Predicción de series de tiempo mediante Deep Learning y aplicación en datos demográficos

Autor:
Lic. Diego José Araujo Arellano

Director:
Título y Nombre del director (FIUBA)

Esta planificación fue realizada en el curso de Gestión de proyectos entre el 20 de agosto de 2024 y el 08 de octubre de 2024.

Índice

Registros de cambios

Revisión	Detalles de los cambios realizados	Fecha
0	Creación del documento	20 de agosto de 2024
1	Se completa hasta el punto 5 inclusive	01 de setiembre de 2024
2	Se completa hasta el punto 9 inclusive Se realizan correcciones de entrega anterior	09 de setiembre de 2024

Acta de constitución del proyecto

Buenos Aires, 20 de agosto de 2024

Por medio de la presente se acuerda con el Lic. Diego José Araujo Arellano que su Trabajo Final de la Carrera de Especialización en Inteligencia Artificial se titulará "Predicción de series de tiempo mediante Deep Learning y aplicación en datos demográficos" y consistirá en desarrollar y evaluar un modelo de predicción de series de tiempo basado en técnicas de deep learning, para su aplicación en problemas demográficos relevantes. El trabajo tendrá un presupuesto preliminar estimado de 600 horas de trabajo, con fecha de inicio el 20 de agosto de 2024 y fecha de presentación pública el 15 de mayo de 2025.

Se adjunta a esta acta la planificación inicial.

Dr. Ing. Ariel Lutenberg Director posgrado FIUBA	Nombre del cliente FIUBA

Título y Nombre del director
Director del Trabajo Final

1. Descripción técnica-conceptual del proyecto a realizar

La predicción de series de tiempo es una herramienta fundamental en numerosos campos, entre ellos la demografía. Un tema relevante de analizar en este ámbito, es el de la Tasa Global de Fecundidad (TGF). Este indicador no solo afecta el crecimiento poblacional, sino que también influye en la estructura de edad de la población, la oferta laboral futura, y la demanda de servicios como educación y salud. Cambios en la TGF pueden tener implicaciones significativas para la planificación de políticas públicas, ya que un aumento o disminución abrupta en la TGF puede alterar la demanda de recursos y servicios en el corto y largo plazo. Por lo tanto, una predicción precisa de dicha tasa es esencial para una planificación demográfica y socioeconómica eficaz.

El valor de la TGF se interpreta como el número de hijos que, en promedio, tendría cada mujer de una cohorte hipotética de mujeres no expuestas al riesgo de muerte desde el inicio hasta el fin del periodo fértil y que, a partir del momento en que se inicia la reproducción, están expuestas a las tasas de fecundidad por edad de la población en estudio.

Son varios los métodos clásicos que se utilizan para predecir series temporales, por ejemplo: modelo ARIMA, regresiones lineales y no lineales, entre otros. Sin embargo estos métodos por lo general presentan grandes limitaciones, ya que a menudo realizan supuestos fuertes sobre los datos (por ejemplo, que sean estacionarios) o pueden no capturar bien la complejidad de las relaciones entre la TGF y los factores predictivos. En este sentido, se torna relevante poder utilizar técnicas de deep learning que puedan superar estas restricciones.

Últimamente los modelos seq2seq se han hecho populares en el tratamiento de datos secuenciales o temporales en el contexto de *deep learning*. Si bien en un inicio se utilizaban principalmente para procesamiento de lenguaje natural (NLP) por ejemplo, en traducción de textos, se ha demostrado que se pueden adaptar al análisis de series temporales, para obtener mejores predicciones con respecto a los modelos clásicos.

Los modelos *seq2seq* por lo general se componen de un codificador (*encoder*) que genera una nueva representación de los datos, conocida como "vector de contexto" y un decodificador (*decoder*) que recibe como entrada ese vector y genera otra secuencia de salida. En la Figura 1 se puede ver una abstracción de este proceso.

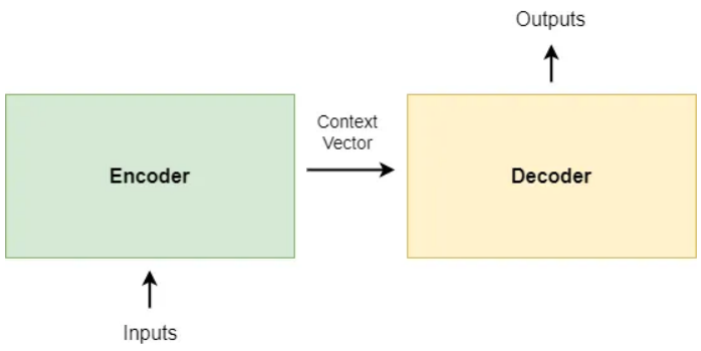

Figura 1. Diagrama en bloques del sistema.

Una ventaja de esta arquitectura es que permite que los componentes internos del modelo (*encoder-decoder*), pueden variar de recurrentes, a convolucionales, a transformers, o incluso a modelos híbridos.

En la Figura 2 se muestra está lógica en el contexto de series de tiempo. Se observa que el codificador toma una secuencia (o ventana) de series de tiempo y el decodificador intenta predecir múltiples pasos hacia adelante (n-pasos).

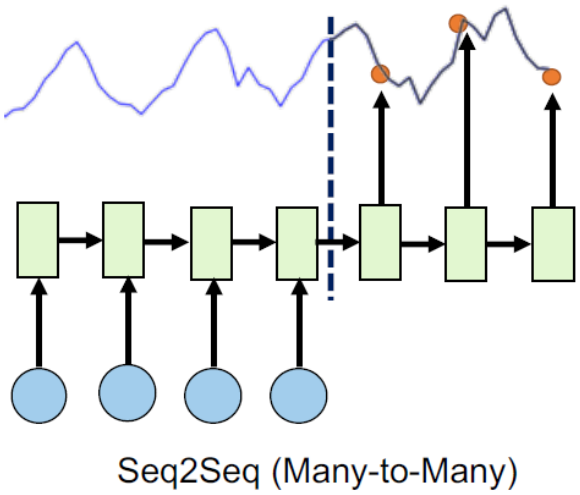

Figura 2. Proceso seq2seq en series de tiempo.

El presente trabajo se enmarca como un emprendimiento personal y que, si bien está enfocado a su aplicación en datos demográficos, es extrapolable a cualquier otro ámbito donde se necesite realizar predicciones de series temporales de manera más precisa.

2. Identificación y análisis de los interesados

En el cuadro 1 se puede ver una lista de los interesados en el proyecto, su rol de participación, nombre y apellido, organización a la cuál pertenece y puesto.

Rol	Nombre y Apellido	Organización	Puesto
Cliente	Miembros del jurado	FIUBA	-
Responsable	Lic. Diego José Araujo Arellano	FIUBA	Alumno
Orientador	Título y Nombre del director	FIUBA	Director del Trabajo Final
Usuario final	Demógrafos, investigadores, analistas	-	-

A continuación, se brinda detalle de nivel de participación de cada uno de los interesados mencionados en el cuadro previo:

- Cliente: serán los encargados de evaluar el resultado del proyecto y que los trabajos realizados fueron acordes para lograr el objetivo propuesto.
- Responsable: es el encargado del proyecto, trabajará siguiendo la planificación establecida para lograr el objetivo establecido.

- Orientador: participará en la definición de objetivos y cumplirá un rol de consultor de las técnicas utilizadas a lo largo del proyecto.
- Usuario Final: serán aquellas personas que puedan hacer usufructo de las técnicas y resultados obtenidos en este proyecto.

3. Propósito del proyecto

Este proyecto se basa en un emprendimiento personal. Su objetivo principal es desarrollar y evaluar un modelo de predicción de series de tiempo basado en técnicas de *deep learning*. Se propone utilizar un modelo de secuencia a secuencia (*seq2seq*) para poder predecir, más específicamente, tasas de fecundidad en distintos países. La intención es mostrar las ventajas que presenta este algoritmo en el abordaje de un problema demográfico relevante y poder medir su potencial y efectividad.

4. Alcance del proyecto

Dentro del alcance del proyecto se incluye:

- Análisis, estudio y evaluación de la técnica *seq2seq*, originalmente utilizada a un problema de procesamiento de lenguaje natural, adaptando al problema de la predicción de series de tiempo.
- Escribir todo el código necesario en un lenguaje de programación seleccionado para poder llevar a cabo el modelo. Este una vez concluido, quedará disponible en un repositorio para su consulta, junto a ejemplos reproducibles presentados en el proyecto.
- El armado del dataset necesario para llevar a cabo el entrenamiento y la validación del modelo, que se conformará no soolo de series históricas de la variable objetivo (TGF), sino de covariables que puedan incidir en la predicción de la misma.
- Presentar los resultados, evidenciando ventajas e inconvenientes de aplicar este modelo.

El presente proyecto no contempla la profundización en modelos clásicos de predicción de series de tiempo (por ejemplo ARIMA), ni comparación con los mismos, salvo mención en características y limitaciones ampliamente conocidas.

Los puntos no incluidos en el alcance del proyecto serán evaluados en coordinación del alumno con el orientador.

5. Supuestos del proyecto

Para el desarrollo del presente proyecto se supone que:

- Se contará con la base teórica necesaria y aprendida a lo largo de la Carrera de Especialización en Inteligencia Artificial para poder implementar el modelo aplicado a este problema.
- A su vez se contará con la base teórica mínima en el problema demográfico de fondo, para poder validar resultados.
- Se contará con el equipo necesario para poder realizar la creación del modelo y los conocimientos técnicos en el lenguaje de programación utilizado, para llevar a cabo el mismo.
- De existir procesos que requieran GPU, podrán ser ejecutados en herramientas virtuales y gratuitas tales como Google Colaboratory.
- Se dispondrá del tiempo necesario para poder realizar el proyecto y en todo caso estará ajustado a lo planificado sin mayor extensión del mismo.
- El orientador brindará apoyo y dispondrá del tiempo necesario para poder guiar en cuestiones técnicas del proyecto.

6. Requerimientos

1. Requerimientos funcionales:
   1.1. El modelo de predicción de la TGF debe ser capaz de predecir con un error menor al 10 % las tasas futuras a partir de los datos de series de tiempo históricos y covariables relevantes seleccionadas.
   1.2. El sistema debe ser capaz de procesar y normalizar de forma automatizada los datos de entrada, asegurando que estén listos para el modelo de predicción sin tener que intervenir manualmente.
2. Requerimientos de documentación:
   2.1. Los resultados del modelo seq2seq deben ser validados y aprobados por los usuarios finales y expertos en demografía antes de su implementación en entornos reales.
   2.2. El modelo debe ser probado en diferentes escenarios, incluyendo datos de series de tiempo de diferentes países y períodos, para asegurar estabilidad y robustez en la predicción.
3. Requerimiento de testing
   3.1. El código desarrollado deberá estar correctamente comentado y estructurado, para facilitar su comprensión y reproducibilidad por parte de terceros.
   3.2. La documentación generada deberá incluir una guía detallada para la configuración y ejecución del modelo, así como un repositorio en GitHub que contenga todo el código, datos de ejemplo y una guía de usuario para la correcta interpretación de los resultados.

7. Historias de usuarios (*Product backlog*)

Las historias de usuarios se ponderan en base a las siguientes categorías:

- Cantidad de trabajo requerido para completar la tarea.
- Complejidad del trabajo a realizar.
- Incertidumbre asociada a la actividad, es decir el nivel de riesgo que involucra realizar la tarea.

A cada categoría se le asignaran valores de ponderación entre 1 (bajo) y 5 (alto), siendo 3 el valor medio. A las historias de usuario se les asigna un peso igual a la suma de cada categoría y luego se reemplazara por el número de la serie de Fibonacci siguiente inmediato, de forma de aproximar la estimación de velocidad y salida del trabajo.

- Como investigador interesado en la predicción de series de tiempo, quiero una introducción clara y detallada sobre los modelos seq2seq y sus aplicaciones, para entender cómo estos modelos pueden mejorar las predicciones en contextos demográficos y otros campos.
  - Dificultad: 1
  - Complejidad: 1
  - Incertidumbre: 2
  - **Total:** 4 | **Ponderación final:** 5 *story points*

- Como analista de datos, quiero un conjunto de datos procesados que incluya la TGF y covariables relevantes que incidan en la misma, para asegurar que los datos están listos para ser utilizados en el modelo y maximizar la precisión.
  - Dificultad: 3
  - Complejidad: 4
  - Incertidumbre: 5
  - **Total:** 12 | **Ponderación final:** 13 *story points*

- Como demógrafo, quiero un análisis de las predicciones y su interpretación en un contexto demográfico, para entender las implicaciones de los resultados y evaluar su aplicación en la planificación de políticas públicas.
  - Dificultad: 3
  - Complejidad: 3
  - Incertidumbre: 3
  - **Total:** 9 | **Ponderación final:** 13 *story points*

- Como analista de modelos predictivos, quiero evaluar el rendimiento del modelo utilizando métricas apropiadas, para verificar la efectividad del modelo y compararlo con otros enfoques clásicos.

- Dificultad: 2
- Complejidad: 3
- Incertidumbre: 3
- **Total:** 8 | **Ponderación final:** 8 *story points*

- Como investigador, quiero poder acceder al código y tener documentación de todas las fases del proyecto, además de reporte final con conclusiones y recomendaciones, para ver los resultados de manera efectiva y replicar en futuras investigaciones.
  - Dificultad: 2
  - Complejidad: 2
  - Incertidumbre: 3
  - **Total:** 7 | **Ponderación final:** 8 *story points*

8. Entregables principales del proyecto

Los entregables del proyecto son:

- Código del modelo.
- Plan del proyecto.
- Informe de avance.
- Memoria del trabajo final.

9. Desglose del trabajo en tareas

1. Planificación y gestión del proyecto (60 h)
   1.1. Descripción del proyecto. (5 h)
   1.2. Definición de alcance y requerimientos. (10 h)
   1.3. Planificación de recursos, cronograma y gestión de Riesgos. (25 h)
   1.4. Seguimiento y control del proyecto. (20 h)
2. Selección y preprocesamiento de datos (100 h)
   2.1. Recolección de datos y selección de covariables. (40 h)
   2.2. Limpieza, transformación y normalización de datos. (40 h)
   2.3. Análisis Exploratorio de los datos (EDA). (20 h)
3. Desarrollo del modelo seq2seq (120 h)
   3.1. Investigación sobre arquitecturas y técnicas de deep learning. (40 h)
   3.2. Diseño y configuración del modelo seq2seq con encoder-decoder. (40 h)
   3.3. Implementación de modelos LSTM y GRU, y mecanismos de atención. (40 h)

4. Entrenamiento y evaluación del modelo (120 h)
   4.1. Preparación de datos para entrenamiento, validación y testeo. (40 h)
   4.2. Evaluación del modelo con métricas clave y validación cruzada. (40 h)
   4.3. Ajuste de hiperparámetros y optimización del modelo. (40 h)
5. Análisis de Resultados (65 h)
   5.1. Análisis de resultados e interpretación en contexto demográfico. (30 h)
   5.2. Revisión de resultados y ajustes finales del modelo. (35 h)
6. Documentación y Reporte Final (95 h)
   6.1. Documentación técnica del modelo y proceso completo. (30 h)
   6.2. Creación de un pepositorio en GitHub para publicar código y documentación. (30 h)
   6.3. Generación de reportes detallados de resultados y conclusiones. (35 h)
7. Presentación del trabajo (50 h)
   7.1. Armar presentación. (20 h)
   7.2. Preparación para la defensa pública. (30 h)

Cantidad de horas total: 610.

10. Diagrama de Activity On Node

Armar el AoN a partir del WBS definido en la etapa anterior.

Una herramienta simple para desarrollar los diagramas es el Draw.io (`https://app.diagrams.net/`). Draw.io

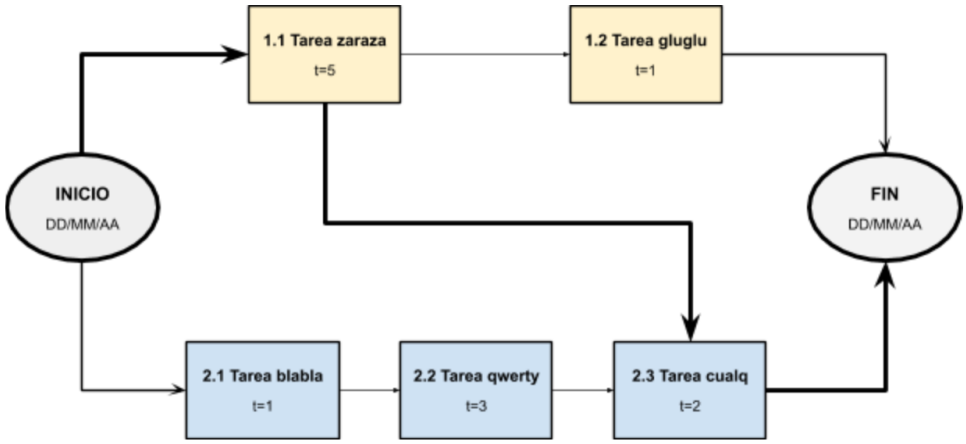

Figura 3. Diagrama de *Activity on Node*.

Indicar claramente en qué unidades están expresados los tiempos. De ser necesario indicar los caminos semi críticos y analizar sus tiempos mediante un cuadro. Es recomendable usar colores y un cuadro indicativo describiendo qué representa cada color.

11. Diagrama de Gantt

Existen muchos programas y recursos *online* para hacer diagramas de Gantt, entre los cuales destacamos:

- Planner
- GanttProject
- Trello + *plugins*. En el siguiente link hay un tutorial oficial:
  `https://blog.trello.com/es/diagrama-de-gantt-de-un-proyecto`
- Creately, herramienta online colaborativa.
  `https://creately.com/diagram/example/ieb3p3ml/LaTeX`
- Se puede hacer en latex con el paquete *pgfgantt*
  `http://ctan.dcc.uchile.cl/graphics/pgf/contrib/pgfgantt/pgfgantt.pdf`

Pegar acá una captura de pantalla del diagrama de Gantt, cuidando que la letra sea suficientemente grande como para ser legible. Si el diagrama queda demasiado ancho, se puede pegar primero la "tabla" del Gantt y luego pegar la parte del diagrama de barras del diagrama de Gantt.

Configurar el software para que en la parte de la tabla muestre los códigos del EDT (WBS).
Configurar el software para que al lado de cada barra muestre el nombre de cada tarea.
Revisar que la fecha de finalización coincida con lo indicado en el Acta Constitutiva.

En la figura 4, se muestra un ejemplo de diagrama de gantt realizado con el paquete de *pgfgantt*. En la plantilla pueden ver el código que lo genera y usarlo de base para construir el propio.

Las fechas pueden ser calculadas utilizando alguna de las herramientas antes citadas. Sin embargo, el siguiente ejemplo fue elaborado utilizando esta hoja de cálculo.

Es importante destacar que el ancho del diagrama estará dado por la longitud del texto utilizado para las tareas (Ejemplo: tarea 1, tarea 2, etcétera) y el valor *x unit*. Para mejorar la apariencia del diagrama, es necesario ajustar este valor y, quizás, acortar los nombres de las tareas.

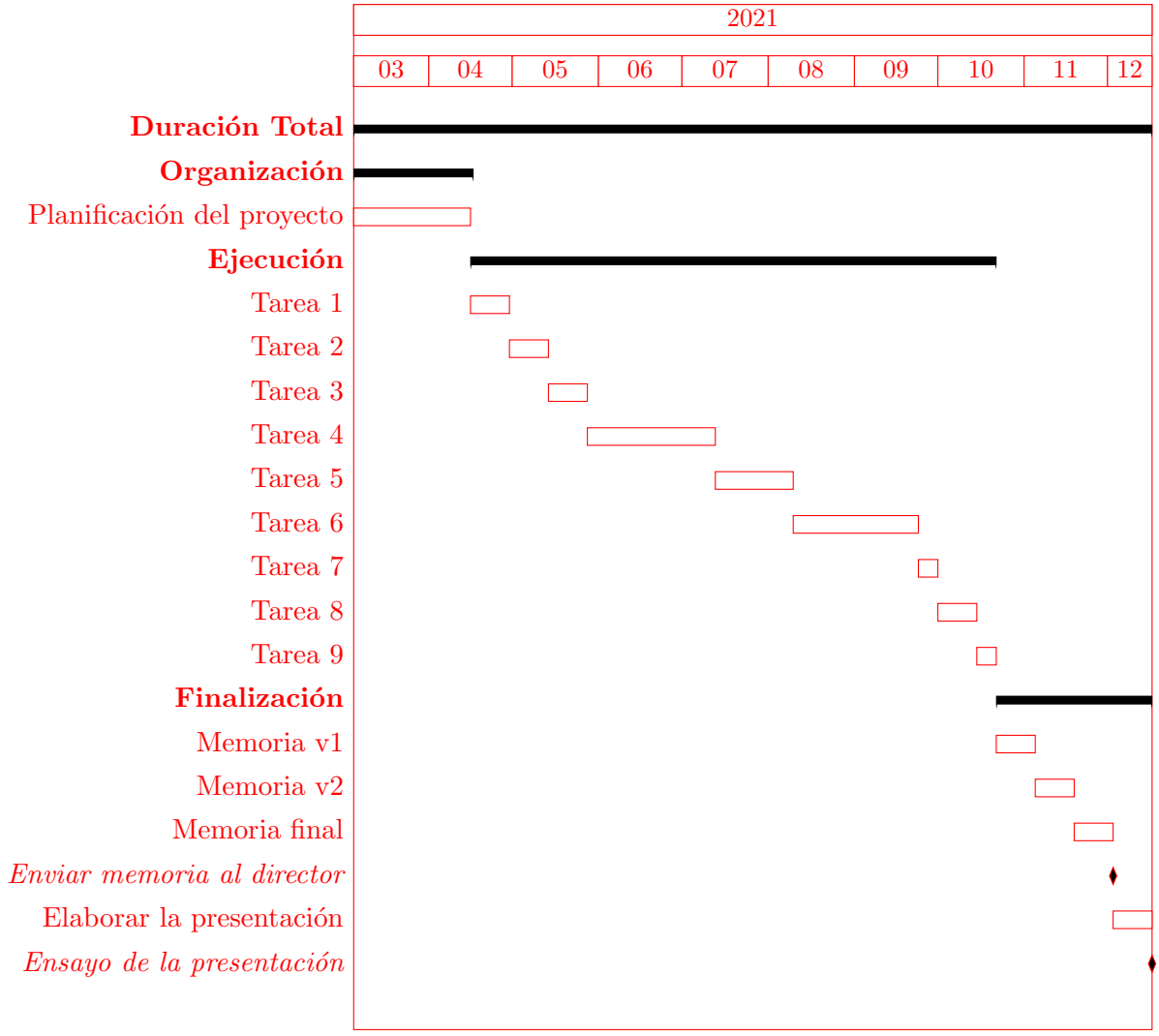

Figura 4. Diagrama de gantt de ejemplo

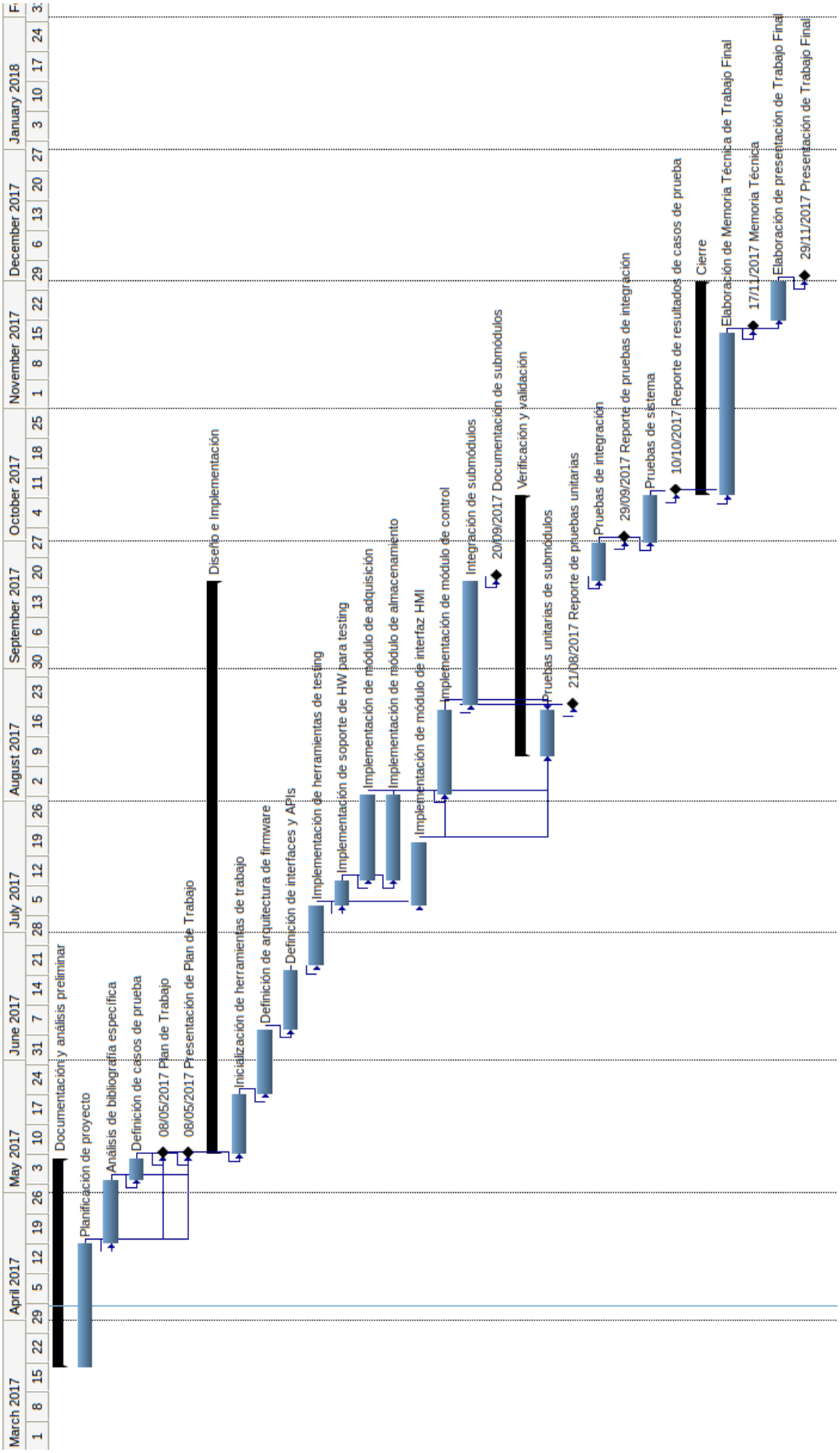

Figura 5. Ejemplo de diagrama de Gantt (apaisado).

12. Presupuesto detallado del proyecto

Si el proyecto es complejo entonces separarlo en partes:

- Un total global, indicando el subtotal acumulado por cada una de las áreas.
- El desglose detallado del subtotal de cada una de las áreas.

IMPORTANTE: No olvidarse de considerar los COSTOS INDIRECTOS.

Incluir la aclaración de si se emplea como moneda el peso argentino (ARS) o si se usa moneda extranjera (USD, EUR, etc). Si es en moneda extranjera se debe indicar la tasa de conversión respecto a la moneda local en una fecha dada.

COSTOS DIRECTOS			
Descripción	Cantidad	Valor unitario	Valor total
SUBTOTAL			
COSTOS INDIRECTOS			
Descripción	Cantidad	Valor unitario	Valor total
SUBTOTAL			
TOTAL			

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:

- 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. Gestión de la calidad

Elija al menos diez requerimientos que a su criterio sean los más importantes/críticos/que aportan más valor y para cada uno de ellos indique las acciones de verificación y validación que permitan asegurar su cumplimiento.

- Req #1: copiar acá el requerimiento con su correspondiente número.

- Verificación para confirmar si se cumplió con lo requerido antes de mostrar el sistema al cliente. Detallar.
- Validación con el cliente para confirmar que está de acuerdo en que se cumplió con lo requerido. Detallar.

Tener en cuenta que en este contexto se pueden mencionar simulaciones, cálculos, revisión de hojas de datos, consulta con expertos, mediciones, etc.

Las acciones de verificación suelen considerar al entregable como "caja blanca", es decir se conoce en profundidad su funcionamiento interno.

En cambio, las acciones de validación suelen considerar al entregable como "caja negra", es decir, que no se conocen los detalles de su funcionamiento interno.

15. Procesos de cierre

Establecer las pautas de trabajo para realizar una reunión final de evaluación del proyecto, tal que contemple las siguientes actividades:

- Pautas de trabajo que se seguirán para analizar si se respetó el Plan de Proyecto original:
  - Indicar quién se ocupará de hacer esto y cuál será el procedimiento a aplicar.
- Identificación de las técnicas y procedimientos útiles e inútiles que se emplearon, los problemas que surgieron y cómo se solucionaron:
  - Indicar quién se ocupará de hacer esto y cuál será el procedimiento para dejar registro.
- Indicar quién organizará el acto de agradecimiento a todos los interesados, y en especial al equipo de trabajo y colaboradores:
  - Indicar esto y quién financiará los gastos correspondientes.