Índice

Índice

Registros de cambios

Revisión	Detalles de los cambios realizados	Fecha
0	Creación del documento	24 de junio de 2025
1	Se completa hasta el punto 5 inclusive	6 de Julio de 2025
2	Se completa hasta el punto 9 inclusive	15 de Julio de 2025

.UBAfiuba FACULTAD DE INGENIERÍA
Plan de proyecto del Trabajo Final
Carrera de Especialización en Inteligencia Artificial
Ing. Gaspar Acevedo Zain

Registros de cambios

Revisión	Detalles de los cambios realizados	Fecha
0	Creación del documento	24 de junio de 2025
1	Se completa hasta el punto 5 inclusive	6 de Julio de 2025

Acta de constitución del proyecto

Buenos Aires, 24 de junio de 2025

Por medio de la presente se acuerda con el Ing. Gaspar Acevedo Zain que su Trabajo Final de la Carrera de Especialización en Inteligencia Artificial se titulará "Predicción de quiebra de empresas mediante técnicas de Machine Learning" y consistirá en el desarrollo de una herramienta basada en Machine Learning que permitirá predecir si una empresa puede entrar en quiebra o no. El trabajo tendrá un presupuesto preliminar estimado de 604 horas y un costo estimado de $ XXX, con fecha de inicio el 24 de junio de 2025 y fecha de presentación pública el a definir.

Se adjunta a esta acta la planificación inicial.

Dr. Ing. Ariel Lutenberg
Director posgrado FIUBA

Nombre del cliente
Empresa del cliente

Título y Nombre del director
Director del Trabajo Final

Acta de constitución del proyecto

Buenos Aires, 24 de junio de 2025

Por medio de la presente se acuerda con el Ing. Gaspar Acevedo Zain que su Trabajo Final de la Carrera de Especialización en Inteligencia Artificial se titulará "Predicción de quiebra de empresas mediante técnicas de Machine Learning" y consistirá en el desarrollo de una herramienta basada en Machine Learning que permita predecir si una empresa puede entrar en quiebra o no. El trabajo tendrá un presupuesto preliminar estimado de 600 horas y un costo estimado de $ XXX, con fecha de inicio el 24 de junio de 2025 y fecha de presentación pública el a definir.

Se adjunta a esta acta la planificación inicial.

Dr. Ing. Ariel Lutenberg
Director posgrado FIUBA

Nombre del cliente
Empresa del cliente

Título y Nombre del director
Director del Trabajo Final

1. Descripción técnica-conceptual del proyecto a realizar

Este proyecto consiste en un emprendimiento personal cuyo objetivo es utilizar técnicas de aprendizaje de máquina para detectar si una empresa puede entrar en quiebra o no. Este tipo de análisis puede resultar de gran interés y utilidad para distintos actores del mercado financiero, tales como bancos, compañías aseguradoras, fondos de inversión o consultoras especializadas en riesgo crediticio. Por ello, estos se considerarán como potenciales clientes.

Para llevarlo a cabo, se utilizará un *dataset* publicado por el Taiwan Economic Journal, que contiene información financiera de empresas del mercado de Taiwán entre los años 1999 y 2009. Al ser estos datos públicos, hoy en día existen soluciones que exploran esta temática. Algunas de ellas hacen uso de modelos de *machine learning* tales como *SVM* y *XGBoost*, junto con algunas técnicas de preprocesamiento de datos como *Smote* y de búsqueda de hiperparámetros como *Random Search*.

Con el fin de diferenciarse de estas soluciones, se propone implementar el marco de trabajo basado en *MLFlow* definido en la figura 1. Se detalla una serie de etapas cuyas salidas se refinarán durante distintas iteraciones. Esto permitirá a los usuarios finales trabajar en un entorno seguro, robusto, y reproducible.

El proyecto se encuentra en la etapa de planificación. El desarrollo e implementación se realizará en distintas etapas. Se comenzará con un análisis exploratorio de datos, que nos permitirá conocer mejor al dataset en cuestión. Luego, se realizarán iteraciones sobre las siguientes etapas:

- Preprocesamiento de datos: en la primer iteración se implementarán técnicas de tratamiento de nulos y desbalance de clases. En las siguientes iteraciones, se estudiarán técnicas de extracción e ingeniería de features.
- Entrenamiento de modelos: se implementará un modelo distinto en cada iteración. Los modelos a explorar son regresión logística, *SVM* y *XGBoost*. También, se explorará la optimización de hiperparámetros mediante búsqueda bayesiana.
- Evaluación y refinamiento: en esta etapa se evaluará al modelo entrenado en la etapa anterior. Se generarán métricas que permitirán compararlo con resultados obtenidos en otras iteraciones.

La innovación de este proyecto radica en el uso del marco de trabajo definido en la figura 1. Éste proporciona un ambiente productivo, reproducible y escalable, en donde se podrán analizar diversas técnicas de aprendizaje de máquina para detectar si una empresa puede entrar en quiebra o no.

.UBAfiuba FACULTAD DE INGENIERÍA
Plan de proyecto del Trabajo Final
Carrera de Especialización en Inteligencia Artificial
Ing. Gaspar Acevedo Zain

1. Descripción técnica-conceptual del proyecto a realizar

Este proyecto consiste en un emprendimiento personal cuyo objetivo es utilizar técnicas de aprendizaje de máquina para detectar si una empresa puede entrar en quiebra o no. Este análisis puede resultar de gran interés y utilidad para distintos participantes del mercado financiero. Por ello, se considerarán como potenciales usuarios finales a empresas especializadas en finanzas, en inversiones, en prestación de seguros, entre otras.

Para llevar a cabo al mismo, se utilizará un *dataset* publicado por el Taiwan Economic Journal, el cual contiene información financiera de empresas del mercado de Taiwán entre los años 1999 y 2009. Al ser estos datos públicos, hoy en día existen soluciones que exploran esta temática. Algunas de ellas hacen uso de modelos de *machine learning* tales como *SVM* y *XGBoost*, junto con algunas técnicas de preprocesamiento de datos como *Smote* y de búsqueda de hiperparámetros como *Random Search*.

Con el fin de diferenciarse de estas soluciones, se propone implementar el marco de trabajo basado en *MLFlow* definido en la figura 1. En el mismo se detallan una serie de etapas cuyas salidas se refinarán durante distintas iteraciones. Esto permitirá explorar diversas técnicas de preprocesamiento de datos, búsqueda de hiperparámetros y modelos de *machine learning* que ayuden a lograr el objetivo de este proyecto.

El proyecto se encuentra en la etapa de planificación. El desarrollo e implementación del mismo se realizará en distintas etapas. Se comenzará con un análisis exploratorio de datos, que nos permitirá conocer mejor al dataset en cuestión. Luego, se realizarán diversas iteraciones sobre las siguientes etapas:

- Preprocesamiento de datos: en la primer iteración se implementarán técnicas de tratamiento de nulos y desbalance de clases. En las siguientes iteraciones, se estudiarán técnicas de extracción e ingeniería de features.
- Entrenamiento de modelos: se implementará un modelo distinto en cada iteración. Los modelos a explorar son regresión logística, *SVM* y *XGBoost*. También se explorará la optimización de hiperparámetros mediante búsqueda bayesiana.
- Evaluación y refinamiento: en esta etapa se evaluará al modelo entrenado en la etapa anterior. Se generarán métricas que permitirán compararlo con resultados obtenidos en otras iteraciones.

La innovación de este proyecto radica en el uso del marco de trabajo definido en la figura 1. Éste proporciona un ambiente productivo, reproducible y escalable, en donde se podrán analizar diversas técnicas de aprendizaje de máquina para detectar si una empresa puede entrar en quiebra o no.

Figura 1. Diagrama en bloques del sistema.

2. Identificación y análisis de los interesados

Rol	Nombre y Apellido	Organización	Puesto
Responsable	Ing. Gaspar Acevedo Zain	FIUBA	Alumno
Orientador	Título y Nombre del director	pertenencia	Director del Trabajo Final
Cliente	Actores del mercado financiero	-	-
Usuario final	Trabajadores de clientes	-	-

- Orientador: podrán ayudar en la recomendación y evaluación de técnicas a explorar en las diferentes etapas del proyecto.
- Cliente: si bien es un proyecto personal, se considerarán como potenciales clientes a distintos actores del mercado financiero, tales como bancos, compañías aseguradoras, fondos de inversión o consultoras especializadas en riesgo crediticio.
- Usuario final: analistas de riesgos, ejecutivo de créditos, entre otros integrantes que trabajan para los potenciales clientes.

3. Propósito del proyecto

Predecir si una empresa puede entrar en quiebra o no, al explorar técnicas de *machine learning* en un marco de trabajo productivo, reproducible y escalable.

4. Alcance del proyecto

El alcance del proyecto incluye:

- Análisis exploratorio de datos: se analizarán las distintas variables presentes en el *dataset* de estudio, con el fin de conocer sus características y poder tomar decisiones con base en ellas.

Figura 1. Diagrama en bloques del sistema.

2. Identificación y análisis de los interesados

Rol	Nombre y Apellido	Organización	Puesto
Responsable	Ing. Gaspar Acevedo Zain	FIUBA	Alumno
Orientador	Título y Nombre del director	pertenencia	Director del Trabajo Final
Usuario final	Mercado financiero	-	-

- Orientador: podrán ayudar en la recomendación y evaluación de técnicas a explorar en las diferentes etapas del proyecto.
- Usuario final: si bien es un proyecto personal, se considerarán como potenciales usuarios finales a empresas del sector financiero. Estas podrán encontrar interés y utilidad en las predicciones de los modelos explorados.

3. Propósito del proyecto

Predecir si una empresa puede entrar en quiebra o no, al explorar técnicas de *machine learning* en un marco de trabajo productivo, reproducible y escalable.

4. Alcance del proyecto

El alcance del proyecto incluye:

- Análisis exploratorio de datos: se analizarán las distintas variables presentes en el *dataset* de estudio, con el fin de conocer sus características y poder tomar decisiones en base a ellas.
- Preprocesamiento de datos: se realizarán técnicas de tratamiento de datos faltantes, selección y/o extracción de features, como así también de ingeniería de features.
- Implementación de modelos de *machine learning*: se estudiarán diversos modelos de aprendizaje de máquina sobre los datos procesados, tales como *logistic regression*, *SVM*

- Preprocesamiento de datos: se realizarán técnicas de tratamiento de datos faltantes, selección y/o extracción de variables, como así también de ingeniería de features.
- Implementación de modelos de *machine learning*: se estudiarán diversos modelos de aprendizaje de máquina sobre los datos procesados, tales como *logistic regression*, *SVM* y *XGBoost*. Además, se optimizarán los hiperparámetros de estos modelos mediante búsqueda bayesiana.
- Evaluación y comparación de modelos: se obtendrán métricas relacionadas a los modelos explorados, con el fin de poder determinar cuál de ellos realiza una mejor predicción.
- Implementación de un entorno basado en *MLFlow*: este entornó facilitará la realización, la reproducibilidad y la escalabilidad de las distintas etapas de trabajo que se realizarán en este proyecto. Este será de caracter local, es decir, no se implementará en una plataforma de *cloud computing*.

No se incluye:

- El despliegue del entorno de trabajo en una plataforma de *cloud computing*, tales como *Azure*, *AWS*, entre otros.
- El análisis de otros *datasets* distintos al propuesto.

5. Supuestos del proyecto

Para el desarrollo del presente proyecto se supone que:

- Supuesto 1: el *dataset* de estudio presenta datos fiables, y no tiene restricciones en cuanto a licencias de uso.
- Supuesto 2: una *laptop* como equipo de trabajo es más que suficiente para realizar el preprocesamiento y entrenamiento de los modelos de aprendizaje automático.
- Supuesto 3: el entorno de *MLFlow* podrá desarrollarse en etapas futuras del proyecto, posteriores a la exploración de los modelos de aprendizaje automático.
- Supuesto 4: el entorno de *MLFlow* podrá desplegarse de manera local, sin necesidad de recurrir a plataforma de *cloud computing*, tales como *Azure*, *AWS*, entre otros.
- Supuesto 5: se disponen de al menos 15 horas semanales para realizar el proyecto.

6. Product Backlog

Roles

- *Ingeniero del proyecto*: es quien se encarga del análisis, diseño, desarrollo y despliegue del proyecto.
- *Usuario final*: es quien consulta y analiza las predicciones de los modelos explorados en el proyecto.

y *XGBoost*. Además, se optimizarán los hiperparámetros de estos modelos mediante búsqueda bayesiana.

- Evaluación y comparación de modelos: se obtendrán métricas relacionadas a los modelos explorados, con el fin de poder determinar cual de ellos realiza una mejor predicción.
- Implementación de un entorno basado en *MLFlow*: este entornó facilitará la realización, la reproducibilidad y la escalabilidad de las distintas etapas de trabajo que se realizarán en este proyecto. El mismo será de caracter local, es decir, no se implementará en una plataforma de *cloud computing*.

No se incluye:

- El despliegue del entorno de trabajo en una plataforma de *cloud computing*, tales como *Azure*, *AWS*, entre otros.
- El análisis de otros *datasets* distintos al propuesto.

5. Supuestos del proyecto

Para el desarrollo del presente proyecto se supone que:

- Supuesto 1: el *dataset* de estudio presenta datos fiables, y no tiene restricciones en cuanto a licencias de uso.
- Supuesto 2: una *laptop* como equipo de trabajo es más que suficiente para realizar el preprocesamiento y entrenamiento de los modelos de aprendizaje automático.
- Supuesto 3: el entorno de *MLFlow* podrá desarrollarse en etapas futuras del proyecto, posteriores a la exploración de los modelos de aprendizaje automático.
- Supuesto 4: el entorno de *MLFlow* podrá desplegarse de manera local, sin necesidad de recurrir a plataforma de *cloud computing*, tales como *Azure*, *AWS*, entre otros.
- Supuesto 5: se disponen de al menos 15 horas semanales para realizar el proyecto.

6. Product Backlog

El Product Backlog debe organizarse en cuatro ***épicas*** fundamentales del proyecto. Cada épica debe contener al menos dos historias de usuario que describan funcionalidades clave.

El Product Backlog debe permitir interpretar cómo será el proyecto y su funcionalidad. Se deben indicar claramente las prioridades entre las historias de usuario y si hay alguna opcional.

Las historias de usuario deben ser breves, claras y medibles, expresando el rol, la necesidad y el propósito de cada funcionalidad. También deben tener una prioridad definida para facilitar la planificación de los sprints.

Cada historia de usuario debe incluir una ponderación en *Story Points*, un número entero que representa el tamaño relativo de la historia. El criterio para calcular los Story Points debe indicarse explícitamente.

Criterios de ponderación de historias de usuario

Esto son los criterios que se utilizan para ponderar a las historias de usuario mediante *Story Points*:

- Dificultad: representa la cantidad de trabajo estimado que requiere la historia de usuario para realizarse.
- Complejidad: representa la dificultad de realizar la historia de usuario a nivel técnico.
- Incertidumbre: representa el riesgo asociado a la historia de usuario.

Cada criterio tiene asociado las ponderaciones *baja*, *media* y *alta*, que se detallan en el cuadro 1. Los *Story Points* de una historia de usuario quedan definidos por la suma de los valores de estas ponderaciones redondeada hacia el número superior más próximo en la serie de *Fibonacci*.

Criterio\Ponderación	Baja	Media	Alta
Dificultad	1	3	5
Complejidad	1	3	5
Incertidumbre	1	5	8

Cuadro 1. Tabla de ponderaciones de historia de usuario.

Épicas

- **Épica 1 - Análisis y procesamiento de datos**
  - HU1 - Análisis exploratorio
    - Como ingeniero del proyecto, quiero realizar un análisis exploratorio de datos para conocer las distribuciones, formas y otras particularidades de las variables del dataset con el que se trabajará.
    - Ponderación
      - Dificultad: Media - 3 *Story Points*
      - Complejidad: Baja - 1 *Story Points*
      - Incertidumbre: Baja - 1 *Story Points*
      - Suma: 5
      - Total: 5 *Story Points*
  - HU2 - Procesamiento de datos faltantes y datos atípicos
    - Como ingeniero del proyecto, quiero realizar un procesamiento de datos faltantes y de datos atípicos con el fin de asegurar la calidad del dataset.
    - Ponderación
      - Dificultad: Media - 3 *Story Points*
      - Complejidad: Media - 3 *Story Points*
      - Incertidumbre: Baja - 1 *Story Points*
      - Suma: 7
      - Total: 8 *Story Points*
  - HU3 - *Feature Engineering*
    - Como ingeniero del proyecto, quiero implementar *Feature Engineering* con el fin de crear nuevos atributos en el dataset.

Las historias deben seguir el formato: "*Como {rol}, quiero {tal cosa} para {tal otra cosa}*".

Las épicas deben estructurarse de la siguiente forma:

- **Épica 1**
  - HU1
  - HU2
- **Épica 2**
  - HU3
  - HU4
- **Épica 3**
  - HU5
  - HU6
- **Épica 4**
  - HU7
  - HU8

Reglas para definir historias de usuario:

- Ser concisas y claras.
- Expresarlas en términos cuantificables y medibles.
- No dejar margen para interpretaciones ambiguas.
- Indicar claramente su prioridad y si son opcionales.
- Considerar regulaciones y normas vigentes.

7. Criterios de aceptación de historias de usuario

Los criterios de aceptación deben establecerse para cada historia de usuario, asegurando que se cumplan las condiciones necesarias para que la funcionalidad sea validada correctamente.

Cada historia debe tener criterios medibles, específicos y verificables. Deben permitir validar que se cumple con las necesidades del usuario.

Se estructuran de forma análoga a las épicas del backlog:

- **Épica 1**
  - Criterios de aceptación HU1
  - Criterios de aceptación HU2

- Ponderación
  - Dificultad: Media - 3 *Story Points*
  - Complejidad: Media - 3 *Story Points*
  - Incertidumbre: Baja - 1 *Story Points*
  - Suma: 7
  - Total: 8 *Story Points*

- **Épica 2 - Implementación y comparación de modelos**
  - HU4 - Implementación de modelos de *Machine Learning*
    - Como ingeniero del proyecto, quiero implementar distintos modelos de *Machine Learning* que permitan predecir si una empresa entra en quiebra o no.
    - Ponderación
      - Dificultad: Media - 3 *Story Points*
      - Complejidad: Media - 3 *Story Points*
      - Incertidumbre: Media - 5 *Story Points*
      - Suma: 11
      - Total: 13 *Story Points*
  - HU5 - Optimización de hiperparámetros
    - Como ingeniero del proyecto, quiero implementar técnicas de optimización de hiperparámetros y aplicarlas a los modelos de *Machine Learning* implementados.
    - Ponderación
      - Dificultad: Media - 3 *Story Points*
      - Complejidad: Media - 3 *Story Points*
      - Incertidumbre: Media - 5 *Story Points*
      - Suma: 11
      - Total: 13 *Story Points*
  - HU6 - Métricas de modelos
    - Como ingeniero del proyecto, quiero calcular métricas en cada modelo de *Machine Learning* implementado y comparar sus resultados.
    - Ponderación
      - Dificultad: Media - 3 *Story Points*
      - Complejidad: Media - 3 *Story Points*
      - Incertidumbre: Baja - 1 *Story Points*
      - Suma: 7
      - Total: 8 *Story Points*

- **Épica 3 - Despliegue en entorno *MLFlow***
  - HU7 - Despliegue en *MLFlow*
    - Como ingeniero del proyecto, quiero desplegar un entorno local de *MLFLow* en donde se repliquen los pasos de procesamiento de datos e implementación y comparación de modelos.
    - Ponderación
      - Dificultad: Alta - 5 *Story Points*
      - Complejidad: Media - 3 *Story Points*
      - Incertidumbre: Media - 5 *Story Points*

- **Épica 2**
  - Criterios de aceptación HU3
  - Criterios de aceptación HU4
- **Épica 3**
  - Criterios de aceptación HU5
  - Criterios de aceptación HU6
- **Épica 4**
  - Criterios de aceptación HU7
  - Criterios de aceptación HU8

Reglas para definir criterios de aceptación:

- Medibles y verificables.
- Especificar cuándo una historia se considera completada.
- Incluir condiciones específicas.
- No ambiguos.
- Probables de testear funcional o técnicamente.
- Mínimo 3 criterios por HU.

8. Fases de CRISP-DM

1. **Comprensión del negocio:** objetivo, valor agregado de IA, métricas de éxito.
2. **Comprensión de los datos:** tipo, origen, cantidad, calidad.
3. **Preparación de los datos:** características clave, transformaciones necesarias.
4. **Modelado:** tipo de problema, algoritmos posibles.
5. **Evaluación del modelo:** métricas de rendimiento.
6. **Despliegue del modelo (opcional):** tipo de despliegue y herramientas.

9. Desglose del trabajo en tareas

A partir de cada HU, descomponer en tareas concretas, técnicas y medibles:

- Duración estimada: entre 2 y 8 h. Evitar tareas genéricas.
- Si una tarea excede 8 h, dividirla.
- Indicar prioridad relativa (Alta, Media, Baja).

- Suma: 13
- Total: 13 *Story Points*

- HU8 - *API* para entorno *MLFlow*
  - Como ingeniero del proyecto, quiero exponer el entorno de *MLFlow* mediante una *API* para facilitar el acceso y su consumo.
  - Ponderación
    - Dificultad: Baja - 1 *Story Points*
    - Complejidad: Media - 3 *Story Points*
    - Incertidumbre: Baja - 1 *Story Points*
    - Suma: 5
    - Total: 5 *Story Points*

- **Épica 4 - Documentación y calidad**
  - HU9 - Implementación de buenas prácticas
    - Como ingeniero del proyecto, quiero asegurar que el código siga las buenas prácticas y estándares de la industria.
    - Ponderación
      - Dificultad: Media - 3 *Story Points*
      - Complejidad: Baja - 1 *Story Points*
      - Incertidumbre: Baja - 1 *Story Points*
      - Suma: 5
      - Total: 5 *Story Points*
  - HU10 - Documentación
    - Como ingeniero del proyecto, quiero documentar todos los pasos realizados durante el proyecto.
    - Ponderación
      - Dificultad: Baja - 1 *Story Points*
      - Complejidad: Baja - 1 *Story Points*
      - Incertidumbre: Baja - 1 *Story Points*
      - Suma: 3
      - Total: 3 *Story Points*
  - HU11 - Validación de *API* de *MLFlow*
    - Como usuario final, quiero consultar los resultados y comparaciones de los modelos mediante la *API* del entorno de **MLFlow**, para poder analizarlos.
    - Ponderación
      - Dificultad: Media - 3 *Story Points*
      - Complejidad: Baja - 1 *Story Points*
      - Incertidumbre: Baja - 1 *Story Points*
      - Suma: 5
      - Total: 5 *Story Points*

7. Criterios de aceptación de historias de usuario

- **Épica 1 - Análisis y procesamiento de datos**

Historia de usuario	Tarea técnica	Estimación	Prioridad
HU1	Tarea 1 HU1	6 h	Alta
HU1	Tarea 2 HU1	8 h	Alta
HU2	Tarea 1 HU2	5 h	Media
HU2	Tarea 2 HU2	6 h	Alta
...	...	...	...

Criterios para estimar tiempos:

- Considerar dificultad, complejidad técnica y grado de incertidumbre de cada tarea.
- Evitar subestimar el esfuerzo requerido. Si una tarea supera las 8 h, dividirla en subtareas.
- Basar la estimación en la experiencia propia o en referencias de tareas similares.

Sobre la prioridad:

- Asignar una prioridad relativa (Alta, Media o Baja) según la relevancia funcional de la tarea y su impacto en los entregables.
- Priorizar tareas que estén vinculadas a criterios de aceptación de las HU o que sean necesarias para desbloquear otras.
- Incluir tareas opcionales solo si están bien justificadas.

Recomendaciones generales:

- Incluir al menos dos tareas por historia de usuario.
- El total estimado debe ser coherente con la planificación global de unas 600 horas (la cantidad de horas sugeridas para un trabajo de posgrado).
- Este desglose se utilizará en las secciones siguientes para armar el diagrama de Gantt (sección 10) y la planificación de sprints (sección 11).
- Recordar que la calidad del desglose impacta directamente en la calidad de la planificación del proyecto.

10. Diagrama de Gantt

El diagrama de Gantt debe representar de forma visual y cronológica todas las tareas del proyecto, abarcando aproximadamente 600 horas totales, de las cuales entre 480 y 500 deben destinarse a tareas técnicas (desarrollo, pruebas, implementación) y entre 100 y 120 a tareas no técnicas (planificación, documentación, escritura de memoria y preparación de la defensa).

Consignas y recomendaciones:

- Incluir tanto tareas técnicas derivadas de las HU como tareas no técnicas generales del proyecto.

- Criterios de aceptación HU1 - Análisis exploratorio
  - Se estudia la presencia de datos atípicos y de datos faltantes para cada variable.
  - Se grafican las distribuciones de las variables del dataset.
  - Se realiza un estudio de correlaciones entre variables numéricas.
  - Se documentan los hallazgos del análisis de cada variable.
- Criterios de aceptación HU2 - Procesamiento de datos faltantes y datos atípicos
  - Se realiza una imputación de datos faltantes a las variables del dataset.
  - Se justifican los métodos de imputación utilizados.
  - Se ajustan los datos atípicos de las variables del dataset.
  - Se justifican los métodos de ajuste utilizados.
  - Se justifican los casos en donde se decide no imputar ni ajustar.
- Criterios de aceptación HU3 - *Feature Engineering*
  - Se crean nuevas variables en el dataset a partir de las existentes.
  - Se estudia el impacto por separado de estas variables en los modelos generados, a partir de sus métricas.
  - Se justifica la inclusión o no en el modelo de cada variable generada.

- **Épica 2 - Implementación y comparación de modelos**
  - Criterios de aceptación HU4 - Implementación de modelos de *Machine Learning*
    - Se implementan distintos modelos de *Machine Learning*.
    - Se justifica el uso de cada uno de los modelos implementados.
    - Se persisten los modelos generados en GitHub, para futuros análisis y comparaciones.
  - Criterios de aceptación HU5 - Optimización de hiperparámetros
    - Se seleccionan los hiperparámetros de cada modelo a optimizar.
    - Se define el rango sobre el que se optimizará cada hiperparámetro.
    - Se realiza una búsqueda del valor óptimo de los hiperparámetros en los rangos definidos.
    - Se justifican las decisiones tomadas en cada paso.
  - Criterios de aceptación HU6 - Métricas de modelos
    - Se definen las métricas de análisis para cada modelo.
    - Se justifica la selección de cada métrica para cada modelo.
    - Se obtienen las métricas de análisis de cada modelo.
    - Se comparan los distintos modelos mediante las métricas definidas.

- **Épica 3 - Despliegue en entorno *MLFlow***
  - Criterios de aceptación HU7 - Despliegue en *MLFlow*
    - Se crea un entorno *MLFlow* local desde cero
    - Se configura el paso correspondiente al análisis de datos en el entorno.
    - Se replican las técnicas exploradas de análisis de datos en el paso correspondiente.
    - Se configura el paso de entrenamiento de modelos en el entorno.
    - Se replican las técnicas exploradas de entrenamiento de modelos en el paso correspondiente.
    - Se configura el paso de evaluación de modelos en el entorno.

- El eje vertical debe listar las tareas y el eje horizontal representar el tiempo en semanas o fechas.
- Utilizar colores diferenciados para distinguir tareas técnicas y no técnicas.
- Las tareas deben estar ordenadas cronológicamente y reflejar todo el ciclo del proyecto.
- Iniciar con la planificación del proyecto (coincidente con el inicio de Gestión de Proyectos) y finalizar con la defensa, próxima a la fecha de cierre del trabajo.
- Configurar el software para mostrar los códigos del desglose de tareas y los nombres junto a cada barra.
- Asegurarse de que la fecha final coincida con la del Acta Constitutiva.
- Evitar tareas genéricas o ambiguas y asegurar una secuencia lógica y realista.
- Las fechas pueden ser aproximadas; ajustar el ancho del diagrama según el texto y el parámetro `x unit`. Para mejorar la apariencia del diagrama, es necesario ajustar este valor y, quizás, acortar los nombres de las tareas.

Herramientas sugeridas:

- Planner, GanttProject, Trello + plugins
  https://blog.trello.com/es/diagrama-de-gantt-de-un-proyecto
- Creately (colaborativa online)
  https://creately.com/diagram/example/ieb3p3ml/LaTeX
- LaTeX con `pgfgantt`:
  http://ctan.dcc.uchile.cl/graphics/pgf/contrib/pgfgantt/pgfgantt.pdf

Incluir una imagen legible del diagrama de Gantt. Si es muy ancho, presentar primero la tabla y luego el gráfico de barras.

11. Planificación de Sprints

Organizar las tareas técnicas del proyecto en sprints de trabajo que permitan distribuir de forma equilibrada la carga horaria total, estimada en 600 horas.

Consigna:

- Completar una tabla que relacione sprints con HU y tareas técnicas correspondientes.
- Incluir estimación en horas para cada tarea.
- Indicar responsable y porcentaje de avance estimado o completado.
- Contemplar también tareas de planificación, documentación, redacción de memoria y preparación de defensa.

Conceptos clave:

- Se replican las técnicas exploradas de evaluación de modelos en el paso correspondiente.

- Criterios de aceptación HU8 - *API* para entorno *MLFlow*
  - Se exponen los resultados de los modelos explorados en el entorno de *MLFlow* mediante una *API*.
  - Se exponen las comparaciones de los modelos explorados en el entorno de *MLFlow* mediante una *API*.

- **Épica 4 - Documentación y calidad**

- Criterios de aceptación HU9 - Implementación de buenas prácticas
  - Se implementan buenas prácticas de código *Python* en el proyecto.
- Criterios de aceptación HU10 - Documentación
  - Se documentan todos los pasos realizados durante el desarrollo del proyecto.
  - Se valida que cada paso realizado esté correctamente justificado.
- Criterios de aceptación HU11 - Validación de *API* de *MLFlow*
  - Se valida el acceso a los resultados de los modelos mediante la *API* del entorno *MLFlow*.
  - Se valida el acceso a la comparación de los modelos mediante la *API* del entorno *MLFlow*.

8. Fases de CRISP-DM

1. **Comprensión del negocio:**
   - *Objetivo:* predecir si una empresa va a entrar en quiebra o no.
   - *Impacto:* ayudar en la toma de decisiones a empresas especializadas en finanzas, en inversiones, en prestación de seguros, entre otras, permitiéndoles saber si una empresa sobre la que se quiere invertir o a la que se le quiere otorgar un préstamo puede entrar en quiebra o no.
   - *Métricas:* se predice correctamente si la empresa quiebra o no.
2. **Comprensión de los datos**
   - *Tipos de datos:* datos tabulares.
   - *Fuente de datos:* datos publicados por el Taiwan Economic Journal.
   - *Cantidad de datos:* 6819 registros con 96 columnas.
3. **Preparación de los datos**
   - *Transformaciones*
     - Análisis y ajuste de datos atípicos.
     - Análisis y ajuste de datos faltantes.
     - Creación de nuevas variables al combinar las variables existentes.
     - Normalización de datos.
   - *Características clave*
     - Indicador de si la empresa entró en quiebra o variable *target*.

- Una épica es una unidad funcional amplia; una historia de usuario es una funcionalidad concreta; un sprint es una unidad de tiempo donde se ejecutan tareas.
- Las tareas son el nivel más desagregado: permiten estimar tiempos, asignar responsables y monitorear progreso.

Duración sugerida:

- Para un proyecto de 600 h, se recomienda planificar entre 10 y 12 sprints de aproximadamente 2 semanas cada uno.
- Asignar entre 45 y 50 horas efectivas por sprint a tareas técnicas.
- Reservar 100 a 120 h para actividades no técnicas (planificación, escritura, reuniones, defensa).

Importante:

- En proyectos individuales, el responsable suele ser el propio autor.
- Aun así, desagregar tareas facilita el seguimiento y mejora continua.

Conversión opcional de Story Points a horas:

- 1 SP ≈ 2 h como referencia flexible.
- Tener en cuenta aproximaciones tipo Fibonacci.

Cuadro 1. Formato sugerido

Sprint	HU o fase	Tarea	Horas / SP	Responsable	% Completado
Sprint 0	Planificación	Definir alcance y cronograma	10 h	Alumno	100 %
Sprint 0	Planificación	Reunión con tutor/cliente	5 h	Alumno	50 %
Sprint 0	Planificación	Ajuste de entregables	6 h	Alumno	25 %
Sprint 1	HU1	Tarea 1 HU1	6 h / 3 SP	Alumno	0 %
Sprint 1	HU1	Tarea 2 HU1	10 h / 5 SP	Alumno	0 %
Sprint 2	HU2	Tarea 1 HU2	7 h / 5 SP	Alumno	0 %
...	...	...	...	...	...
Sprint 5	Escritura	Redacción memoria	50 h / 34 SP	Alumno	0 %
Sprint 6	Defensa	Preparación exposición	20 h / 13 SP	Alumno	0 %

Recomendaciones:

- Verificar que la carga horaria por sprint sea equilibrada.
- Usar sprints de 1 a 3 semanas, acordes al cronograma general.
- Actualizar el % completado durante el seguimiento del proyecto.
- Considerar un sprint final exclusivo para pruebas, revisión y ajustes antes de la defensa.

- Distintas métricas del desempeño de la empresa a nivel económico y contable.

4. **Modelado**
   - *Tipo de problema:* Clasificación.
   - *Arquitecturas posibles:* Modelos de clasificación como *Logistic Regression*, *Support Vector Machines* y *XGBoost*.
5. **Evaluación del modelo**
   - *F1-score* y *AUC-ROC*.
6. **Despliegue del modelo**
   - Despliegue local usando *MLFlow*.

9. Desglose del trabajo en tareas

- HU1 - Análisis exploratorio (21 h).
  - Identificar variables *categóricas*.
    - *Estimación:* 4 h.
    - *Prioridad:* Media.
  - Identificar variables *numéricas*.
    - *Estimación:* 4 h.
    - *Prioridad:* Media.
  - Graficar la distribución de las variables *numéricas*.
    - *Estimación:* 6 h.
    - *Prioridad:* Media.
  - Realizar análisis de correlaciones entre variables *numéricas*.
    - *Estimación:* 4 h.
    - *Prioridad:* Media.
  - Documentar pasos y decisiones tomadas.
    - *Estimación:* 3 h.
    - *Prioridad:* Media.
- HU2 - Procesamiento de datos faltantes y datos atípicos (64 h).
  - Investigar técnicas de balanceo de clases para algoritmos de clasificación.
    - *Estimación:* 6 h.
    - *Prioridad:* Media.
  - Implementar técnicas de balanceo de clases para algoritmos de clasificación.
    - *Estimación:* 5 h.
    - *Prioridad:* Media.
  - Separar *dataset* en *train* y *test*.
    - *Estimación:* 3 h.
    - *Prioridad:* Media.

.UBAfiuba FACULTAD DE INGENIERÍA
Plan de proyecto del Trabajo Final
Carrera de Especialización en Inteligencia Artificial
Ing. Gaspar Acevedo Zain

12. Normativa y cumplimiento de datos (gobernanza)

En esta sección se debe analizar si los datos utilizados en el proyecto están sujetos a normativas de protección de datos y privacidad, y en qué condiciones se pueden emplear.

Aspectos a considerar:

- Evaluar si los datos están regulados por normativas como GDPR, Ley 25.326 de Protección de Datos Personales en Argentina, HIPAA u otras según jurisdicción y temática.
- Determinar si el uso de los datos requiere consentimiento explícito de los usuarios involucrados.
- Indicar si existen restricciones legales, técnicas o contractuales sobre el uso, compartición o publicación de los datos.
- Aclarar si los datos provienen de fuentes licenciadas, de acceso público o bajo algún tipo de autorización especial.
- Analizar la viabilidad del proyecto desde el punto de vista legal y ético, considerando la gobernanza de los datos.

Este análisis es clave para garantizar el cumplimiento normativo y evitar conflictos legales durante el desarrollo y publicación del proyecto.

13. Gestión de riesgos

a) Identificación de los riesgos (al menos cinco) y estimación de sus consecuencias:

Riesgo 1: detallar el riesgo (riesgo es algo que si ocurre altera los planes previstos de forma negativa)

- Severidad (S): mientras más severo, más alto es el número (usar números del 1 al 10).
  Justificar el motivo por el cual se asigna determinado número de severidad (S).
- Probabilidad de ocurrencia (O): mientras más probable, más alto es el número (usar del 1 al 10).
  Justificar el motivo por el cual se asigna determinado número de (O).

Riesgo 2:

- Severidad (S): X.
  Justificación...
- Ocurrencia (O): Y.
  Justificación...

Riesgo 3:

- Identificar variables con datos faltantes.
  - *Estimación:* 4 h.
  - *Prioridad:* Alta.
- Analizar causas de datos faltantes.
  - *Estimación:* 6 h.
  - *Prioridad:* Alta.
- Corregir datos faltantes.
  - *Estimación:* 8 h.
  - *Prioridad:* Alta.
- Identificar datos con valores atípicos.
  - *Estimación:* 6 h.
  - *Prioridad:* Alta.
- Analizar causas de datos atípicos.
  - *Estimación:* 8 h.
  - *Prioridad:* Alta.
- Graficar variables que presentan de datos atípicos.
  - *Estimación:* 5 h.
  - *Prioridad:* Media.
- Corregir datos atípicos.
  - *Estimación:* 8 h.
  - *Prioridad:* Alta.
- Documentar pasos y decisiones tomadas.
  - *Estimación:* 5 h.
  - *Prioridad:* Media.

- HU3 - *Feature Engineering* (38 h).
  - Identificar variables menos importantes para eliminarlas.
    - *Estimación:* 5 h.
    - *Prioridad:* Alta.
  - Implementar técnicas de eliminación de features.
    - *Estimación:* 5 h.
    - *Prioridad:* Media.
  - Crear nuevas variables mediante combinaciones lineales de variables existentes.
    - *Estimación:* 7 h.
    - *Prioridad:* Alta.
  - Investigar otras técnicas de creación de variables.
    - *Estimación:* 5 h.
    - *Prioridad:* Media.
  - Aplicar otras técnicas de creación de variables.
    - *Estimación:* 8 h.
    - *Prioridad:* Alta.
  - Evaluar nuevas variables en modelos.

.UBAfiuba FACULTAD DE INGENIERÍA
Plan de proyecto del Trabajo Final
Carrera de Especialización en Inteligencia Artificial
Ing. Gaspar Acevedo Zain

- Severidad (S): X.
  Justificación...
- Ocurrencia (O): Y.
  Justificación...

b) Tabla de gestión de riesgos: (El RPN se calcula como RPN=SxO)

Riesgo	S	O	RPN	S*	O*	RPN*

Criterio adoptado:

Se tomarán medidas de mitigación en los riesgos cuyos números de RPN sean mayores a...

Nota: los valores marcados con (*) en la tabla corresponden luego de haber aplicado la mitigación.

c) Plan de mitigación de los riesgos que originalmente excedían el RPN máximo establecido:

Riesgo 1: plan de mitigación (si por el RPN fuera necesario elaborar un plan de mitigación). Nueva asignación de S y O, con su respectiva justificación:

- Severidad (S*): mientras más severo, más alto es el número (usar números del 1 al 10). Justificar el motivo por el cual se asigna determinado número de severidad (S).
- Probabilidad de ocurrencia (O*): mientras más probable, más alto es el número (usar del 1 al 10). Justificar el motivo por el cual se asigna determinado número de (O).

Riesgo 2: plan de mitigación (si por el RPN fuera necesario elaborar un plan de mitigación).

Riesgo 3: plan de mitigación (si por el RPN fuera necesario elaborar un plan de mitigación).

14. Sprint Review

La revisión de sprint (*Sprint Review*) es una práctica fundamental en metodologías ágiles. Consiste en revisar y evaluar lo que se ha completado al finalizar un sprint. En esta instancia, se presentan los avances y se verifica si las funcionalidades cumplen con los criterios de aceptación establecidos. También se identifican entregables parciales y se consideran ajustes si es necesario.

Aunque el proyecto aún se encuentre en etapa de planificación, esta sección permite proyectar cómo se evaluarán las funcionalidades más importantes del backlog. Esta mirada anticipada favorece la planificación enfocada en valor y permite reflexionar sobre posibles obstáculos.

Objetivo: anticipar cómo se evaluará el avance del proyecto a medida que se desarrollen las funcionalidades, utilizando como base al menos cuatro historias de usuario del *Product Backlog*.

- *Estimación:* 5 h.
- *Prioridad:* Alta.

- Documentar pasos y decisiones tomadas.
  - *Estimación:* 3 h.
  - *Prioridad:* Baja.

- HU4 - Implementación de modelos de *Machine Learning* (89 h).
  - Implementar código de validación cruzada para *Logistic Regression*.
    - *Estimación:* 4 h.
    - *Prioridad:* Media.
  - Implementar modelo *Logistic Regression*, sin considerar *feature engineering*.
    - *Estimación:* 6 h.
    - *Prioridad:* Alta.
  - Evaluar modelo *Logistic Regression*, sin considerar *feature engineering*.
    - *Estimación:* 4 h.
    - *Prioridad:* Media.
  - Implementar modelo *Logistic Regression*, considerando *feature engineering*.
    - *Estimación:* 6 h.
    - *Prioridad:* Alta.
  - Evaluar modelo *Logistic Regression*, considerando *feature engineering*.
    - *Estimación:* 4 h.
    - *Prioridad:* Media.
  - Implementar código de validación cruzada para *SVM*.
    - *Estimación:* 4 h.
    - *Prioridad:* Media.
  - Implementar modelo *SVM*, sin considerar *feature engineering*.
    - *Estimación:* 6 h.
    - *Prioridad:* Alta.
  - Evaluar modelo *SVM*, sin considerar *feature engineering*.
    - *Estimación:* 4 h.
    - *Prioridad:* Media.
  - Implementar modelo *SVM*, considerando *feature engineering*.
    - *Estimación:* 6 h.
    - *Prioridad:* Alta.
  - Evaluar modelo *SVM*, considerando *feature engineering*.
    - *Estimación:* 4 h.
    - *Prioridad:* Media.
  - Implementar código de validación cruzada para *XGBoost*.
    - *Estimación:* 4 h.
    - *Prioridad:* Media.
  - Implementar modelo *XGBoost*, sin considerar *feature engineering*.
    - *Estimación:* 8 h.

Seleccionar al menos 4 HU del Product Backlog. Para cada una, completar la siguiente tabla de revisión proyectada:

Formato sugerido:

HU seleccionada	Tareas asociadas	Entregable esperado	¿Cómo sabrás que está cumplida?	Observaciones o riesgos
HU1	Tarea 1 Tarea 2	Módulo funcional	Cumple criterios de aceptación definidos	Falta validar con el tutor
HU3	Tarea 1 Tarea 2	Reporte generado	Exportación disponible y clara	Requiere datos reales
HU5	Tarea 1 Tarea 2	Panel de gestión	Roles diferenciados operativos	Riesgo en integración
HU7	Tarea 1 Tarea 2	Informe trimestral	PDF con gráficos y evolución	Puede faltar tiempo para ajustes

15. Sprint Retrospective

La retrospectiva de sprint es una práctica orientada a la mejora continua. Al finalizar un sprint, el equipo (o el alumno, si trabaja de forma individual) reflexiona sobre lo que funcionó bien, lo que puede mejorarse y qué acciones concretas pueden implementarse para trabajar mejor en el futuro.

Durante la cursada se propuso el uso de la **Estrella de la Retrospectiva**, que organiza la reflexión en torno a cinco ejes:

- ¿Qué hacer más?
- ¿Qué hacer menos?
- ¿Qué mantener?
- ¿Qué empezar a hacer?
- ¿Qué dejar de hacer?

Aun en una etapa temprana, esta herramienta permite que el alumno planifique su forma de trabajar, identifique anticipadamente posibles dificultades y diseñe estrategias de organización personal.

Objetivo: reflexionar sobre las condiciones iniciales del proyecto, identificando fortalezas, posibles dificultades y estrategias de mejora, incluso antes del inicio del desarrollo.

Completar la siguiente tabla tomando como referencia los cinco ejes de la Estrella de la Retrospectiva (*Starfish* o estrella de mar). Esta instancia te ayudará a definir buenas prácticas

- *Prioridad:* Alta.
- Evaluar modelo *XGBoost*, sin considerar *feature engineering*.
  - *Estimación:* 6 h.
  - *Prioridad:* Media.
- Implementar modelo *XGBoost*, considerando *feature engineering*.
  - *Estimación:* 8 h.
  - *Prioridad:* Alta.
- Evaluar modelo *XGBoost*, considerando *feature engineering*.
  - *Estimación:* 6 h.
  - *Prioridad:* Media.
- Persistir modelos en *GitHub*
  - *Estimación:* 4 h
  - *Prioridad:* Media
- Documentar pasos y decisiones tomadas.
  - *Estimación:* 5 h.
  - *Prioridad:* Media.

- HU5 - Optimización de hiperparámetros (118 h).
  - Identificar hiperparámetros y rangos de *Logistic Regression*.
    - *Estimación:* 5 h.
    - *Prioridad:* Media.
  - Optimizar hiperparámetros de *Logistic Regression*, sin considerar *feature engineering*.
    - *Estimación:* 6 h.
    - *Prioridad:* Media.
  - Implementar hiperparámetros más óptimos en *Logistic Regression*, sin considerar *feature engineering*.
    - *Estimación:* 4 h.
    - *Prioridad:* Media.
  - Evaluar modelo de *Logistic Regression* con hiperparámetros óptimos, sin considerar *feature engineering*.
    - *Estimación:* 4 h.
    - *Prioridad:* Media.
  - Optimizar hiperparámetros de *Logistic Regression*, considerando *feature engineering*.
    - *Estimación:* 6 h.
    - *Prioridad:* Media.
  - Implementar hiperparámetros más óptimos en *Logistic Regression*, considerando *feature engineering*.
    - *Estimación:* 4 h.
    - *Prioridad:* Media.
  - Evaluar modelo de *Logistic Regression* con hiperparámetros óptimos, considerando *feature engineering*.
    - *Estimación:* 4 h.
    - *Prioridad:* Media.

.UBAfiuba FACULTAD DE INGENIERÍA
Plan de proyecto del Trabajo Final
Carrera de Especialización en Inteligencia Artificial
Ing. Gaspar Acevedo Zain

desde el inicio y prepararte para enfrentar el trabajo de forma organizada y flexible. Se deberá completar la tabla al menos para 3 sprints técnicos y 1 no técnico.

Formato sugerido:

Sprint tipo y N°	**¿Qué hacer más?**	**¿Qué hacer menos?**	**¿Qué mantener?**	**¿Qué empezar a hacer?**	**¿Qué dejar de hacer?**
Sprint técnico - 1	Validaciones continuas con el alumno	Cambios sin versión registrada	Pruebas con datos simulados	Documentar cambios propuestos	Ajustes sin análisis de impacto
Sprint técnico - 2	Verificar configuraciones en múltiples escenarios	Modificar parámetros sin guardar historial	Perfiles reutilizables	Usar logs para configuración	Repetir pruebas manuales innecesarias
Sprint técnico - 8	Comparar correlaciones con casos previos	Cambiar parámetros sin justificar	Revisión cruzada de métricas	Anotar configuraciones usadas	Trabajar sin respaldo de datos
Sprint no técnico - 12 (por ej.: "Defensa")	Ensayos orales con feedback	Cambiar contenidos en la memoria	Material visual claro	Dividir la presentación por bloques	Agregar gráficos difíciles de explicar