

13MBID_10_A Metodologías de Gestión y Diseño de Proyectos Big Data

Actividad 3. Sprint 2 – Preparación. Modelado. Evaluación. Despliegue

Máster en Big Data y Data Science

Profesor: Dr. Horacio Kuna

Alumno: Victor David Betancourt Leal

Fecha: 23 Mayo-2023

Curso 2022 – Ed. Octubre

Contenido

1	Introducción	7
1.1	Objetivo	8
2	Sprint 1: Fase Preparación de los Datos	9
2.1	Selección de Atributos	10
2.2	Verificación de Valores Nulos	11
2.3	Limpieza de los Datos	12
2.4	Construcción de Datos	12
2.5	Integración de los Datos	13
2.5.1	Dataset Datos Completos	13
2.5.2	Meta-Datos (DTale)	13
2.5.3	Meta-Datos (SweetViz)	20
2.6	Formateo de los Datos	20
2.7	Finaliza Fase de Preparación de los Datos	21
3	Sprint 2: Fase Modelado	23
3.1	Selección de la Técnica de Modelado	24
3.1.1	Planteo Inicial de las Técnicas de Modelado	24
3.1.2	Adaptación de los Datos	25
3.1.3	Datos Completos Filtrados	26
3.1.4	Meta-Datos (SweetViz)	27
3.1.5	Selección de Técnicas Aplicables	28
3.2	Generación del Plan de Pruebas	28
3.2.1	Determinar tipo y cantidad de pruebas	28
3.3	Construcción del Modelo	29
3.3.1	Configuración de Parámetros	29
3.3.2	Datos Procesados y Dummies	30
3.3.3	Generación de los Modelos	30
3.3.4	Descripción de los Modelos	31
3.4	Evaluación del Modelo	31
3.4.1	Prueba 1	31
3.4.2	Prueba 2	35

3.5	Finaliza Fase de Modelado	36
4	Sprint 2: Fase Evaluación.....	37
4.1	Evaluación de los Resultados	37
4.2	Proceso de Revisión	37
4.3	Determinación de Futuras Tareas	38
4.4	Finaliza Fase Evaluación	39
5	Sprint 2: Fase de Despliegue	40
5.1	Plan de Implementación	40
5.2	Supervisión y Mantenimiento	40
5.3	Informe Final	41
5.3.1	Datos Nuevos	41
5.3.2	Datasets para Despliegue.....	42
5.3.3	Variables del Despliegue	42
5.3.4	Modelos para Predicción.....	43
5.3.5	Resultados	44
5.3.6	Datasets de Predicciones	45
5.4	Revisión del Proyecto	46
5.5	Finaliza Fase Despliegue.....	47
6	Conclusiones.....	49
6.1	Áreas de Oportunidad	49
6.2	Recomendaciones	49
6.3	Sprint Retrospective	50
7	Glosario	51
8	Bibliografía	52

Índice de Ilustraciones

Ilustración 1. Azure: Finaliza Selección de Atributos	10
Ilustración 2. SweetViz: Datos Completos.....	20
Ilustración 3. Azure: Finaliza Sprint 1 - Vista Backlogs	21
Ilustración 4. Azure: Finaliza Sprint 1 - Vista Sprints Backlog.....	21
Ilustración 5. Azure: Finaliza Sprint 1 - Vista Sprints Taskboard	22
Ilustración 6. Azure: Inicia Sprint 2.....	23
Ilustración 7. SweetViz: Datos Completos Filtrados.....	27
Ilustración 8. Azure: Finaliza Sprint 2 Fase Modelado	36
Ilustración 9. Azure: Finaliza Sprint 2 Fase Evaluación.....	39
Ilustración 10. Azure Backlogs/Backlog: Finaliza Sprint 2 Fase Despliegue	47
Ilustración 11. Azure Sprints/Backlog: Finaliza Sprint 2 Fase Despliegue	47
Ilustración 12. Azure Sprints/Taskboard: Finaliza Sprint 2 Fase Despliegue.....	48

Índice de Tablas

Tabla 1. Limpieza de los Datos	12
Tabla 2. Dataset Datos Completos	13
Tabla 3. Meta-Datos del Dataset Completo	19
Tabla 4. Adaptación de los Datos	26
Tabla 5. Dataset Datos Completos Filtrados	27
Tabla 6. Dataset Datos Procesados	30
Tabla 7. Dataset Datos Dummies	30
Tabla 8. Evaluación Prueba 1	34
Tabla 9. Evaluación Prueba 1 RandomizedSearch	34
Tabla 10. Evaluación Prueba 2	36
Tabla 11. Dataset Datos Nuevos	41
Tabla 12. Datos Nuevos Codificados Matched.....	41
Tabla 13. Data Despliegue.....	42
Tabla 14. Datos Nuevos Codificados Despliegue	42
Tabla 15. Resultados Predicción.....	45
Tabla 16. Datasets Resultados Predicción.....	45

1 Introducción

Las autoridades de una Facultad desean **obtener conocimiento** a partir de los datos disponibles de los alumnos inscritos durante el ciclo lectivo 2020, siendo el reto principal poder predecir con un margen de confianza considerable la situación de los nuevos alumnos inscritos para el periodo 2022.

Para gestionar este proyecto, se ha tomado como base la **Metodología CRISP-DM** que consta de 6 fases:

- 1) Comprensión del Negocio
- 2) Comprensión de los Datos
- 3) Preparación de los Datos
- 4) Modelado
- 5) Evaluación
- 6) Implementación

Adicionalmente, se han incorporado elementos característicos de la Metodología Ágil **SCRUM**, mediante el uso de la tecnología **Azure Boards**.

Las autoridades de una Facultad desean **obtener conocimiento** a partir de los datos disponibles de los alumnos inscritos durante el ciclo lectivo 2020, principalmente en lo que respecta a su situación como estudiante y su grado de avance en la carrera a la que se hayan inscrito. La situación de cada estudiante podrá ser:

- **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.

1.1 Objetivo

El **objetivo** del presente documento consiste en concluir con las Fases pendientes de la **Metodología CRISP-DM** en seguimiento a la Actividad/Seminario 2 de la asignatura, a saber:

- Preparación de los Datos
- Modelado
- Evaluación
- Despliegue/Implantación

Para tal fin, se empleará la herramienta [Azure DevOps](#), así como código en Python, a través de Google Colab. Las tareas corresponderán al **segundo Sprint** del proyecto. Y se dispone los siguientes datasets:

- `datos_inscripciones_mod.csv`
- `datos_cursado_mod.csv`
- `datos_academico_mod.csv`
- `datos_nuevos_22.csv`

Adicionalmente, se ha puesto a disposición el material utilizado a lo largo de este documento, en un repositorio de [GitHub](#).

2 Sprint 1: Fase Preparación de los Datos

Para poder realizar la gestión del proyecto, se empleará la herramienta [Azure DevOps](#), en particular, **Azure Boards**, que cuenta con vistas importantes, tales como:

- **Product Backlog**
- **Sprints Backlog**
- **Sprints Tasks**

En “**Backlogs**” se definirán los siguientes elementos:

- 1) Identificar las historias de usuario **épicas** (*Epics*) y sus dependencias (serán equivalentes a las fases de la metodología **CRISP-DM**).
- 2) Describir brevemente los objetivos y las fechas de trabajo de cada épica.

Dentro de **Azure Boards**, se procederá de la siguiente manera:

- 1) Se creará una nueva historia de usuario (*Issues*) por cada tarea genérica de la metodología CRISP-DM.
- 2) Se vinculará a cada historia de usuario con su correspondiente épica generada previamente.
- 3) Se cargarán las tareas (*Tasks*) necesarias para dar cumplimiento a las fases de la metodología agregadas en el paso anterior, pudiendo ser una o más según se requiera.

La fase de Preparación de los Datos prosigue a la Comprensión de los Datos, y consta de 5 etapas:

- 1) Selección de Atributos
- 2) Limpieza de los Datos
- 3) Construcción de Datos
- 4) Integración de los Datos
- 5) Formateo de los Datos

2.1 Selección de Atributos

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' (*)



Ilustración 1. Azure: Finaliza Selección de Atributos

2.2 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
- Total = 612

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.

2.3 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: 46 (3.45%)
[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: 12 (0.9%)
[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: 197 (24.17%)

Tabla 1. Limpieza de los Datos

2.4 Construcción de Datos

Se ha generado el siguiente atributo relativo al grado de cumplimiento de actividades:

- “pct_avance_carrera” [datos_academicos]: se calcula con la siguiente fórmula:
 - $\text{actividades_aprobadas} / \text{total_actividades}$

2.5 Integración de los Datos

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

2.5.1 Dataset Datos Completos



datos_completos.csv

Tabla 2. Dataset Datos Completos

Resultados obtenidos:

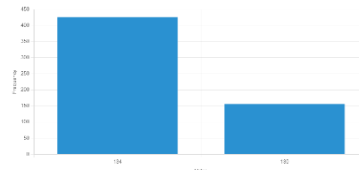


- Datos de inscripciones: **1286**
- Datos de cursado: **1320**
- Coincidencias entre inscripciones y cursado: **1275**
- Filas Datos completos: **582** integrando todas las fuentes
- Columnas Datos completos: **24** integrando todas las fuentes

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.

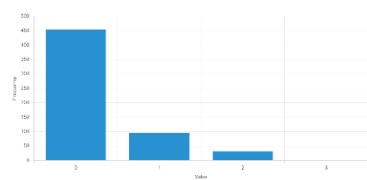
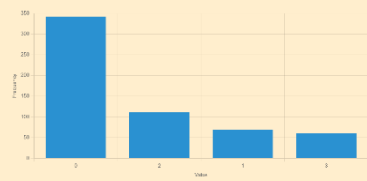
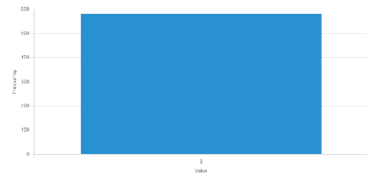
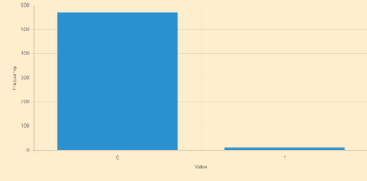
2.5.2 Meta-Datos (DTale)

Registro de **meta-datos** del dataset completo, con ayuda de [Dtale](#):

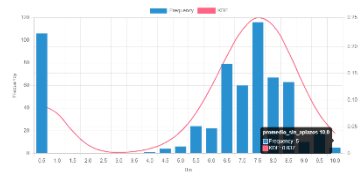
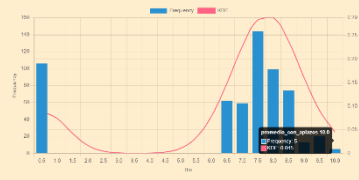
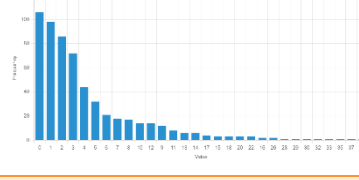
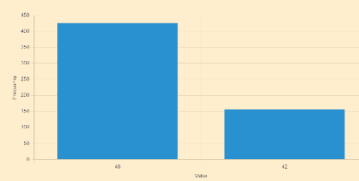
Dataset	Atributo	Tipo de Datos	Observaciones
[completo]	id_estudiante	String: Alfanumérico	<p>Atributo con un formato especial: ##-#####-#</p> <p>Valores de ejemplo: CA-003269-2 (1), CA-005300-2 (1), CA-006491-8 (1), CA-007459-0 (1), CA-008313-1 (1), CA-008316-5 (1), CA-008422-6 (1), ...</p>

			<p>Hace referencia al número de matrícula de cada estudiante.</p> <p>Cantidad de nulos = 0</p>
[completo]	propuesta	N Numérico: Entero	<p>Descripción y distribución de valores (cantidad):</p> <p>134 (426), 130 (156)</p> 
[completo]	estado_inscripcion	String	<p>Descripción y distribución de valores (cantidad):</p> <p>Pendiente (331), Aceptado (251)</p> 
[completo]	fecha_ingreso	String	<p>Descripción y distribución de valores (cantidad):</p> <p>01/04/2020 (582)</p> 
[completo]	fecha_inscripcion	String	<p>Descripción y distribución de valores (cantidad):</p> <p>04/02/2020 (43), 18/12/2019 (39), 11/12/2019 (30), 17/12/2019 (30), 03/12/2019 (27), 13/12/2019 (26), 16/12/2019 (25), ...</p>

[completo]	ingreso_aprobadas	Numérico: Entero	<p>Descripción y distribución de valores (cantidad):</p> <p>1 (296), 2 (204), 0 (82)</p>
[completo]	ingreso_libres	Numérico: Entero	<p>Descripción y distribución de valores (cantidad):</p> <p>1 (296), 0 (204), 2 (82)</p>
[completo]	ingreso_totales	Numérico: Entero	<p>Descripción y distribución de valores (cantidad):</p> <p>2 (582)</p>
[completo]	cursadas_aprobadas	Numérico: Entero	<p>Descripción y distribución de valores (cantidad):</p> <p>0 (480), 1 (86), 2 (13), 3 (3)</p>
[completo]	cursadas_regulares	Numérico: Entero	<p>Descripción y distribución de</p>

			<p>valores (cantidad):</p> <p>0 (454), 1 (96), 2 (31), 3 (1)</p> 
[completo]	cursadas_libres	Numérico: Entero	<p>Descripción y distribución de valores (cantidad):</p> <p>0 (342), 2 (111), 1 (69), 3 (60)</p> 
[completo]	cursadas_totales	Numérico: Entero	<p>Descripción y distribución de valores (cantidad):</p> <p>3 (582)</p> 
[completo]	inscripciones_examenes	Numérico: Entero	<p>Descripción y distribución de valores (cantidad):</p> <p>0 (571), 1 (11)</p> 
[completo]	exámenes_aprobados	Numérico: Entero	<p>Descripción y distribución de valores (cantidad):</p> <p>0 (577), 1 (5)</p>

			
[completo]	anio_ingreso	Numérico: Entero	<p>Descripción y distribución de valores (cantidad):</p> <p>2020 (582)</p> 
[completo]	fecha_ultimo_examen	String	<p>Descripción y distribución de valores (cantidad):</p> <p>18/08/2021 (46), 20/08/2021 (39), 23/08/2021 (34), 19/08/2021 (15), 05/08/2021 (11), 16/03/2021 (10), ...</p> 
[completo]	anio_ultima_reinscripcion	Numérico: Flotante	<p>Descripción y distribución de valores (cantidad):</p> <p>2021 (378)</p> 
[completo]	promedio_sin_aplazos	Numérico: Flotante	<p>Descripción y distribución de valores (cantidad):</p> <p>0 (106), 7 (69), 8 (53), 6 (51), 7.5 (29), 7.33 (17), 9 (17), 6.5 (16), 7.67 (12), 8.5 (8), 5.5 (7), ...</p>

			
[completo]	promedio_con_aplazos	Númérico: Flotante	<p>Descripción y distribución de valores (cantidad):</p> <p>0 (106), 7 (80), 8 (59), 6 (46), 7.5 (36), 7.33 (20), 9 (18), 7.67 (17), 6.5 (16), ...</p> 
[completo]	actividades_aprobadas	Númérico: Entero	<p>Descripción y distribución de valores (cantidad):</p> <p>0 (106), 1 (98), 2 (86), 3 (72), 4 (44), 5 (32), 6 (21), 7 (18), 8 (17), 10 (14), ...</p> 
[completo]	total_actividades	Númérico: Entero	<p>Descripción y distribución de valores (cantidad):</p> <p>40 (426), 42 (156)</p> 
[completo]	regular	String	<p>Descripción y distribución de valores (cantidad):</p> <p>S (378), N (204)</p>

[completo]	calidad	String	<p>Descripción y distribución de valores (cantidad):</p> <p>A (378), P (204)</p>
[completo]	segundo_anio	String	<p>Descripción y distribución de valores (cantidad):</p> <p>N (406), S (176)</p>

Tabla 3. Meta-Datos del Dataset Completo

2.5.3 Meta-Datos (SweetViz)

Adicionalmente se han generado los reportes de los **Datos Completos** utilizando la librería [SweetViz](#).

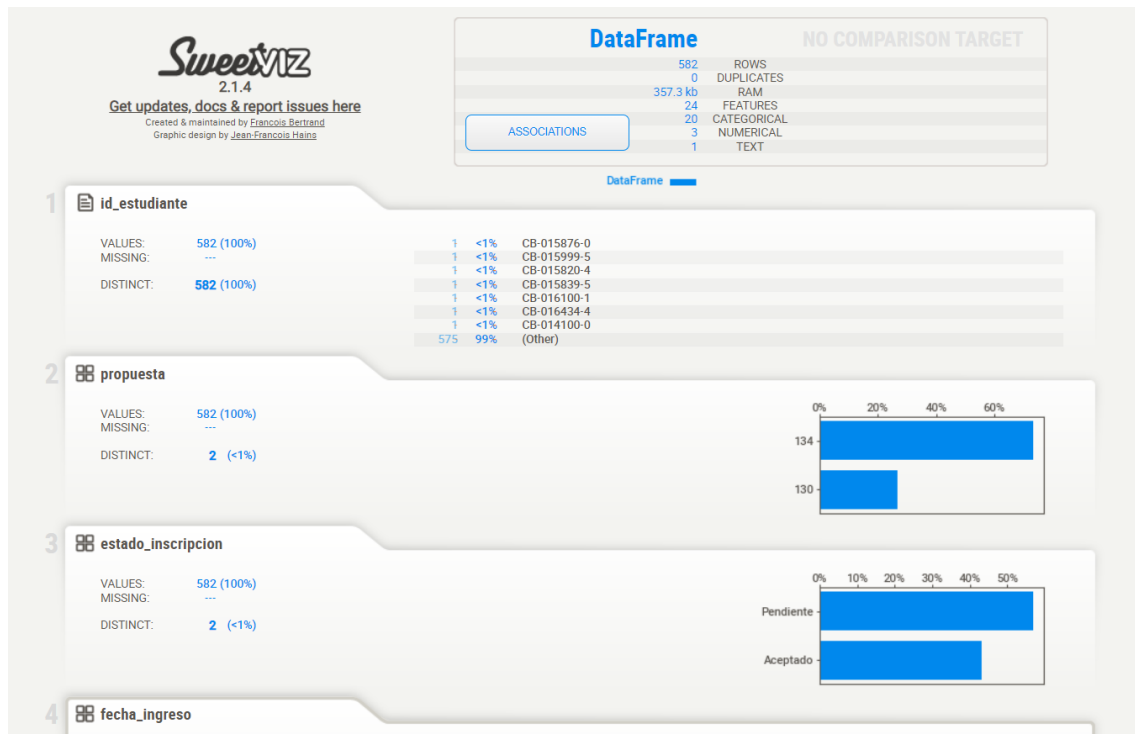


Ilustración 2. SweetViz: Datos Completos

2.6 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.

2.7 Finaliza Fase de Preparación de los Datos

Se registra en **Azure Boards** la finalización de la Fase de Preparación de los Datos.

- **Vista Backlogs / Backlog**

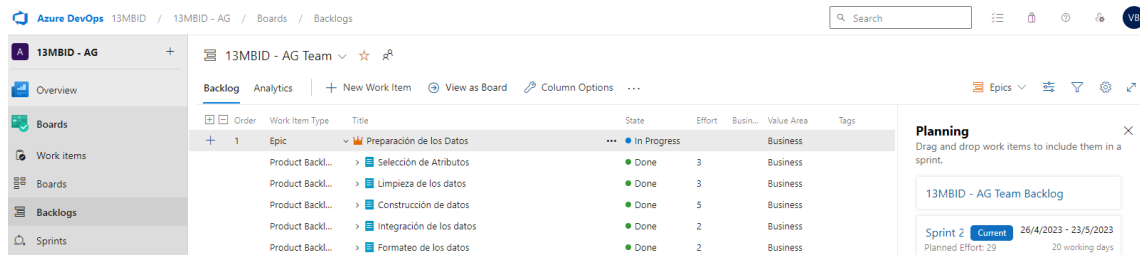


Ilustración 3. Azure: Finaliza Sprint 1 - Vista Backlogs

- **Vista Sprints / Backlog**

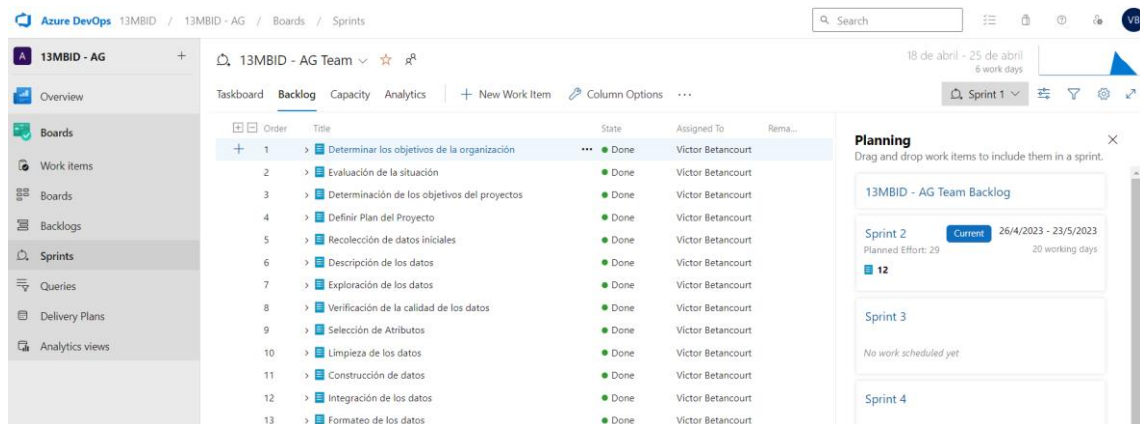


Ilustración 4. Azure: Finaliza Sprint 1 - Vista Sprints Backlog

• Vista Sprints / Taskboard

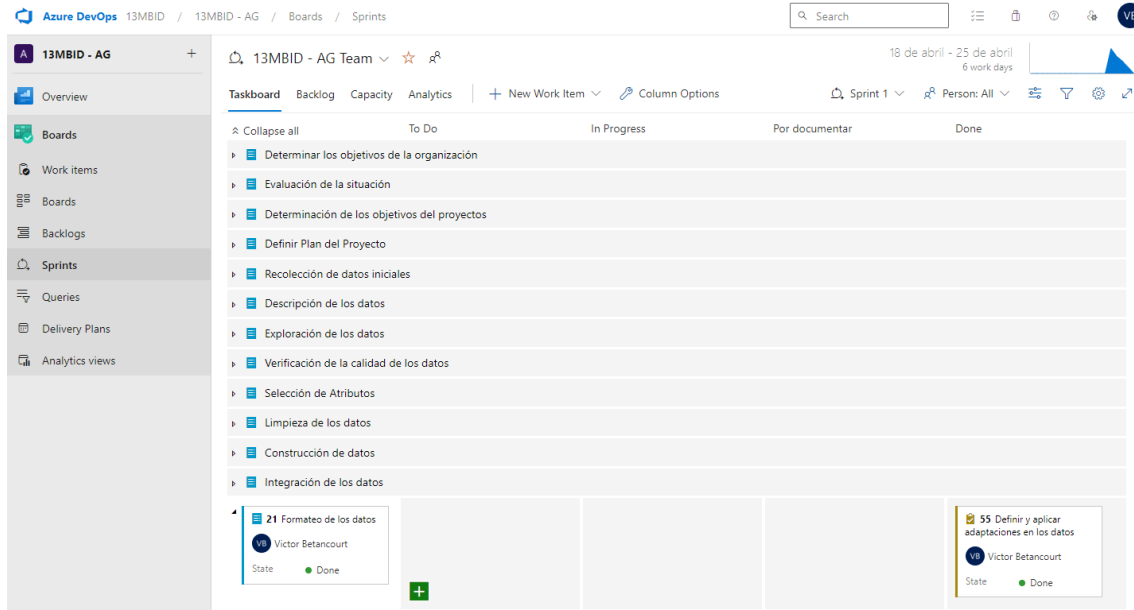


Ilustración 5. Azure: Finaliza Sprint 1 - Vista Sprints Taskboard

3 Sprint 2: Fase Modelado

El [Modelado](#) generalmente se lleva a cabo en múltiples iteraciones. Típicamente, los Científicos de Datos ejecutan varios modelos utilizando los parámetros predeterminados y luego ajustan los parámetros o regresan a la fase de preparación de datos para las manipulaciones requeridas por su modelo de elección.

En el presente proyecto, se han considerado las siguientes etapas para la fase de Modelado:

- 1) Selección de la Técnica de Modelado
- 2) Adaptación de los Datos
- 3) Generación del Plan de Pruebas
- 4) Construcción del Modelo
- 5) Evaluación del Modelo

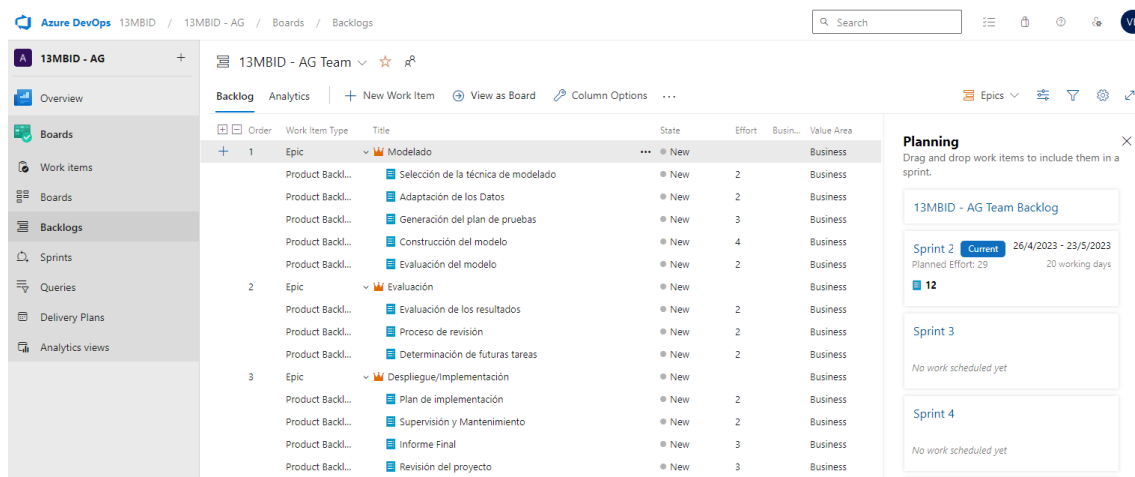


Ilustración 6. Azure: Inicia Sprint 2

3.1 Selección de la Técnica de Modelado

3.1.1 Planteo Inicial de las Técnicas 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:

- Regresión Logística
- K-Nearest Neighbors
- Árboles de Decisión (TDIDT, *Top Down Induction of Decision Trees*)
- Métodos de Ensamblado
 - Random Forest
 - Gradient Boosting
- Red Neuronal
- Support Vector Machine (SVM)

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.

3.1.2 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
Generación de un atributo	pct_avance_ingreso	<p>Marca el % de avance logrado en las asignaturas de ingreso a la carrera.</p> <p>Fórmula:</p> <ul style="list-style-type: none"> - ingreso_aprobadas / ingreso_totales
Generación de un atributo	pct_avance_semestre	<p>Marca el % de asignaturas aprobadas en el cursado del 1er semestre de la carrera.</p> <p>Fórmula:</p> <ul style="list-style-type: none"> - (cursadas_aprobadas + cursadas_regulares) / cursadas_totales
Generación de un atributo	pct_avance_carrera(row): return row.actividades_aprobadas / row.total_actividades	<p>Marca el % de actividades aprobadas en el cursado del 1er semestre de la carrera.</p> <p>Fórmula:</p> <ul style="list-style-type: none"> - actividades_aprobadas / total_actividades
Generación de un atributo	exámenes_1er_semestre(row): if row.inscripciones_exámenes > 0 and row.exámenes_aprobados > 0: return 'A' elif row.inscripciones_exámenes > 0: return 'I' else: return 'N'	<p>Marca la etiqueta sobre exámenes aprobados en el cursado del 1er semestre de la carrera.</p> <p>Fórmula:</p> <ul style="list-style-type: none"> - Si inscripciones_exámenes y exámenes_aprobados son mayor que 0, entonces "A" - Si sólo inscripciones_exámenes es mayor que 0, entonces

		<p>"I"</p> <ul style="list-style-type: none"> - En cualquier otro caso, "N"
Generación de un atributo	rango_promedios	<p>Marca la etiqueta de "Bajo", "Medio" y "Alto" con respecto al atributo "promedio_con_plazos" en el cursado del 1er semestre de la carrera.</p> <p>rangos=[0.0, 5.0, 7.0, 10.0]</p>
Eliminación de atributos	'actividades_aprobadas' 'total_actividades' 'ingreso_aprobadas' 'ingreso_libres' 'ingreso_totales' 'cursadas_aprobadas' 'cursadas_regulares' 'cursadas_libres' 'cursadas_totales' 'fecha_ingreso' 'fecha_inscripcion' 'anio_ingreso' 'inscripciones_exámenes' 'exámenes_aprobados' 'promedio_sin_aplazos' 'promedio_con_aplazos' 'id_estudiante'	<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>

Tabla 4. Adaptación de los Datos

3.1.3 Datos Completos Filtrados

Datos Completos Filtrados, columnas finales:

- 1) 'propuesta',
- 2) 'estado_inscripcion',
- 3) 'fecha_ultimo_examen',
- 4) 'anio_ultima_reinscripcion',
- 5) 'regular',
- 6) 'calidad',
- 7) 'segundo_anio',
- 8) 'rango_promedios',

- 9) 'exámenes_1er_semestre',
- 10) 'avance_ingreso',
- 11) 'avance_1er_semestre',
- 12) 'avance_carrera'



datos_completos_filtrados.csv

Tabla 5. Dataset Datos Completos Filtrados

3.1.4 Meta-Datos (SweetViz)

Se han generado los reportes de los **Datos Completos Filtrados** utilizando la librería [SweetViz](#).

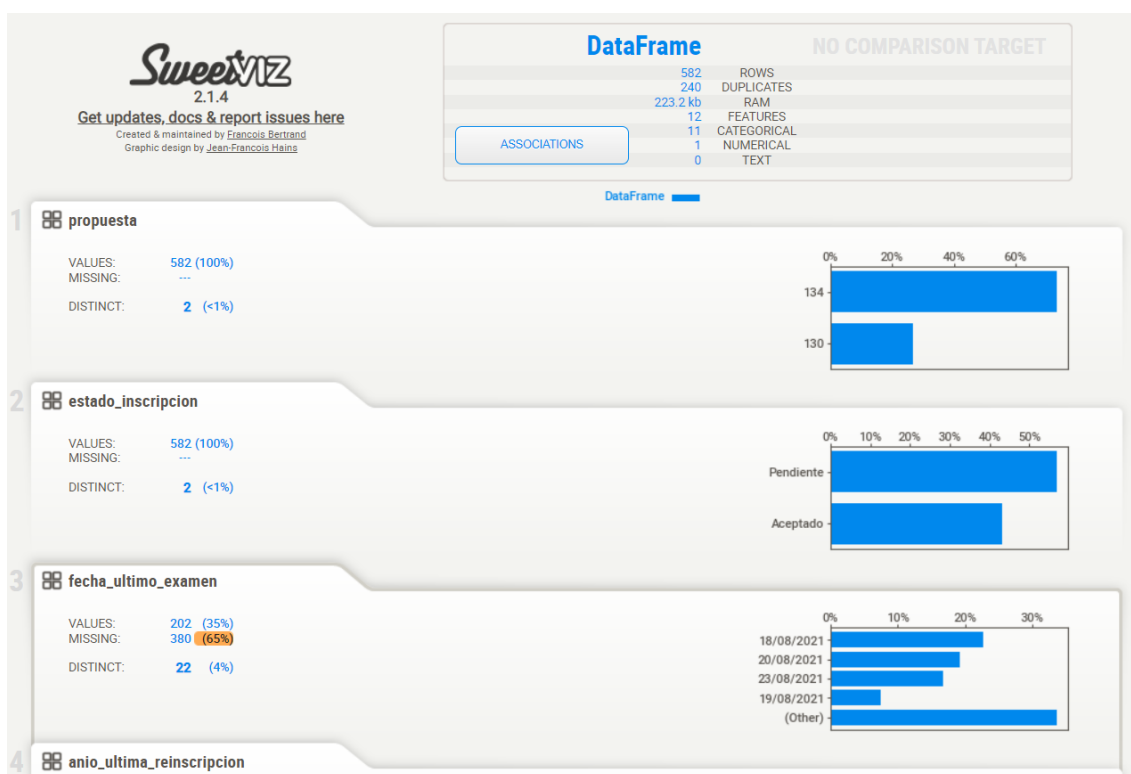


Ilustración 7. SweetViz: Datos Completos Filtrados

3.1.5 Selección de Técnicas Aplicables

Como ya se comentó, entre las técnicas que se aplicarán, se encuentran:

- Regresión Logística
- K-Nearest Neighbors
- Árboles de Decisión (TDIDT, *Top Down Induction of Decision Trees*)
- Métodos de Ensamblado
 - Random Forest
 - Gradient Boosting
- Red Neuronal
- Support Vector Machine (SVM)

Se realizarán distintas “**pruebas**” para poder variar los hiperparámetros de estos modelos y poder seleccionar el más adecuado.

3.2 Generación del Plan de Pruebas

3.2.1 Determinar tipo y cantidad 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.

3.3 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 notebooks de Google Colab disponibles en la siguiente ubicación: [GitHub](#).

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

3.3.1 Configuración de Parámetros

Antes de comenzar con el procesamiento del dataset mediante las técnicas seleccionadas se han realizado los siguientes ajustes:

- Sobre el atributo '**estado_inscripcion**' se han transformado sus valores según el siguiente criterio:
 - "Pendiente" = "P"
 - "Aceptada" = "A"
- Se genera un atributo '**rindio_examen**' que obtiene sus valores a partir del atributo 'fecha_ultimo_examen':
 - Si es una fecha nula = "N"
 - Si es una fecha concreta = "S" indicando que sí ha rendido exámenes
- Se genera un atributo '**inscripto_ult_ciclo**' que obtiene sus valores a partir del atributo 'anio_ultima_reinscripcion':
 - Si el valor es 2021 = "S"
 - En caso de valor diferente o nulo = "N" indicando que no se ha inscripto a cursar nuevamente

3.3.2 Datos Procesados y Dummies

Al ejecutar los ajustes anteriores sobre los **Datos Completos Filtrados**, se obtendrá un dataset de **Datos Procesados**. Este último será convertido en *dummies* y resultará en un dataset de **Data Dummies**.



data_procesados.csv

Tabla 6. Dataset Datos Procesados



data_dummies.csv

Tabla 7. Dataset Datos Dummies

3.3.3 Generación de los Modelos

Se definirán los datasets correspondientes al entrenamiento y testing, con ayuda de la biblioteca **train_test_split** de **Scikit-Learn**. Se ha optado por usar una partición correspondiente a:

- 75% de Entrenamiento
- 25% de Testing

Lo cual es totalmente parametrizable con el objeto **test_size**.

Adicionalmente, se definirá una función que permita evaluar los modelos utilizando varias métricas. Se usará la biblioteca **sklearn.metrics** para calcular tales métricas.

3.3.4 Descripción de los Modelos

Los modelos que están contemplados son los siguientes:

- Regresión Logística
- K-Nearest Neighbors
- Árboles de Decisión (TDIDT, *Top Down Induction of Decision Trees*)
- Métodos de Ensamblado
 - Random Forest
 - Gradient Boosting
- Red Neuronal
- Support Vector Machine (SVM)

3.4 Evaluación del Modelo

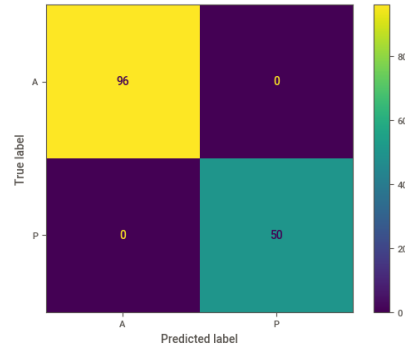
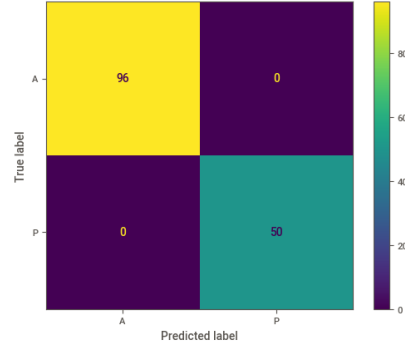
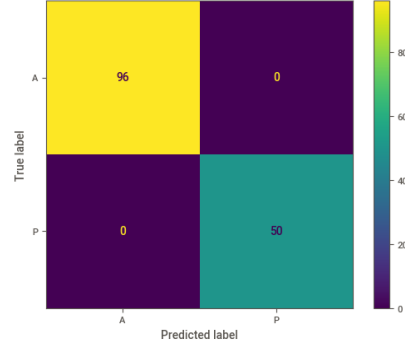
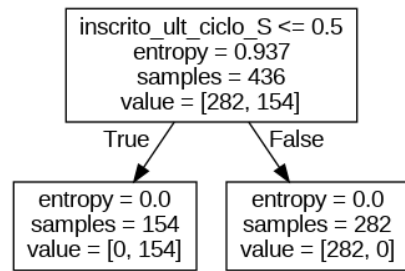
En esta etapa se procederá a la ejecución de 2 bloques de Pruebas, que incluirán modelos y parámetros como se especifica a continuación.

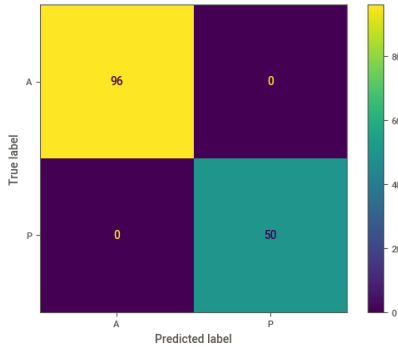
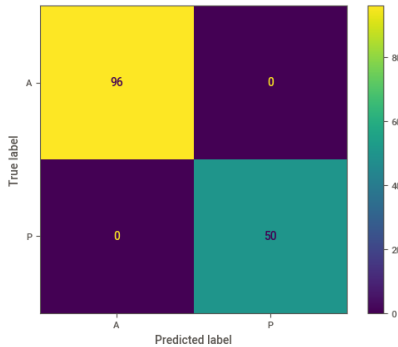
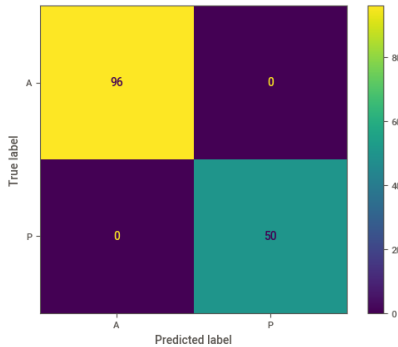
3.4.1 Prueba 1

Se ejecutarán los siguientes modelos:

- Regresión Logística
- K-Nearest Neighbors
- Árboles de Decisión (TDIDT, *Top Down Induction of Decision Trees*)
- Métodos de Ensamblado
 - Random Forest
 - Gradient Boosting
- Red Neuronal
- Support Vector Machine (SVM)

Técnica	Hiperparámetros	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,	Rendimiento obtenido: 1.0 Matriz de confusión:

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

	<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> 
GradientBoostingClassifier (Gradient Boosting)	<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> 
MLPClassifier (Red Neuronal)	<pre>MLPClassifier(activation='relu', alpha=0.0001, batch_size='auto', beta_1=0.9, beta_2=0.999, early_stopping=False, epsilon=1e-08, hidden_layer_sizes=(100,), learning_rate='constant', learning_rate_init=0.001, max_fun=15000, max_iter=300, momentum=0.9, n_iter_no_change=10, nesterovs_momentum=True, power_t=0.5, random_state=0, shuffle=True, solver='adam', tol=0.0001, validation_fraction=0.1, verbose=False,</pre>	<p>Rendimiento obtenido: 1.0</p> <p>Matriz de confusión:</p> 

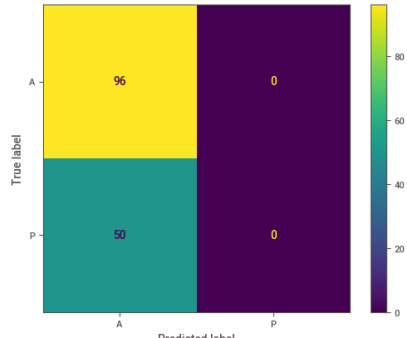
	warm_start=False)	
SVC (Support Vector Classifier)	SVC(C=1.0, break_ties=False, cache_size=200, class_weight=None, coef0=0.0, decision_function_shape='ovr', degree=3, gamma='scale', kernel='sigmoid', max_iter=-1, probability=False, random_state=0, shrinking=True, tol=0.001, verbose=False)	<p>Rendimiento obtenido: 0.66</p> <p>Matriz de confusión:</p> 

Tabla 8. Evaluación Prueba 1

Adicionalmente, para el modelo **SVC** se aplicó un **RandomizedSearchCV** para obtener mejores parámetros. Estos fueron los resultados:

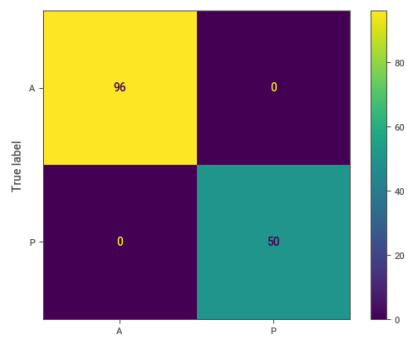
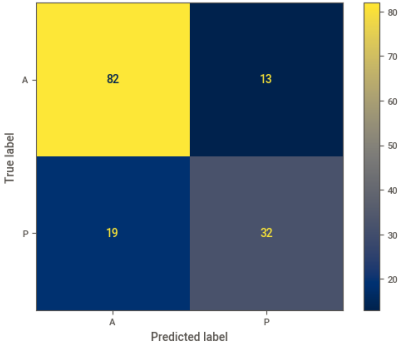
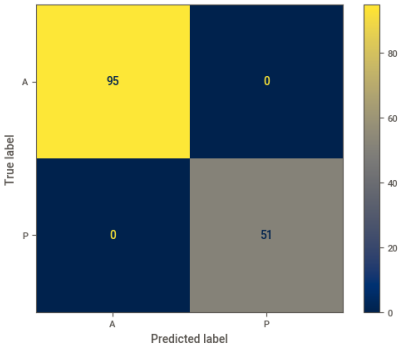
Técnica	Hiperparámetros	Resultados
SVC	SVC(C=0.1, break_ties=False, cache_size=200, class_weight=None, coef0=0.0, decision_function_shape='ovr', degree=3, gamma=0.01, kernel='linear', max_iter=-1, probability=False, random_state=0, shrinking=True, tol=0.001, verbose=False)	<p>Rendimiento obtenido: 1.0</p> <p>Matriz de confusión:</p> 

Tabla 9. Evaluación Prueba 1 RandomizedSearch

3.4.2 Prueba 2

Se ejecutarán los siguientes modelos:

- Métodos de Ensamblado
 - Gradient Boosting 2 (es decir, una variante del usado en la Prueba 1)
- Red Neuronal 2 (es decir, una variante del usado en la Prueba 1)

Técnica	Hiperparámetros	Resultados
MLPClassifier (Red Neuronal 2)	MLPClassifier(activation='tanh', alpha=0.0001, batch_size='auto', beta_1=0.9, beta_2=0.999, early_stopping=True, epsilon=1e-08, hidden_layer_sizes=(100, 50), learning_rate='adaptive', learning_rate_init=0.001, max_fun=15000, max_iter=500, momentum=0.9, n_iter_no_change=10, nesterovs_momentum=True, power_t=0.5, random_state=0, shuffle=True, solver='sgd', tol=0.0001, validation_fraction=0.1, verbose=False, warm_start=False)	Rendimiento obtenido: 0.78 Matriz de confusión: 
GradientBoostingClassifier (Gradient Boosting 2)	GradientBoostingClassifier(ccp_alpha=0.0, criterion='friedman_mse', init=None, learning_rate=0.1, loss='log_loss', max_depth=3, max_features='sqrt', max_leaf_nodes=None, min_impurity_decrease=0.0, min_samples_leaf=2, min_samples_split=4, min_weight_fraction_leaf=0.0, n_estimators=200, n_iter_no_change=None,	Rendimiento obtenido: 1.0 Matriz de confusión: 

	<pre> random_state=0, subsample=1.0, tol=0.0001, validation_fraction=0.1, verbose=0, warm_start=False) </pre>	
--	--	--

Tabla 10. Evaluación Prueba 2

3.5 Finaliza Fase de Modelado

Se registra en el **Tasksboard** la finalización de la Fase de Modelado.

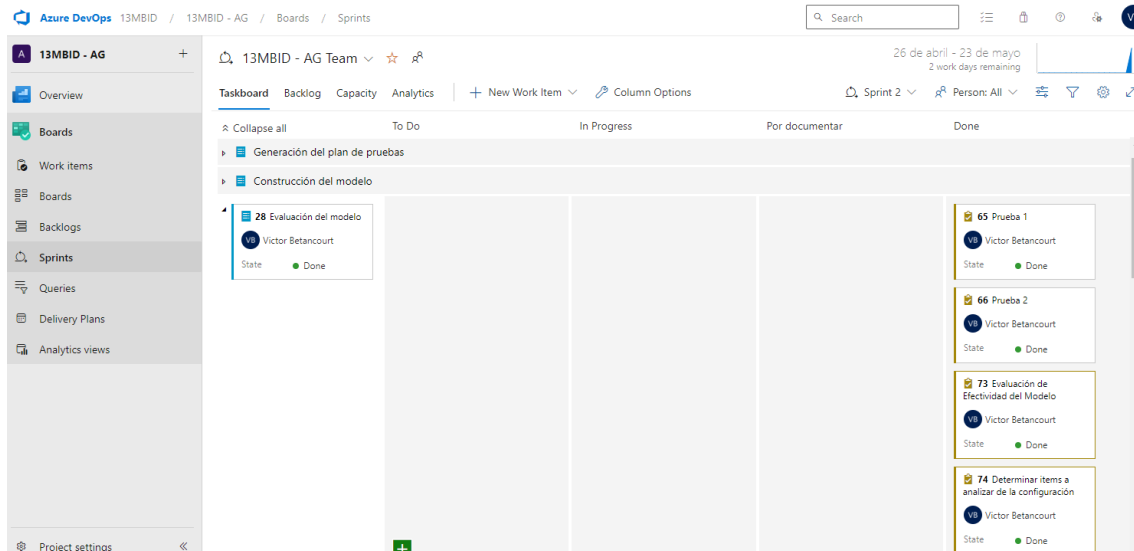


Ilustración 8. Azure: Finaliza Sprint 2 Fase Modelado

4 Sprint 2: Fase Evaluación

Después del Modelado, es necesario **evaluar los resultados** utilizando los criterios de negocio establecidos al inicio del proyecto. Esta es la clave para asegurar que la institución pueda hacer uso de los resultados que se han obtenido.

4.1 Evaluación de los Resultados

En función de los resultados obtenidos en la ejecución del plan de pruebas documentado en la sección anterior se han seleccionado las técnicas:

- Random Forest
- Gradient Boosting 2
- Red Neuronal 2

para ser utilizada sobre los datos de estudiantes nuevos, con base en los rendimientos obtenidos: 1.0, 1.0, y 0.78, respectivamente.

4.2 Proceso de Revisión

Con base en los resultados obtenidos de las Pruebas 1 y 2, se optará por aplicar los siguientes modelos en la Fase de Despliegue/Implantación:

- Random Forest
- Gradient Boosting 2
- Red Neuronal 2

Es decir, estos modelos se usarán para procesar los datos de los **estudiantes de 2022** y determinar el valor de su atributo “calidad” con base en sus datos almacenados en el sistema de gestión académica. Los **resultados de la predicción** se encuentran en el apartado “[Informe final](#)” del presente documento.

4.3 Determinación de Futuras Tareas

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

- integrar datos pertinentes a la condición socioeconómica de cada estudiante. Esto incluiría factores tales como: su situación laboral, si es responsable de otras personas, entre otras consideraciones.

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

- Como un ejemplo inicial, sería beneficioso revisar la integridad referencial de los datos disponibles, con el objetivo de incrementar el volumen de información utilizada en el entrenamiento de los modelos de aprendizaje automático.

Además, podemos agregar algunas tareas adicionales que podrían ser útiles:

- Experimentar con diferentes algoritmos de aprendizaje automático para ver si alguno ofrece mejores resultados.
- Asegurarse de que los datos estén balanceados, ya que los desequilibrios en las clases pueden afectar el rendimiento del modelo.
- Considerar la utilización de métodos de selección de características para eliminar características no esenciales y reducir la dimensionalidad de los datos, mejorando así la eficiencia del modelo.
- Evaluar y mejorar la métrica de rendimiento del modelo actual, lo que puede llevar a mejoras en la predicción.
- Explorar la posibilidad de incorporar más datos demográficos, lo que podría enriquecer el análisis y las predicciones.

4.4 Finaliza Fase Evaluación

Se registra en el **Tasksboard** la finalización de la Fase de Evaluación.

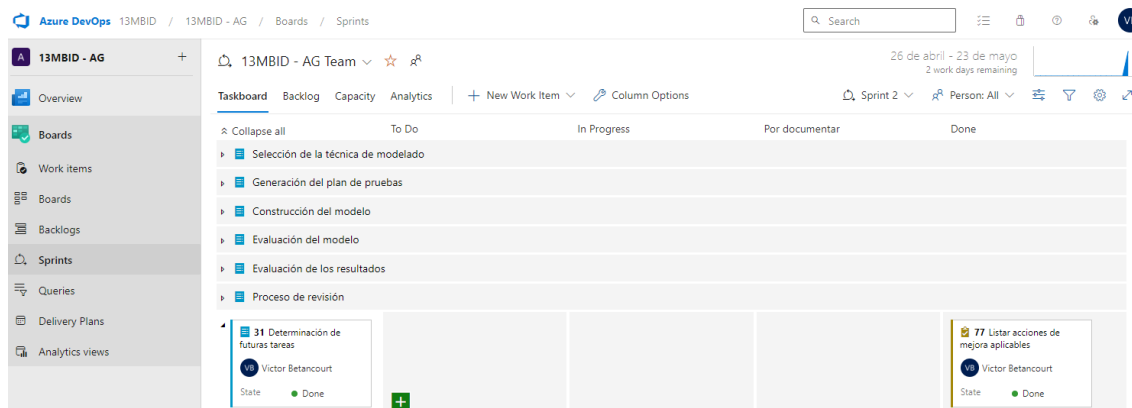


Ilustración 9. Azure: Finaliza Sprint 2 Fase Evaluación

5 Sprint 2: Fase de Despliegue

En general, la fase de [despliegue](#) de CRISP-DM se desarrollará con base en las siguientes etapas:

- 1) Planificación de la implementación
- 2) Supervisión y mantenimiento
- 3) Informe Final
- 4) Revisión del Proyecto

5.1 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.

5.2 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**.
- Evaluación de **Feedback de Usuarios**. Dado que el sistema se implementará en un entorno educativo y será utilizado por diferentes partes interesadas (como administradores, profesores y estudiantes), es fundamental recoger y evaluar el feedback de los usuarios. Esto permitirá realizar ajustes o mejoras al modelo en base a las experiencias prácticas y las necesidades del usuario.
- Revisión y **Actualización Regular del Modelo**. Los modelos de Machine Learning pueden volverse obsoletos con el tiempo debido a los cambios en las tendencias y patrones subyacentes en los datos. Por lo tanto, es crucial realizar revisiones regulares

del modelo y, si es necesario, actualizarlo con nuevos datos y ajustar sus parámetros para garantizar su rendimiento y precisión continuos. Este proceso puede incluir el reentrenamiento del modelo con nuevos datos, la comparación del rendimiento con versiones anteriores del modelo y la aplicación de nuevas técnicas o algoritmos si se vuelven disponibles.

5.3 Informe Final

En esta sección, se presentarán los resultados de aplicación del modelo generado con la técnica de mejor rendimiento sobre los datos de estudiantes nuevos para el periodo 2022.

5.3.1 Datos Nuevos



datos_nuevos_22.csv

Tabla 11. Dataset Datos Nuevos

Estos datos tendrán que ser procesados con los mismos ajustes que se aplicaron en la [Configuración de Parámetros](#). Luego se codificarán, es decir, se convertirán en Dummies. Y finalmente, se hará un match entre este dataset de Dummies y el dataset de Entrenamiento.

A este último dataset le llamaremos Nuevos Codificados Matched:



nuevos_codif_matched.csv

Tabla 12. Datos Nuevos Codificados Matched

5.3.2 Datasets para Despliegue

Los datasets que se utilizarán para la Fase del Despliegue son 2:

- **Data Despliegue:** Es una copia de Data Dummies
- **Nuevos Codificados Despliegue:** Es una copia de Nuevos Codificados Matched



Tabla 13. Data Despliegue



Tabla 14. Datos Nuevos Codificados Despliegue

5.3.3 Variables del Despliegue

Data Despliegue

- 1) 'propuesta',
- 2) 'avance_ingreso',
- 3) 'avance_1er_semestre',
- 4) 'avance_carrera',
- 5) 'segundo_anio_N',
- 6) 'segundo_anio_S',
- 7) 'rango_promedios_Alto',
- 8) 'rango_promedios_Bajo',
- 9) 'rango_promedios_Medio',
- 10) 'exámenes_1er_semestre_A',
- 11) 'exámenes_1er_semestre_I',
- 12) 'exámenes_1er_semestre_N',
- 13) 'estadoN_A',
- 14) 'estadoN_P',
- 15) 'rindio_examen_N',
- 16) 'rindio_examen_S'

Nuevos Codificados Despliegue

```
1) 'propuesta',  
2) 'avance_ingreso',  
3) 'avance_1er_semestre',  
4) 'avance_carrera',  
5) 'segundo_anio_N',  
6) 'segundo_anio_S',  
7) 'rango_promedios_Alto',  
8) 'rango_promedios_Bajo',  
9) 'rango_promedios_Medio',  
10) 'exámenes_1er_semestre_A',  
11) 'exámenes_1er_semestre_I',  
12) 'exámenes_1er_semestre_N',  
13) 'estadoN_A',  
14) 'estadoN_P',  
15) 'rindio_examen_N',  
16) 'rindio_examen_S'
```

5.3.4 Modelos para Predicción

- Métodos de Ensamblado
 - Random Forest
 - Gradient Boosting 2
- Red Neuronal 2

5.3.5 Resultados

Técnica	Hiperparámetros	Resultados
Random Forest	<p>RandomForestClassifier</p> <p>RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None, criterion='gini', max_depth=None, max_features='sqrt',</p> <p>max_leaf_nodes=None, max_samples=None,</p> <p>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)</p>	<p>A 780 (80.6%)</p> <p>P 187 (19.4%)</p>
GradientBoostingClassifier (Gradient Boosting 2)	<p>GradientBoostingClassifier(ccp_alpha=0.0, criterion='friedman_mse', init=None, learning_rate=0.1, loss='log_loss', max_depth=3, max_features='sqrt', max_leaf_nodes=None,</p> <p>min_impurity_decrease=0.0, min_samples_leaf=2,</p> <p>min_samples_split=4, min_weight_fraction_leaf=0.0, n_estimators=200, n_iter_no_change=None, random_state=0, subsample=1.0, tol=0.0001,</p> <p>validation_fraction=0.1, verbose=0, warm_start=False)</p>	<p>A 757 (78.2%)</p> <p>P 210 (21.8%)</p>
MLPClassifier (Red Neuronal 2)	<p>MLPClassifier(activation='tanh', alpha=0.0001, batch_size='auto', beta_1=0.9, beta_2=0.999,</p>	<p>A 967 (100%)</p> <p>P 0 (0%)</p>

	<pre> early_stopping=True, epsilon=1e-08, hidden_layer_sizes=(100, 50), learning_rate='adaptive', learning_rate_init=0.001, max_fun=15000, max_iter=500, momentum=0.9, n_iter_no_change=10, nesterovs_momentum=True, power_t=0.5, random_state=0, shuffle=True, solver='sgd', tol=0.0001, validation_fraction=0.1, verbose=False, warm_start=False) </pre>	
--	--	--

Tabla 15. Resultados Predicción

5.3.6 Datasets de Predicciones

A continuación, se hacen disponibles los datasets resultantes de la predicción por cada modelo.




Random Forest	Gradient Boosting 2	Red Neuronal 2
 datos_nuevos_calidad_rndf.csv	 datos_nuevos_calidad_gbc.csv	 datos_nuevos_calidad_mlp.csv

Tabla 16. Datasets Resultados Predicción

5.4 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.
- Implantar una solución de **gestión de proyectos** para facilitar la planificación, seguimiento y evaluación del rendimiento del equipo y del proyecto en general.
- Desarrollar una mejor **estrategia de gestión de datos**, que pueda incluir una política de respaldo de datos y un plan para la gestión de datos faltantes o erróneos.
- Establecer una **metodología más rigurosa para el diseño de experimentos**, lo que podría ayudar a mejorar la calidad y la eficacia de las pruebas realizadas.
- Introducir más herramientas de **visualización de datos** para facilitar la interpretación de los resultados y el descubrimiento de patrones.
- Promover sesiones de **formación continua** para el equipo de trabajo, centradas en las últimas tendencias y técnicas en la ciencia de datos, lo que podría mejorar la eficacia del equipo.
- Implementar un sistema de **control de versiones** para el código y los modelos desarrollados, lo que puede facilitar el seguimiento de los cambios y permitir la reproducción de los resultados.
- Considerar la posibilidad de **expandir el equipo** de trabajo para incluir roles adicionales, como un ingeniero de datos o un especialista en aprendizaje automático, dependiendo de las necesidades del proyecto.
- Planificar **revisiones periódicas post-despliegue** para asegurar que el modelo continúa funcionando como se espera a medida que los datos evolucionan con el tiempo.

5.5 Finaliza Fase Despliegue

Se registra en el **Tasksboard** la finalización de la Fase de Despliegue.

- **Vista Backlogs / Backlog**

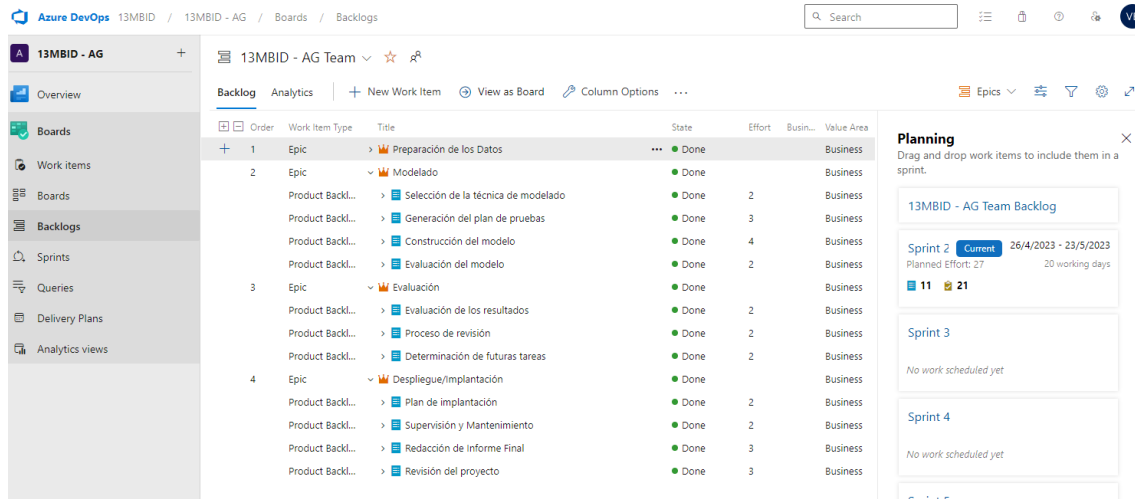


Ilustración 10. Azure Backlogs/Backlog: Finaliza Sprint 2 Fase Despliegue

- **Vista Sprints / Backlog**

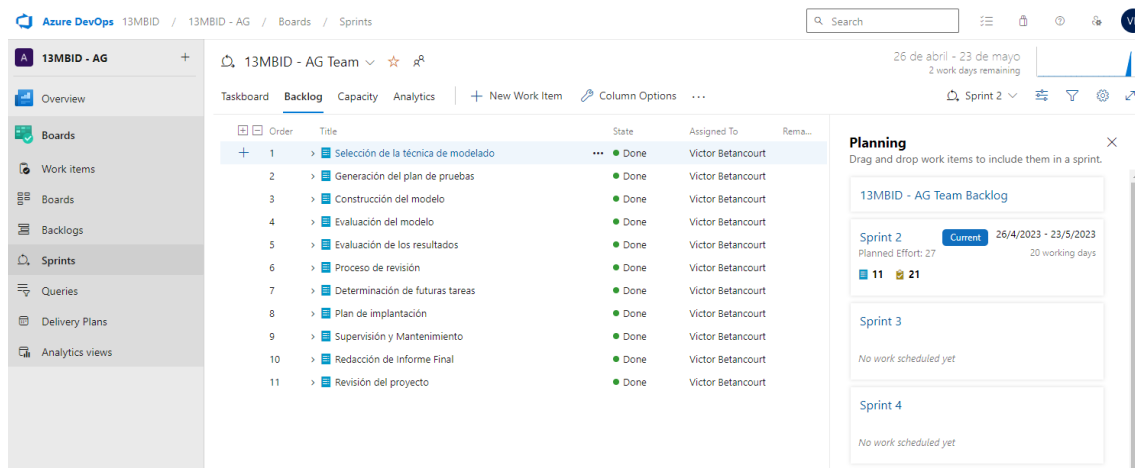


Ilustración 11. Azure Sprints/Backlog: Finaliza Sprint 2 Fase Despliegue

- **Vista Sprints / Taskboard**

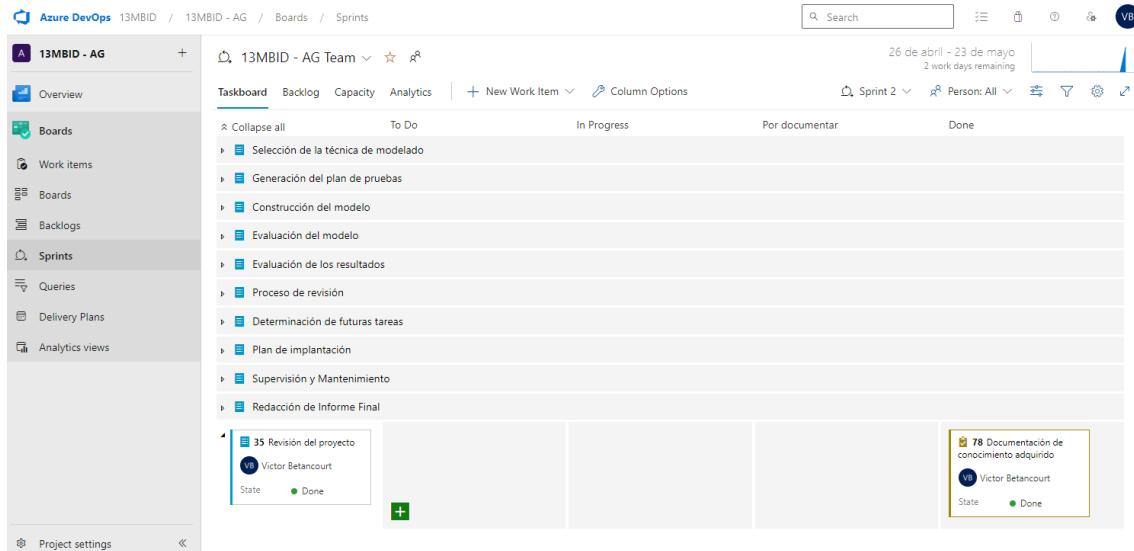


Ilustración 12. Azure Sprints/Taskboard: Finaliza Sprint 2 Fase Despliegue

6 Conclusiones

Se han completado todas etapas de la **Metodología CRISP-DM**, a saber:

- **Comprensión del Negocio**
- **Comprensión de los Datos**
- **Preparación de los Datos**
- **Modelado**
- **Evaluación**
- **Despliegue**

La herramienta [Azure DevOps](#), en particular, **Azure Boards**, ha resultado de gran aliado en el desarrollo de las fases antes descritas, y se han documentado las tareas correspondientes a los **2 Sprints**.

6.1 Áreas de Oportunidad

- Se detectaron en el dataset **datos_academicos** los siguientes atributos presentan valores nulos:
 - fecha_ultimo_examen = 406 filas nulas
 - anio_ultima_reinscripcion = 206 filas nulas
 - Total = 612

La calidad de estos atributos debería formar parte de la estrategia de calidad de datos de la institución.

6.2 Recomendaciones

- Implementar un sistema de **control de versiones para el código** y los modelos desarrollados, lo que puede facilitar el seguimiento de los cambios y permitir la reproducción de los resultados.
- Establecer una **metodología más rigurosa para el diseño de experimentos**, lo que podría ayudar a mejorar la calidad y la eficacia de las pruebas realizadas.
- Planificar **revisiones periódicas post-despliegue** para asegurar que el modelo continúa funcionando como se espera a medida que los datos evolucionan con el tiempo.

6.3 Sprint Retrospective

- Es recomendable considerar la posibilidad de:
 - Robustecer las Fases de **Modelado y Evaluación**, aplicando los elementos propuestos en Metodologías como **CRISP-ML**.
 - Fortalecer la Fase de **Despliegue** utilizando los componentes planteados en la metodología **MLOps**, la cual busca mejorar la colaboración y comunicación entre los equipos de desarrollo y operaciones de aprendizaje automático, proporcionando un marco para la automatización y el control de calidad.
- Es crucial mantener un equilibrio en la documentación del proyecto. A pesar de su importancia, una **excesiva documentación** puede desviar de los principios de las Metodologías Ágiles, que promueven respuestas adaptativas a los cambios sobre la planificación extensa.
- En cuanto a la gestión de **reuniones**, se sugiere limitarlas solo a aquellas esenciales para la continuidad del proyecto.
- Se propone la inclusión de puntos "**pivote**" o "de corte" (como la generación de datasets), desde los cuales se pueda probar el código existente o generar nuevo código sin la necesidad de reiniciar la ejecución completa.
 - Esta estrategia requiere consideración cuidadosa, ya que podría ser una forma de **refactorización de código**, permitiendo el manejo de bloques de código independientes.
 - Sin embargo, al mismo tiempo, este enfoque podría conducir a la **redundancia** en bloques de código, lo cual podría afectar la eficiencia y la manejabilidad del mismo.

7 Glosario

Azure Boards	Es un servicio de seguimiento de trabajo de Microsoft que permite a los equipos planificar, rastrear y discutir el trabajo a lo largo de todo el ciclo de vida del desarrollo. Ofrece herramientas de seguimiento de trabajos personalizables, incluyendo tableros Kanban, backlogs, sprint planning tools, consultas de trabajo y gráficos.
Backlog	En la gestión de proyectos ágil, el Backlog es una lista de tareas pendientes que se necesitan para completar el proyecto. En Scrum, el Backlog del Producto (véase Product Backlog) es una lista de características o mejoras deseadas para el producto, priorizadas en función de su valor para el negocio.
Epic	Historia de Usuario o Epic. En la Metodología Ágil, una historia de usuario es una descripción breve y simple de una característica contada desde la perspectiva del usuario que desea la nueva capacidad. Los epics son un grupo de historias de usuario relacionadas que se pueden desglosar en tareas más pequeñas. Un epic puede abarcar múltiples equipos, puede ser entregado en múltiples sprints, y puede requerir el trabajo de múltiples usuarios.
Historia de Usuario	Véase Epic.
Product Backlog	Es una lista priorizada de características deseadas para el producto.
Sprint	En Scrum, un Sprint es un período de tiempo fijo (normalmente de dos a cuatro semanas) durante el cual se completa un conjunto definido de trabajo. Cada Sprint comienza con una reunión de planificación y termina con una revisión y una retrospectiva.

8 Bibliografía

- [1] Chandrasekara, Chaminda; Herath, Pushpa (2019). *Hands-On Azure Boards. Configuring and Customizing Process Workflows in Azure Devops Services*. Apress.
- [2] Gothelf, Jeff (2017). *Lean vs Agile vs Design Thinking*. Glenn Rock, NJ.
- [3] Green, David (2016). *SCRUM. Novice to Ninja*. SitePoint Pty. Ltd.
- [4] Hundhausen, Richard (2021). *Professional Scrum Development with Azure Devops*. Microsoft Press.
- [5] Rossberg, Joachim (2019). *Agile Project Management with Azure Devops: Concepts, Templates, and Metrics*. Apress.
- [6] Scotcher, Edward (2015). *Brilliant Agile Project Management. A Practical Guide to Using Agile, Scrum and Kanban*. Pearson.
- [7] Sutherland, Jeff (2014). *SCRUM. The Art of Doing Twice the Work in Half the Time*. Crown Business.
- [8] IBM. (s.f.). *SPSS Modeler: Documentación*. Recuperado de <https://www.ibm.com/docs/en/spss-modeler/saas>.