204</td><td>35%</td></tr></table>	TOP CATEGORIES			S	378	65%	N	204	35%																													
TOP CATEGORIES																																									
S	378	65%																																							
N	204	35%																																							
	calidad	string	<table><tr><th colspan="3">TOP CATEGORIES</th></tr><tr><td>A</td><td>378</td><td>65%</td></tr><tr><td>P</td><td>204</td><td>35%</td></tr></table>	TOP CATEGORIES			A	378	65%	P	204	35%																													
TOP CATEGORIES																																									
A	378	65%																																							
P	204	35%																																							
	segundo_anio	string	<table><tr><th colspan="3">TOP CATEGORIES</th></tr><tr><td>N</td><td>406</td><td>70%</td></tr><tr><td>S</td><td>176</td><td>30%</td></tr></table>	TOP CATEGORIES			N	406	70%	S	176	30%																													
TOP CATEGORIES																																									
N	406	70%																																							
S	176	30%																																							

- Formateo de los datos

Esta operación va a ser realizada en forma previa al inicio del proceso de generación de modelos sobre los datos disponibles. Será documentada oportunamente.

**Figura 6 Product backlog correspondientes a la finalización de la 1ra. Iteración (Sprint 1 en la herramienta)**

Work Item Type	Title	State	Effort
Epic	Compreñsion del negocio	New	
Product Backl...	Determinar los objetivos de la organización	Done	1
Product Backl...	Evaluación de la situación	Done	2
Product Backl...	Determinar los objetivos del Proyecto	Done	3
Product Backl...	Definir plan del proyecto	Done	8
Epic	Compreñsion de los Datos	New	
Product Backl...	Recolección de datos iniciales	Done	2
Product Backl...	Descripción de los datos	Done	2
Product Backl...	Exploración de los datos	Done	3
Product Backl...	Verificación de la calidad de los datos	Done	8
Epic	Preparación de los datos	New	
Product Backl...	Selección de los datos	Done	3
Product Backl...	Limpieza de los datos	Done	3
Product Backl...	Construcción de datos	Done	5
Product Backl...	Integración de los datos	Done	2
Product Backl...	Formateo de datos	Done	3

**Figura 7. Sprint backlog de la 2da. Iteración (Sprint 2 en la herramienta)**

4	Epic	Modelado	New	3	Business
	Product Backl...	Selección de las técnicas de modelado	New		Business
	Task	Planteo inicial de técnicas	To Do		
	Task	Selección de técnicas aplicables	To Do		
	Product Backl...	Generación de un plan de pruebas	New	2	Business
	Task	Definir tipo y cantidad de pruebas	To Do		
	Product Backl...	Construcción del modelo	New	8	Business
	Task	Configuración de parámetros	To Do		
	Task	Generación de los modelos	To Do		
	Task	Descripción de los modelos	To Do		
	Product Backl...	Evaluación del modelo	New	5	Business
	Task	Evaluación de efectividad del Modelo	To Do		

5	Epic	▼  Evaluación	● New		Business
	Product Backl...	▼  Evaluación de los resultados	... ● New	3	Business
	Task	Evaluar resultados con los criterios del cliente	● To Do		
	Task	Aprobar o descartar modelos	● To Do		
	Product Backl...	▼  proceso de revisión	● New	3	Business
	Task	Revisión y definición de mejoras aplicables al proceso	● To Do		
	Product Backl...	▼  Determinación de futuras líneas de trabajo	● New	3	Business
	Task	Listar acciones de mejoras aplicables	● To Do		
	Task	Determinar acciones en próximas iteraciones	● To Do		
6	Epic	▼  Despliegue	● New		Business
	Product Backl...	▼  Plan de Implantación	● New	2	Business
	Task	Determinar esquema de implementación	● To Do		
	Product Backl...	▼  Supervisión y mantenimiento	● New	1	Business
	Task	Plan de actualización y mantenimiento	● To Do		
	Task	Determinación y aplicación de acciones de monitoreo	● To Do		
	Product Backl...	▼  Redacción del Informe final	● New	5	Business
	Task	Confección del reporte final	● To Do		
	Task	Presentación del reporte	● To Do		
	Product Backl...	▼  Revisión del proyecto	● New	2	Business
	Task	Documentación del conocimiento adquirido	● To Do		

## [D] Fase de modelado

- Selección de la técnica de modelado

Con base en los objetivos planteados al inicio del proyecto, las técnicas a utilizar para generar el/los modelos que conformarán el producto de datos final son:

- ✓ Árboles de decisión
- ✓ Redes neuronales
- ✓ Métodos de ensamblado de modelos
- ✓ Redes bayesianas

Además, se van a considerar técnicas o métodos de gestión de parámetros para las técnicas involucradas, así como también para la fase de evaluación de los resultados de cada modelo.

- Adaptación de los datos

Una vez iniciado el trabajo de la fase de Modelado se ha determinado realizar las siguientes modificaciones sobre el dataset disponible:

Operación	Atributo/s	Descripción / Observaciones
Generación de un atributo	pct_avance_ingreso	Marca el % de avance logrado en las asignaturas de ingreso a la carrera.  Fórmula: - ingreso_aprobadas / ingreso_totales
Generación de un atributo	pct_avance_semestre	Marca el % de asignaturas aprobadas en el cursado del 1er semestre de la carrera.  Fórmula: - (cursadas_aprobadas + cursadas_regulares) / cursadas_totales
Generación de un atributo	pct_avance_carrera	Marca el porcentaje de avance en la carrera.  Fórmula: actividades_aprobadas / row.total_actividades
Generación de un atributo	exámenes_1er_semestre	Marca los movimientos con respecto a exámenes.  Fórmula:

		<p>If inscripciones_exámenes &gt; 0 and exámenes_aprobados &gt; 0:     'A'</p> <p>elif inscripciones_exámenes &gt; 0:     'I'</p> <p>else:     'N'</p>
Generación de un atributo	rango_promedios	<p>Marca el promedio académico general</p> <p>Fórmula: promedio_con_aplazos</p> <p>0.0-5.0: 'Bajo'</p> <p>5.0, 7.0: 'Medio'</p> <p>7.0, 10.0: 'Alto'</p>
Eliminación de atributos	<p>'actividades_aprobadas'</p> <p>'total_actividades'</p> <p>'ingreso_aprobadas'</p> <p>'ingreso_libres'</p> <p>'ingreso_totales'</p> <p>'cursadas_aprobadas'</p> <p>'cursadas_regulares'</p> <p>'cursadas_libres'</p> <p>'cursadas_totales'</p> <p>'fecha_ingreso'</p> <p>'fecha_inscripcion'</p> <p>'anio_ingreso'</p> <p>'inscripciones_exámenes'</p> <p>'exámenes_aprobados'</p> <p>'promedio_sin_aplazos'</p> <p>'promedio_con_aplazos'</p> <p>'id_estudiante'</p>	<p>Se han eliminado los atributos utilizados para la generación de las columnas nuevas mencionadas en esta misma tabla.</p> <p>Además de otros atributos que presentan valores constantes y/o no son requeridos para el tipo de análisis que se desea realizar.</p>
Transformación de atributo	estado_inscripcion	<p>Sobre el atributo 'estado_inscripcion' se han transformado sus valores según el siguiente criterio:</p> <ul style="list-style-type: none"> <li>• "Pendiente" = "P"</li> <li>• "Aceptada" = "A"</li> </ul>
Generación de un atributo	rindio_examen	<p>obtiene sus valores a partir del atributo 'fecha_ultimo_examen':</p> <p>Si es una fecha nula = "N"</p> <p>Si es una fecha concreta = "S"</p> <p>indicando que sí ha rendido exámenes</p>

Generación de un atributo	inscripto_ult_ciclo	<p>obtiene sus valores a partir del atributo</p> <p>'anio_ultima_reinscripcion':</p> <p>Si el valor es 2021 = "S"</p> <p>En caso de valor diferente o nulo = "N" indicando que no se ha inscripto a cursar nuevamente</p>
---------------------------	---------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

El nuevo dataset con el que se seguirá trabajando tiene las siguientes dimensiones:

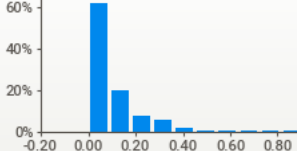
Número de filas: **582**

Número de columnas: **12**

Registro de **meta-datos** del dataset

Dataset	Columna / Atributo	Tipo de datos	Observaciones											
datos_completo filtrados	propuesta	int	<table><tr><th colspan="3">TOP CATEGORIES</th></tr><tr><td>134</td><td>426</td><td>73%</td></tr><tr><td>130</td><td>156</td><td>27%</td></tr></table>	TOP CATEGORIES			134	426	73%	130	156	27%		
	TOP CATEGORIES													
	134	426	73%											
	130	156	27%											
	estado_inscripcion	string	<table><tr><th colspan="3">TOP CATEGORIES</th></tr><tr><td>Pendiente</td><td>331</td><td>57%</td></tr><tr><td>Aceptado</td><td>251</td><td>43%</td></tr></table>	TOP CATEGORIES			Pendiente	331	57%	Aceptado	251	43%		
	TOP CATEGORIES													
	Pendiente	331	57%											
	Aceptado	251	43%											
fecha_ultimo_examen	fecha	<div>Formato de fecha: dd/mm/aaaa</div> <div></div>												
anio_ultima_reinscripcion	int	<table><tr><th colspan="3">TOP CATEGORIES</th></tr><tr><td>2021.0</td><td>378</td><td>100%</td></tr></table>	TOP CATEGORIES			2021.0	378	100%						
TOP CATEGORIES														
2021.0	378	100%												
regular	string	<table><tr><th colspan="3">TOP CATEGORIES</th></tr><tr><td>S</td><td>378</td><td>65%</td></tr><tr><td>N</td><td>204</td><td>35%</td></tr></table>	TOP CATEGORIES			S	378	65%	N	204	35%			
TOP CATEGORIES														
S	378	65%												
N	204	35%												
calidad	string	<table><tr><th colspan="3">TOP CATEGORIES</th></tr><tr><td>A</td><td>378</td><td>65%</td></tr><tr><td>P</td><td>204</td><td>35%</td></tr></table>	TOP CATEGORIES			A	378	65%	P	204	35%			
TOP CATEGORIES														
A	378	65%												
P	204	35%												
segundo_anio	string	<table><tr><th colspan="3">TOP CATEGORIES</th></tr><tr><td>N</td><td>406</td><td>70%</td></tr><tr><td>S</td><td>176</td><td>30%</td></tr></table>	TOP CATEGORIES			N	406	70%	S	176	30%			
TOP CATEGORIES														
N	406	70%												
S	176	30%												
rango_promedios	string	<table><tr><th colspan="3">TOP CATEGORIES</th></tr><tr><td>Alto</td><td>275</td><td>47%</td></tr><tr><td>Medio</td><td>201</td><td>35%</td></tr><tr><td>Bajo</td><td>106</td><td>18%</td></tr></table>	TOP CATEGORIES			Alto	275	47%	Medio	201	35%	Bajo	106	18%
TOP CATEGORIES														
Alto	275	47%												
Medio	201	35%												
Bajo	106	18%												



	exámenes_1er_semestre	string	<table><tr><th colspan="3">TOP CATEGORIES</th></tr><tr><td>N</td><td>571</td><td>98%</td></tr><tr><td>I</td><td>6</td><td>1%</td></tr><tr><td>A</td><td>5</td><td>&lt;1%</td></tr></table>	TOP CATEGORIES			N	571	98%	I	6	1%	A	5	<1%			
TOP CATEGORIES																		
N	571	98%																
I	6	1%																
A	5	<1%																
	avance_ingreso	float	<table><tr><th colspan="3">TOP CATEGORIES</th></tr><tr><td>0.5</td><td>296</td><td>51%</td></tr><tr><td>1.0</td><td>204</td><td>35%</td></tr><tr><td>0.0</td><td>82</td><td>14%</td></tr></table>	TOP CATEGORIES			0.5	296	51%	1.0	204	35%	0.0	82	14%			
TOP CATEGORIES																		
0.5	296	51%																
1.0	204	35%																
0.0	82	14%																
	avance_1er_semestre	float	<table><tr><th colspan="3">TOP CATEGORIES</th></tr><tr><td>0.0</td><td>391</td><td>67%</td></tr><tr><td>0.3333333333333333</td><td>111</td><td>19%</td></tr><tr><td>0.6666666666666666</td><td>69</td><td>12%</td></tr><tr><td>1.0</td><td>11</td><td>2%</td></tr></table>	TOP CATEGORIES			0.0	391	67%	0.3333333333333333	111	19%	0.6666666666666666	69	12%	1.0	11	2%
TOP CATEGORIES																		
0.0	391	67%																
0.3333333333333333	111	19%																
0.6666666666666666	69	12%																
1.0	11	2%																
	avance_carrera	float																

- Generación del plan de pruebas

En primer lugar, a nivel de distribución de las filas del dataset procesado hasta este punto se ha optado por trabajar de la siguiente manera:

- ✓ Se utilizará un **75%** de los datos para tareas de entrenamiento.
- ✓ Se utilizará el restante **25%** para tareas de testeo o evaluación de modelos.

En segunda instancia, se pasará a realizar una experimentación según los siguientes criterios:

- Para cada técnica a emplear se documentarán sus parámetros de entrenamiento con el dataset disponible y, una vez obtenidos los resultados, se registrará la efectividad de su clasificación sobre el dataset de testeo.
- Esto será repetido un mínimo de tres (3) iteraciones a través de las cuales se irán seleccionando aquellas técnicas con un mejor rendimiento. Al pasar de una etapa a otra se podrán realizar modificaciones en los parámetros para optimizar los resultados obtenidos.

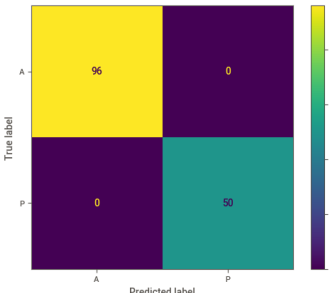
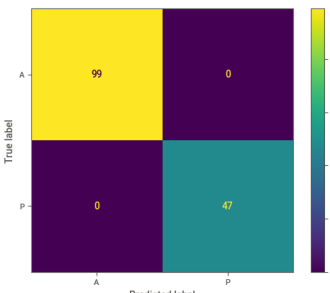
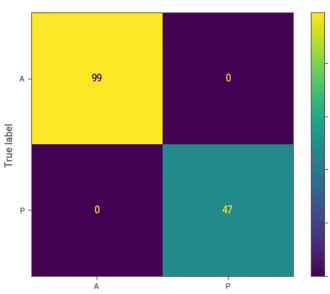
- Construcción del Modelo

En esta oportunidad se van a utilizar librerías implementadas sobre el lenguaje python que proveen diferentes métodos de machine learning para realizar el procesamiento de los datos y la generación de los modelos. El código de estas actividades se encuentra compilado en libretas de jupyter disponibles en la siguiente ubicación: [GitHub](#)

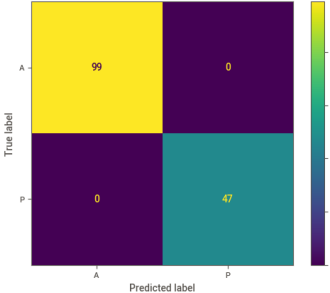
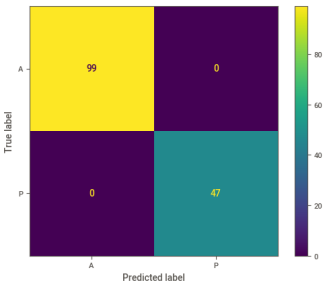
Los resultados de la experimentación serán resumidos en las próximas secciones de este documento.

- Evaluación del modelo

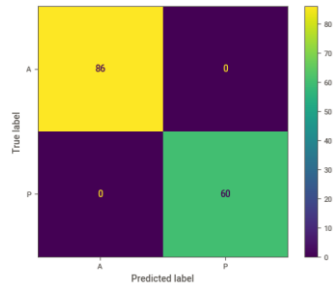
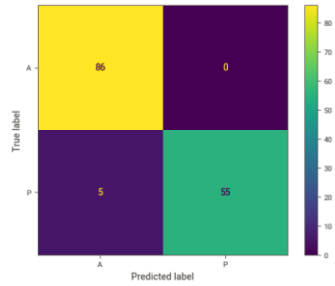
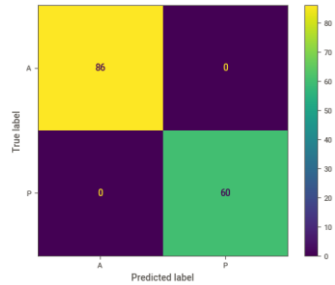
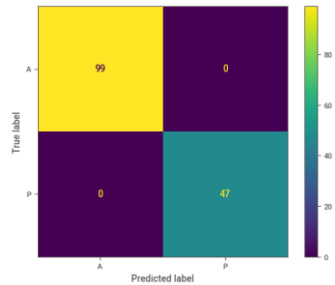
## Prueba #1

Técnica	Parámetros <sup>1</sup>	Resultados
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, <b>solver='liblinear'</b> , tol=0.0001, verbose=0, warm_start=False)	Rendimiento obtenido: 1.0  Matriz de confusión: 
KNeighborsClassifier	KNeighborsClassifier( <b>algorithm='b'</b> , <b>all_tree'</b> , <b>leaf_size=25</b> , metric='minkowski', metric_params=None, n_jobs=None, <b>n_neighbors=50</b> , p=2, weights='uniform')	Rendimiento obtenido: 1.0  Matriz de confusión: 
DecisionTreeClassifier	DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, <b>criterion='entropy'</b> , <b>max_depth=3</b> , max_features=None, max_leaf_nodes=None, min_impurity_decrease=0.0, <b>min_samples_leaf=1</b> , min_samples_split=10, min_weight_fraction_leaf=0.0, random_state=None, splitter='best')	Rendimiento obtenido: 1.0  Matriz de confusión: 
RandomForestClassifier	RandomForestClassifier(bootstrap=True, ccp_alpha=0.0,	Rendimiento obtenido: 1.0

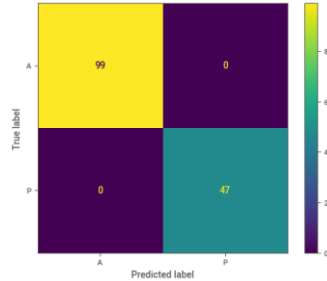
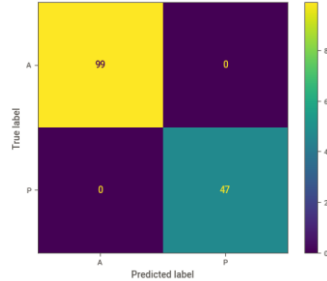
<sup>1</sup> Los nombres y valores de parámetros que se encuentran **destacados** han sido modificados con respecto a su valor por defecto en la ejecución de la técnica.

	<pre> class_weight=None, criterion='gini', max_depth=None, max_features='sqrt', max_leaf_nodes=None, max_samples=None, min_impurity_decrease=0.0, 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) </pre>	<p>Matriz de confusión:</p>  <table border="1"> <thead> <tr> <th></th> <th>Predicted A</th> <th>Predicted P</th> </tr> </thead> <tbody> <tr> <th>True A</th> <td>99</td> <td>0</td> </tr> <tr> <th>True P</th> <td>0</td> <td>47</td> </tr> </tbody> </table>		Predicted A	Predicted P	True A	99	0	True P	0	47
	Predicted A	Predicted P									
True A	99	0									
True P	0	47									
GradientBoostingClassifier	<pre> GradientBoostingClassifier(ccp_alpha=0.0, criterion='friedman_mse', init=None, learning_rate=1.0, loss='log_loss', max_depth=1, max_features=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=100, n_iter_no_change=None, random_state=0, subsample=1.0, tol=0.0001, validation_fraction=0.1, verbose=0, warm_start=False) </pre>	<p>Rendimiento obtenido: 1.0</p> <p>Matriz de confusión:</p>  <table border="1"> <thead> <tr> <th></th> <th>Predicted A</th> <th>Predicted P</th> </tr> </thead> <tbody> <tr> <th>True A</th> <td>99</td> <td>0</td> </tr> <tr> <th>True P</th> <td>0</td> <td>47</td> </tr> </tbody> </table>		Predicted A	Predicted P	True A	99	0	True P	0	47
	Predicted A	Predicted P									
True A	99	0									
True P	0	47									

## Prueba #2

Técnica	Parámetros	Resultados
LogisticRegression	<b>C=1.0</b> , class_weight=None, dual=False, fit_intercept=True, intercept_scaling=1, l1_ratio=None, <b>max_iter=1000</b> , multi_class='auto', n_jobs=None, penalty='l2', random_state=None, <b>solver='lbfgs'</b> , tol=0.0001, verbose=0, warm_start=False	Resultado obtenido: 1.0 Matriz de confusión: 
KNeighborsClassifier	<b>algorithm='kd_tree'</b> , <b>leaf_size=30</b> , metric='minkowski', metric_params=None, n_jobs=None, <b>n_neighbors=10</b> , p=2, weights='uniform'	Resultado obtenido: 1.0 Matriz de confusión: 
DecisionTreeClassifier	<b>n_neighbors=10</b> , <b>algorithm='kd_tree'</b> , <b>leaf_size=30</b>	Resultado obtenido: 1.0 Matriz de confusión: 
RandomForestClassifier	<b>n_estimators=100</b> , <b>max_features='sqrt'</b> , <b>max_depth=None</b>	Resultado obtenido: 1.0 Matriz de confusión: 

### Prueba #3

Técnica	Parámetros	Resultados
KNeighborsClassifier	<b>n_neighbors=5, algorithm='brute', leaf_size=10, weights='distance'</b>	Resultado obtenido: 1.0 Matriz de confusión: 
RandomForestClassifier	<b>n_estimators=100, criterion='gini', max_depth=10, min_samples_split=2, min_samples_leaf=1, max_features='sqrt', random_state=42</b>	Resultado obtenido: 1.0 Matriz de confusión: 

Durante el proceso de prueba de modelos, se emplearon varios modelos y se aplicaron configuraciones específicas para mejorar su rendimiento. Se ajustaron hiperparámetros clave para cada modelo con el objetivo de encontrar combinaciones óptimas.

Para el modelo de regresión logística, se exploraron diferentes solvers y valores de regularización (C) para encontrar la configuración más adecuada.

En el caso del modelo RandomForestClassifier, se ajustaron parámetros como el número de estimadores, la profundidad máxima de los árboles, el tamaño mínimo de muestras para dividir un nodo y la cantidad máxima de características a considerar en cada división.

En resumen, se utilizaron diferentes modelos y se aplicaron configuraciones personalizadas ajustando hiperparámetros logrando en todas las pruebas un rendimiento del 100%.

## [E] Fase de evaluación

- Evaluación de los resultados

En función de los resultados obtenidos en la ejecución del plan de pruebas documentado en la sección anterior se ha seleccionado la técnica **RandomForest** para ser utilizada sobre los datos de estudiantes nuevos. Esto se ha debido a que la efectividad obtenida por la técnica en la fase de entrenamiento y testeo ha sido del **100%**.

- Proceso de revisión

En función de los resultados obtenidos, la técnica mencionada en el paso anterior será utilizada para procesar los datos de los **estudiantes de 2022** y determinar el valor de su atributo “calidad” en base a sus datos almacenados en el sistema de gestión académica. Los resultados de esta predicción se encuentran en el apartado “**Informe final**” del presente documento.

- Determinación de futuras tareas

Como tareas que podrían ser de utilidad para mejorar el rendimiento del modelo de predicción generado se pueden mencionar:

- ✓ Explorar otras técnicas de selección de características e Incrementar el tamaño del conjunto de datos.
- Incorporación de características demográficas y socioeconómicas. Variables como género, edad, etnia, y características socioeconómicas, como el nivel educativo de los padres, el ingreso familiar o el código postal, puede ayudar a capturar factores adicionales que pueden influir en la probabilidad de deserción.

En la **próxima iteración** del proyecto se propone ejecutar las siguientes tareas:

- Verificar cuestiones de integridad referencial en los datos disponibles para incrementar la cantidad de filas utilizadas para el entrenamiento de los modelos de aprendizaje automático.
- Recopilación de más datos: Para aumentar el tamaño del dataset, se debe explorar la posibilidad de obtener datos adicionales de fuentes externas o mediante la recopilación de información directamente de los propios estudiantes. Esto puede incluir encuestas, cuestionarios u otros métodos para obtener datos relevantes que puedan ayudar a mejorar las predicciones.

## [F] Fase de despliegue

- **Plan de implementación**

Las autoridades de la Universidad han determinado que el modelo generado sea utilizado como herramienta de soporte para la definición de políticas de apoyo a los estudiantes que inician sus estudios año a año.

Por tal motivo, se ha dispuesto que el modelo sea actualizado con datos nuevos cada año (*en una fecha a determinar en función de la finalización de la actividad académica de los estudiantes*) y vuelva a ser aplicado a los nuevos estudiantes una vez finalizado su primer semestre de cursado.

- **Supervisión y Mantenimiento**

Una vez que el producto se encuentre en uso por parte de sus usuarios objetivo se podrían realizar las siguientes acciones:

- Monitoreo de las predicciones que se realizan para que sean consistentes con los datos de entrenamiento y testeo a fin de detectar desvíos.
- Contabilizar accesos a la herramienta para obtener métricas de uso.
- Actualización periódica del modelo: A medida que se recopilen nuevos datos y se disponga de más información, es importante realizar actualizaciones periódicas del modelo de predicción.
- Análisis de errores y retroalimentación: Es crucial analizar y comprender los errores del modelo de predicción y recopilar comentarios y retroalimentación de los usuarios.
- Actualización de datos de referencia: Si los datos de referencia utilizados para entrenar y evaluar el modelo están desactualizados o se considera que ya no son representativos de la realidad, se deben realizar esfuerzos para actualizarlos.
- Evaluación de rendimiento continuo: Realizar una evaluación continua del rendimiento del modelo utilizando métricas relevantes, como precisión, sensibilidad, especificidad, entre otras.

- Informe Final

Se presentan los resultados de aplicación del modelo generado con la técnica de mejor rendimiento sobre los datos de estudiantes nuevos para el periodo 2022:

Técnica Seleccionada: **RandomForest**

	Cantidad	Porcentaje
<b>Alumnos Activos</b>	815	84,28%
<b>Alumnos Pasivos</b>	152	15,72%
<b>Total</b>	967	100%

#### Revisión del proyecto

Una vez finalizada la presente iteración de la metodología CRISP-DM se reconocen como mejoras aplicables al proyecto las siguientes:

- Incorporar herramientas que brinden soporte a la comunicación con los expertos en el dominio a fin de solucionar inquietudes del equipo de trabajo con mayor velocidad.
- Documentación del código, es importante documentar y organizar adecuadamente el código, así como los procesos y decisiones tomadas durante el desarrollo del proyecto.
- Identificar oportunidades para automatizar tareas y flujos de trabajo repetitivos puede agilizar el desarrollo y la revisión del proyecto.
- Incorporar pruebas unitarias y de integración en el código del proyecto permite verificar la calidad y funcionalidad del código desarrollado.
- Utilizar herramientas de control de versiones, como Git, y plataformas de alojamiento de repositorios, como GitHub o GitLab, permite mantener un registro y seguimiento de los cambios realizados en el proyecto a lo largo del tiempo.



**Figura 8. Product backlog cierre del 2do sprint**

4	Epic	Modelado	New	3	Business
	Product Backl...	Selección de las técnicas de modelado	Done		Business
	Product Backl...	Generación de un plan de pruebas	Done	2	Business
	Product Backl...	Construcción del modelo	Done	8	Business
	Product Backl...	Evaluación del modelo	Done	5	Business
5	Epic	Evaluación	New		Business
	Product Backl...	Evaluación de los resultados	Done	3	Business
	Product Backl...	proceso de revisión	Done	3	Business
	Product Backl...	Determinación de futuras líneas de trabajo	Done	3	Business
6	Epic	Despliegue	New		Business
	Product Backl...	Plan de Implantación	Done	2	Business
	Product Backl...	Supervisión y mantenimiento	Done	1	Business
	Product Backl...	Redacción del Informe final	Done	5	Business
	Product Backl...	Revisión del proyecto	Done	2	Business

**Figura 9. Sprint backlog del Sprint 2 Completado**

Order	Title	State	Assigned To
1	Selección de las técnicas de modelado	Done	Carlos Mejia Rodriguez
	Planteo inicial de técnicas	Done	Carlos Mejia Rodriguez
	Selección de técnicas aplicables	Done	Carlos Mejia Rodriguez
	Adaptación de los datos	Done	Carlos Mejia Rodriguez
2	Generación de un plan de pruebas	Done	Carlos Mejia Rodriguez
	Definir tipo y cantidad de pruebas	Done	Carlos Mejia Rodriguez
3	Construcción del modelo	Done	Carlos Mejia Rodriguez
	Configuración de parámetros	Done	Carlos Mejia Rodriguez
	Generación de los modelos	Done	Carlos Mejia Rodriguez
	Descripción de los modelos	Done	Carlos Mejia Rodriguez
4	Evaluación del modelo	Done	Carlos Mejia Rodriguez
	Evaluación de efectividad del Modelo	Done	Carlos Mejia Rodriguez
	Determinar items a analizar de la configuración	Done	Carlos Mejia Rodriguez

5	<div> <div> <div></div> <div></div> </div> <div> <div></div> <div></div> </div> </div> <div>Evaluación de los resultados</div> <div> <div> <div></div> <div></div> </div> <div> <div></div> <div></div> </div> </div> <div> <div> <div></div> <div></div> </div> <div> <div></div> <div></div> </div> </div>	<div> <div></div> <div>Done</div> <div>Carlos Mejia Rodriguez</div> </div> <div> <div></div> <div>Done</div> <div>Carlos Mejia Rodriguez</div> </div> <div> <div></div> <div>Done</div> <div>Carlos Mejia Rodriguez</div> </div>
6	<div> <div> <div></div> <div></div> </div> <div> <div></div> <div></div> </div> </div> <div>proceso de revisión</div> <div> <div> <div></div> <div></div> </div> <div> <div></div> <div></div> </div> </div> <div> <div> <div></div> <div></div> </div> <div> <div></div> <div></div> </div> </div>	<div> <div></div> <div>Done</div> <div>Carlos Mejia Rodriguez</div> </div> <div> <div></div> <div>Done</div> <div>Carlos Mejia Rodriguez</div> </div>
7	<div> <div> <div></div> <div></div> </div> <div> <div></div> <div></div> </div> </div> <div>Determinación de futuras líneas de trabajo</div> <div> <div> <div></div> <div></div> </div> <div> <div></div> <div></div> </div> </div> <div> <div> <div></div> <div></div> </div> <div> <div></div> <div></div> </div> </div>	<div> <div></div> <div>Done</div> <div>Carlos Mejia Rodriguez</div> </div> <div> <div></div> <div>Done</div> <div>Carlos Mejia Rodriguez</div> </div> <div> <div></div> <div>Done</div> <div>Carlos Mejia Rodriguez</div> </div>
8	<div> <div> <div></div> <div></div> </div> <div> <div></div> <div></div> </div> </div> <div>Plan de Implantación</div> <div> <div> <div></div> <div></div> </div> <div> <div></div> <div></div> </div> </div> <div> <div> <div></div> <div></div> </div> <div> <div></div> <div></div> </div> </div>	<div> <div></div> <div>Done</div> <div>Carlos Mejia Rodriguez</div> </div> <div> <div></div> <div>Done</div> <div>Carlos Mejia Rodriguez</div> </div>
9	<div> <div> <div></div> <div></div> </div> <div> <div></div> <div></div> </div> </div> <div>Supervisión y mantenimiento</div> <div> <div> <div></div> <div></div> </div> <div> <div></div> <div></div> </div> </div> <div> <div> <div></div> <div></div> </div> <div> <div></div> <div></div> </div> </div>	<div> <div></div> <div>Done</div> <div>Carlos Mejia Rodriguez</div> </div> <div> <div></div> <div>Done</div> <div>Carlos Mejia Rodriguez</div> </div> <div> <div></div> <div>Done</div> <div>Carlos Mejia Rodriguez</div> </div>
10	<div> <div> <div></div> <div></div> </div> <div> <div></div> <div></div> </div> </div> <div>Redacción del Informe final</div> <div> <div> <div></div> <div></div> </div> <div> <div></div> <div></div> </div> </div> <div> <div> <div></div> <div></div> </div> <div> <div></div> <div></div> </div> </div>	<div> <div></div> <div>Done</div> <div>Carlos Mejia Rodriguez</div> </div> <div> <div></div> <div>Done</div> <div>Carlos Mejia Rodriguez</div> </div> <div> <div></div> <div>Done</div> <div>Carlos Mejia Rodriguez</div> </div>
11	<div> <div> <div></div> <div></div> </div> <div> <div></div> <div></div> </div> </div> <div>Revisión del proyecto</div> <div> <div> <div></div> <div></div> </div> <div> <div></div> <div></div> </div> </div>	<div> <div></div> <div>Done</div> <div>Carlos Mejia Rodriguez</div> </div>
	<div> <div> <div></div> <div></div> </div> <div> <div></div> <div></div> </div> </div> <div>Documentación del conocimiento adquirido</div> <div> <div> <div></div> <div></div> </div> <div> <div></div> <div></div> </div> </div>	<div> <div></div> <div>Done</div> <div>Carlos Mejia Rodriguez</div> </div>