PROYECTO FINAL DE INGENIERÍA

PREDICTFLOW: PREDICTOR DE ASISTENCIA A ENCUENTROS FUTBOLÍSTICOS DE LA LIGA PROFESIONAL DE FÚTBOL ARGENTINO EN 2025

Bussolini, Tomas - LU1123735
Ingeniería Informática

Estarli, Juan - LU1123938
Ingeniería Informática

Tutor:
Sabatino, Pablo Luis Esteban, UADE

16 de diciembre de 2025

UNIVERSIDAD ARGENTINA DE LA EMPRESA
FACULTAD DE INGENIERÍA Y CIENCIAS EXACTAS

Agradecimientos

El desarrollo del presente trabajo no habría sido posible sin el acompañamiento y la colaboración de todas las personas que formaron parte de este proceso.

De manera muy especial, queremos agradecer a nuestras familias y parejas, por su apoyo incondicional, contención y por brindarnos la motivación necesaria durante la carrera.

A nuestro tutor, Pablo Luis Esteban Sabatino, por su confianza, orientación y dedicación durante el desarrollo del proyecto.

A nuestros compañeros, quienes nos han acompañado desde el primer día, se han convertido en amigos y han sido un pilar fundamental para conseguir nuestros objetivos.

Finalmente, al cuerpo docente y a la Universidad Argentina de la Empresa (UADE) por habernos formado con excelencia, fomentando en nosotros los valores y conocimientos necesarios para desarrollarnos profesionalmente.

Resumen

El fútbol constituye uno de los movimientos culturales más representativos de la República Argentina y, al mismo tiempo, un desafío organizativo para quienes deben garantizar el desarrollo de dichos eventos. A pesar de la relevancia económica y social, los clubes y autoridades gubernamentales continúan realizando estimaciones empíricas sobre la cantidad de asistentes, ya que no cuentan con herramientas que respalden la toma de decisiones de manera objetiva. Esta limitación suele derivar en una asignación ineficiente de recursos: la sobreestimación genera gastos operativos innecesarios, mientras que la subestimación compromete la seguridad, la logística y la experiencia del espectador.

En respuesta a esta problemática, PredictFlow surge como una plataforma web orientada a prever la asistencia de público a los partidos de fútbol y estimar los recursos necesarios para su organización. La solución implementa modelos de aprendizaje automático entrenados con datos históricos y contextuales, capaces de anticipar los niveles de asistencia y aportar información de valor para la planificación de recursos logísticos, operativos y de seguridad.

La plataforma busca asistir a los organizadores de encuentros deportivos en la toma de decisiones informadas, precisas y eficientes, optimizando el uso de recursos y contribuyendo a mejorar la calidad del espectáculo. En su etapa inicial, el sistema se aplica a los partidos de la Liga Profesional de Fútbol de la Argentina y se encuentra actualmente en fase de pruebas. A futuro, se proyecta su expansión hacia otras ligas nacionales y competiciones con características similares.

Abstract

Football is one of the most representative cultural movements in Argentina and, at the same time, an organizational challenge for those responsible for ensuring the proper development of such events. Despite their economic and social relevance, clubs and governmental authorities still rely on empirical estimations to determine attendance, as they lack tools that support objective decision-making. This limitation often leads to inefficient resource allocation: overestimations generate unnecessary operational expenses, while underestimations compromise safety, logistics, and the spectator experience.

In response to this issue, PredictFlow emerges as a web-based platform designed to forecast football match attendance and estimate the resources required for event organization. The solution implements machine learning models trained on historical and contextual data, capable of anticipating attendance levels and providing valuable insights for the planning of logistical, operational, and security resources.

The platform aims to assist sports event organizers in making informed, accurate, and efficient decisions, optimizing resource usage and contributing to improving the overall quality of the event. In its initial stage, the system focuses on matches of the Liga Profesional de Fútbol and is currently in a testing phase. In the future, it is expected to expand to other national leagues and competitions with similar characteristics.

Contenidos

1. Introducción

Anticipar la cantidad de público que asistirá a un encuentro futbolístico no es solo una curiosidad o un dato estadístico, sino una necesidad concreta para quienes organizan, gestionan y garantizan el correcto desarrollo de ese evento. En particular, en países como Argentina, donde el fútbol tiene un gran peso cultural, económico y social, contar con herramientas que permitan predecir la asistencia a los partidos puede hacer una gran diferencia en la planificación y toma de decisiones.

Conocer de antemano cuántas personas irán a un partido permite, por ejemplo, planificar de manera eficiente el operativo de seguridad, prever el uso del transporte público, gestionar los accesos al estadio, distribuir correctamente el personal de atención o incluso estimar ingresos por venta de entradas y consumos asociados. No se trata solo de mejorar la experiencia del espectador, sino también de optimizar la gestión de los recursos utilizados, aumentando la eficiencia de los mismos.

En este contexto, el uso de herramientas de inteligencia artificial, y en particular de técnicas de machine learning, dan lugar a soluciones innovadoras que prometen mejores resultados. Estas técnicas permiten analizar grandes volúmenes de datos, relacionarlos con ciertas variables dinámicas y descubrir patrones que no siempre son evidentes en los análisis tradicionales. Su aplicación en el ámbito futbolístico ha crecido en los últimos años, demostrando a su vez ser útil para anticipar la asistencia a los partidos y permitir a los clubes optimizar la planificación de recursos y mejorar la experiencia de los aficionados (PERUZZO y ASANI, 2024).

1.1. Objetivos

Este proyecto tiene como objetivo principal desarrollar una plataforma digital que permita anticipar la asistencia a partidos de la Liga Profesional de Fútbol Argentino durante la temporada 2025, con el fin de optimizar la planificación de recursos en cada encuentro.

Para alcanzar ese objetivo general, se plantean una serie de metas específicas:

1. Analizar los procesos actuales de planificación y gestión de recursos utilizados por los clubes.

2. Identificar y evaluar los factores determinantes que influyen en la cantidad de público asistente a los partidos.
3. Recopilar datos históricos de asistencia correspondientes al período 2020-2025.
4. Desarrollar un modelo predictivo basado en técnicas de machine learning que estime la demanda de asistencia.
5. Implementar una aplicación web que visualice las predicciones, permita simular escenarios y facilite la toma de decisiones sobre la asignación de recursos.

1.2. Estructura

En cuanto a la estructura, el trabajo se organiza en distintas secciones. Primero se presentan los Antecedentes, donde se explica el contexto del problema y por qué resulta relevante predecir la asistencia a los encuentros futbolísticos.

Luego, el Marco Teórico introduce los conceptos principales relacionados con la inteligencia artificial y el aprendizaje automático, que son la base teórica para desarrollar el modelo.

Después, en el Estado del Arte, se investigan trabajos previos y estudios similares realizados en contextos internacionales, regionales, nacionales y locales, lo que permite situar a la propuesta dentro de un marco de referencia y analizar las soluciones competidoras.

En la sección de User Research, se incluyen las entrevistas realizadas con actores relacionados con la gestión de los encuentros futbolísticos, que ayudan a entender cómo se toman actualmente las decisiones y qué necesidades podrían cubrirse con esta herramienta. A su vez, se presenta una encuesta destinada a los asistentes que suelen concurrir a estos encuentros, para poder relevar los puntos de dolor que presentan respecto a la organización.

Luego, en la Descripción se detalla el proceso que siguen los datos a lo largo del pipeline, desde su etapa de extracción inicial hasta su utilización en el entrenamiento de los modelos. Se explican también las variables consideradas en el proceso, las conclusiones obtenidas a partir de los entrenamientos realizados, la arquitectura conceptual de la solución propuesta, los atributos de calidad que la caracterizan y el plan de despliegue en la nube.

Posteriormente, se incorpora la de Metodologías de desarrollo, donde se describe el enfoque ágil adoptado y la organización del trabajo a través de iteraciones y entregas incrementales. Luego, el Modelo de negocio analiza la propuesta de valor, las fuentes de ingreso y la viabilidad económica del proyecto en distintos escenarios.

Por último, en la sección de Pruebas se realizan validaciones con los distintos modelos, así como también pruebas funcionales de la solución.

2. Antecedentes

La organización y gestión de eventos futbolísticos representa un desafío constante para los clubes y entidades responsables, debido a la necesidad de garantizar la seguridad, la comodidad y la experiencia general de miles de asistentes. Estas tareas históricamente se han basado en estimaciones empíricas o experiencia previa, lo que puede dar lugar a una asignación inadecuada de recursos o a deficiencias en la operación del evento. Sin embargo, el avance de las tecnologías de la información, y en particular de la Inteligencia Artificial, ha abierto nuevas posibilidades para realizar estimaciones más precisas sobre la asistencia esperada. Esto permite una planificación más eficiente de recursos humanos, logísticos y de seguridad, reduciendo la incertidumbre y mejorando la toma de decisiones.

2.1. Marco Teórico

El presente marco teórico establece los fundamentos teóricos y metodológicos necesarios para comprender las herramientas empleadas en el desarrollo de un modelo predictivo basado en inteligencia artificial, orientado a estimar la asistencia de público en los partidos de la Liga Profesional de Fútbol Argentino (LPF) durante la temporada 2025. Asimismo, se abordan conceptos vinculados a la asistencia y a los factores que tienen incidencia en la misma.

2.1.1. Inteligencia Artificial

La Inteligencia Artificial (IA) es una rama de la informática que se ocupa del diseño de sistemas capaces de realizar tareas que normalmente requieren inteligencia humana,

como la toma de decisiones, el aprendizaje o el reconocimiento de patrones. A su vez, la IA puede definirse como "*el estudio de agentes que perciben su entorno y realizan acciones que maximizan sus posibilidades de éxito*" (RUSSELL y NORVIG, 2020). Esta definición pone el foco en el comportamiento racional de los sistemas, independientemente de si imitan o no el pensamiento humano.

Dentro de la IA, una de las áreas de mayor aplicación actual es el aprendizaje automático o Machine Learning (ML), que permite a los sistemas aprender a partir de datos sin ser programados de forma explícita. Este enfoque resulta particularmente útil en contextos donde el comportamiento humano es complejo y difícil de modelar con reglas fijas, como ocurre con la asistencia a encuentros futbolísticos. En estos escenarios, la IA posibilita el desarrollo de modelos predictivos que analizan múltiples variables: deportivas, temporales, sociales y logísticas. El análisis se realiza para estimar con mayor precisión la cantidad de público que asistirá a un encuentro determinado.

2.1.2. Machine Learning

El aprendizaje automático o Machine Learning (ML) es una subárea de la inteligencia artificial que se enfoca en desarrollar algoritmos capaces de aprender patrones de forma automática a partir de datos, con el objetivo de realizar predicciones o tomar decisiones sin necesidad de ser programados explícitamente. En este sentido, se puede definir como "*el campo de estudio que otorga a las computadoras la capacidad de aprender sin ser explícitamente programadas*" (SAMUEL, 1959). Esta definición ayuda a comprender de manera simple el propósito del ML: permitir que un sistema mejore su rendimiento a partir de la experiencia.

El proceso de aprendizaje consiste en utilizar un conjunto de datos de entrenamiento, es decir, una colección de ejemplos con los que el algoritmo ajusta sus parámetros para minimizar el error entre sus predicciones y los valores reales. Una vez entrenado, el modelo debe ser capaz de generalizar, es decir, ofrecer buenos resultados frente a un conjunto de datos de prueba que no ha visto previamente. En esta instancia, se evalúa la precisión de las predicciones. Si estas son erróneas, pueden darse dos escenarios: *overfitting*, que ocurre cuando el modelo aprende en exceso los datos de entrenamiento y pierde capacidad

predictiva. En caso contrario se produce el *underfitting*, que se da cuando el modelo no logra captar los patrones relevantes y realiza predicciones incorrectas tanto en los datos de entrenamiento como en los de prueba (GÉRON, 2019).

Dentro del proceso de aprendizaje, es posible distinguir tres tipos clasificándolos por la disponibilidad de datos etiquetados:

- Aprendizaje supervisado: el modelo aprende una función que mapea el conjunto de datos que recibe como entrada, con la única salida que espera.
- Aprendizaje no supervisado: el modelo busca identificar patrones en datos no etiquetados, sin tener las salidas esperadas.
- Aprendizaje por refuerzo: el agente aprende mediante la interacción con un entorno, recibiendo recompensas o penalizaciones en función de sus acciones.

2.1.3. Aprendizaje Supervisado

El aprendizaje supervisado es un enfoque dentro del aprendizaje automático en el cual un modelo es entrenado a partir de un conjunto de datos compuestos por pares de entrada y salida. Cada entrada está asociada a una salida esperada, lo que permite que el algoritmo desarrolle una función de mapeo desde el espacio de entrada hacia el espacio de salida (IBM, 2024). El objetivo del entrenamiento es ajustar los parámetros del modelo para que las predicciones realizadas ante nuevas entradas se aproximen, en la medida de lo posible, a las salidas esperadas.

Durante el entrenamiento, el modelo ajusta sus parámetros internos en función del desempeño que obtiene sobre un conjunto de datos etiquetados. Este proceso busca establecer una correspondencia entre entradas y salidas que permita realizar predicciones válidas sobre datos nuevos (MITCHELL, 1997).

A nivel conceptual, los problemas de aprendizaje supervisado pueden clasificarse según la salida obtenida en dos tipos:

- Regresión: en la cual la variable de salida es continua. Un ejemplo es la predicción de un valor numérico como temperatura, precio o cantidad.
- Clasificación: donde la salida pertenece a un conjunto finito de categorías. Por ejemplo, asignar una etiqueta a una observación dentro de un conjunto discreto de clases.

A continuación, se describen algunos modelos basados en aprendizaje supervisado comúnmente aplicados a la predicción de asistencia, utilizando técnicas de regresión.

Regresión Lineal

La regresión lineal es un modelo estadístico utilizado para describir la relación existente entre una variable dependiente continua y un conjunto de variables independientes. Este modelo asume que dicha relación puede aproximarse mediante una combinación lineal de los predictores. Formalmente, un modelo de regresión lineal (1) estima una función f(x) tal que:

$$y = \beta_0 + \beta_1\chi_1 + \beta_2\chi_2 + \cdots + \beta_n\chi_n + \varepsilon \quad (1)$$

Donde y representa la variable objetivo, χ_i son los predictores, β_i los coeficientes del modelo, y ε el término de error aleatorio. Este último se asume con media cero, independencia respecto a las variables explicativas, y varianza constante (supuesto de homocedasticidad), lo que implica que la dispersión de los errores es la misma para todos los valores de las variables independientes.

Desde la perspectiva del aprendizaje automático, la regresión lineal se considera un modelo paramétrico, ya que su forma funcional está definida a priori y su complejidad depende del número de parámetros que deben ajustarse. A pesar de su simplicidad, es frecuentemente utilizado como modelo base por su facilidad de interpretación, eficiencia computacional y utilidad como punto de comparación frente a modelos más complejos (BISHOP, 2006).

En escenarios con datos etiquetados, la regresión lineal permite construir una función de mapeo entre variables de entrada y salida, lo que la hace aplicable a problemas de estimación como la predicción de asistencia a eventos deportivos. En este contexto, se pueden utilizar variables como el equipo local, la instancia del torneo, el horario del partido o las condiciones climáticas como predictores del número de asistentes. No obstante, su capacidad predictiva puede verse limitada cuando la relación entre estas variables y la asistencia no es lineal, existe multicolinealidad entre los predictores (es decir, correlaciones elevadas entre las variables independientes), o el conjunto de datos presenta un alto nivel de ruido, lo cual puede afectar la precisión del modelo y su capacidad de generalización.

Árboles de Decisión

Los árboles de decisión son modelos de aprendizaje supervisado que se utilizan para resolver tareas tanto de clasificación como de regresión. Su principal característica es su estructura jerárquica en forma de árbol binario. En esta estructura, cada nodo interno representa una condición sobre una variable de entrada, y cada rama corresponde a una de las posibles respuestas. Los nodos hoja, por su parte, contienen el resultado de la predicción: una clase en problemas de clasificación o un valor numérico en problemas de regresión.

El proceso de construcción del árbol comienza con el conjunto completo de datos y consiste en seleccionar, en cada paso, la variable y el punto de corte que mejor separen los datos según un criterio de impureza o error. Esta división se realiza de forma recursiva, generando nuevas ramas hasta que se alcanza una condición de parada, como una profundidad máxima del árbol o una cantidad mínima de datos en una hoja. El árbol resultante puede ser posteriormente podado para reducir el sobreajuste, eliminando ramas que aportan poco a la precisión del modelo general.

Entre las ventajas de los árboles de decisión se destaca que no requieren suposiciones sobre la distribución de las variables, lo cual los hace adecuados para datos reales con estructuras complejas o desconocidas. También pueden capturar relaciones no lineales e identificar automáticamente interacciones entre variables. Otra ventaja es su alta interpretabilidad, ya que permiten seguir el camino de decisiones que llevó a una predicción determinada (BREIMAN et al., 1984).

Sin embargo, los árboles de decisión también presentan algunas desventajas. Son susceptibles al sobreajuste, especialmente si se permite que crezcan sin restricciones. Además, pequeñas variaciones en los datos pueden generar árboles muy diferentes, lo cual afecta su estabilidad. Para mitigar estas limitaciones, suelen emplearse métodos de ensamblado como Random Forest o Gradient Boosting, que combinan múltiples árboles para mejorar la robustez y la precisión del modelo.

Métodos de Ensamble

Los métodos de ensamble son una estrategia dentro del aprendizaje automático que busca mejorar el rendimiento predictivo mediante la combinación de múltiples modelos. En lugar de depender de un único modelo, el enfoque de ensamble construye un conjunto de modelos, denominados *weak learners* y luego los combina para obtener una predicción conjunta más robusta y precisa. Esta técnica se fundamenta en la premisa de que diferentes modelos pueden capturar distintos aspectos del conjunto de datos, y que su integración permite reducir errores individuales y mejorar la capacidad de generalización del sistema (ZHOU, 2012). Existen tres enfoques principales para construir modelos de ensamble:

- Bagging: consiste en generar múltiples subconjuntos del conjunto de datos original mediante muestreo aleatorio con reemplazo. Sobre cada subconjunto se entrena un modelo independiente, y las predicciones finales se obtienen a través del promedio (para regresión) o del voto mayoritario (para clasificación).
- Boosting: este enfoque construye modelos de forma secuencial, donde cada nuevo modelo intenta corregir los errores cometidos por los anteriores. Los modelos se combinan asignándoles pesos según su desempeño, lo que permite mejorar la precisión global.
- Stacking: en este caso, se entrenan varios modelos diferentes (de distintas clases o configuraciones), y luego se utiliza un modelo adicional, denominado meta-modelo para aprender cómo combinar sus salidas de manera óptima.

2.1.4. Evaluación de modelos predictivos

La evaluación de modelos predictivos consiste en aplicar métricas cuantitativas que permiten medir la diferencia entre los valores predichos por el modelo y los valores reales observados. En el caso de modelos de regresión, la variable que se busca predecir es continua, por ende, las métricas de error adoptadas deben reflejar con precisión la magnitud de las desviaciones en dicho contexto (SCHLOSSER et al., 2024). Las métricas utilizadas con mayor frecuencia son:

- Error Absoluto Medio (Mean Absolute Error o MAE) (2): el promedio de las diferencias absolutas entre las predicciones y los valores reales. Esta medida indica el tamaño medio de los errores, sin considerar su dirección.

$$MAE = \frac{AE}{n} = \frac{1}{n} \cdot \sum_{i=1}^{n} |GT_i - P_i| \quad (2)$$

- Error Cuadrático Medio (*Mean Squared Error* o MSE) (3): el promedio de los errores elevados al cuadrado. En este evaluador, los errores más grandes tienen un mayor impacto sobre el valor final, lo que hace al MSE más sensible a valores atípicos.

$$MSE = \frac{SE}{n} = \frac{1}{n} \cdot \sum_{i=1}^{n} (GT_i - P_i)^2 \quad (3)$$

- Raíz del Error Cuadrático Medio (*Root Mean Squared Error o* RMSE) (4): es la raíz cuadrada del MSE y se interpreta en las mismas unidades que la variable dependiente.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (GT_i - P_i)^2} \quad (4)$$

- Error Porcentual Absoluto Medio (*Mean Absolute Percentage Error* o MAPE) (5): Es el promedio de todos los errores absolutos expresados en porcentaje.

$$MAPE = \frac{100}{n} \cdot \sum_{i=1}^{n} \frac{|Ei|}{|GTi|} = \frac{100}{n} \cdot \sum_{i=1}^{n} \frac{|GTi - Pi|}{|GTi|} \quad (5)$$

- Coeficiente de determinación (*coefficient of determination* o R^2) (6): mide qué tan bien un modelo estadístico logra predecir un resultado. Dicho resultado está representado por la variable dependiente del modelo.

$$R^2 = 1 - \frac{RSS}{TSS} \quad (6)$$

La selección de una métrica específica depende de las características del problema, del comportamiento esperado del modelo frente a errores extremos, y del equilibrio entre precisión y estabilidad.

2.1.5. Factores que influyen en la asistencia a eventos deportivos

La asistencia a partidos de fútbol en Argentina puede verse condicionada por diversos factores contextuales, logísticos y deportivos. Estos mismos influyen, en mayor o menor medida, en la decisión del público de asistir a los encuentros futbolísticos y se pueden considerar como variables explicativas en los modelos de predicción propuestos en este trabajo. Entre los principales factores identificados que se abordan, podemos distinguir:

- Condiciones climáticas: Las condiciones meteorológicas que se presentan durante el encuentro futbolístico pueden incidir negativamente en la decisión de asistir, tales como lluvias, temperaturas extremas o alertas climáticas pueden representar un deterioro en la experiencia del espectador.
- Día y horario: La programación de los encuentros en determinados días y horarios puede imposibilitar la concurrencia de ciertos grupos de público. El presente estudio busca establecer si determinadas franjas horarias están asociadas a una menor concurrencia, y en qué medida.
- Importancia del partido y rivalidad: Los encuentros disputados ante “clásicos”, los partidos definitorios o los enfrentamientos ante rivales tradicionales suelen generar mayor atracción y entusiasmo en los espectadores. Esta condición es explorada a través del análisis de datos históricos. Se evalúa si este tipo de encuentros tiene incidencia alguna en los niveles de asistencia, y en qué medida.

Estos factores se incorporan en el modelo predictivo desarrollado en el presente trabajo, ya sea mediante variables cuantitativas (temperatura, resultado anterior) o categóricas (rivalidad). Su inclusión permite construir estimaciones más ajustadas al contexto argentino y mejorar la utilidad operativa del sistema propuesto.

2.2. Estado del Arte

En los últimos años, el avance de la inteligencia artificial y el aprendizaje automático ha abierto nuevas posibilidades en distintos ámbitos, incluyendo el deporte. La predicción de la asistencia a eventos deportivos es una de ellas, ya que permite anticiparse a la demanda de público y planificar de forma más eficiente los recursos necesarios. En este sentido, distintos proyectos y estudios han buscado aprovechar datos históricos y variables contextuales para estimar la cantidad de asistentes a partidos, mejorando la toma de decisiones tanto en el ámbito organizativo como en el logístico.

2.2.1. Nivel Internacional

En el ámbito internacional, existen diversos antecedentes que muestran cómo se han utilizado técnicas de machine learning para anticipar la cantidad de público en eventos deportivos. Por ejemplo, en Estados Unidos, se han desarrollado modelos para la NFL (National Football League), utilizando datos históricos de asistencia, popularidad de los equipos, rendimiento, día del evento y presencia en redes sociales. Estos modelos han permitido obtener predicciones bastante cercanas a la realidad y fueron utilizados para fines de gestión y marketing (PANG y WANG, 2024).

Otro caso existente es el de la Copa Mundial de la FIFA de Qatar, en el año 2022, donde se desarrollaron herramientas para estimar la cantidad de público en base a la ubicación del evento, la nacionalidad de los equipos participantes, la distancia geográfica entre estadios y otros factores socioculturales (AL-BUENAIN et al., 2024). Estos estudios muestran cómo, con un enfoque adecuado, es posible anticiparse al comportamiento del público y adaptar la logística de forma eficiente.

También en Europa se ha trabajado en este tipo de predicciones. En Italia, particularmente en la Serie A, se han combinado modelos de análisis de preferencias con técnicas automáticas para anticipar el comportamiento de los hinchas, ayudando a los clubes a mejorar la planificación y la experiencia del espectador (PERUZZO y ASANI, 2024).

2.2.2. Nivel Regional

En Sudamérica, si bien hay menos desarrollos en comparación con otros continentes, se destaca un trabajo en particular. En Brasil, se han utilizado modelos de predicción para estimar la asistencia a partidos de fútbol en ciertas regiones, combinando variables como el horario, el día de la semana, el clima y el desempeño reciente del equipo. Estos enfoques han demostrado ser útiles especialmente en ligas con gran cantidad de partidos y poco margen para organizar operativos personalizados (YAMASHITA et al., 2021).

Estos ejemplos marcan un camino que puede ser seguido en otros países de la región, considerando las similitudes culturales y deportivas que se comparten entre sí.

2.2.3. Nivel Nacional

En Argentina, actualmente no se han identificado proyectos que apliquen modelos de machine learning específicamente para predecir la asistencia a partidos de fútbol. Sin embargo, en los últimos años ha crecido el interés por aplicar tecnologías de datos en el ámbito futbolístico, con avances en el análisis del rendimiento de jugadores, estrategias de juego y gestión de entradas.

La disponibilidad existente, y creciente, de datos estadísticos y plataformas open-source ofrecen oportunidades para avanzar con este tipo de estudios. Surge la necesidad de proyectos que exploren nuevas formas de análisis, orientadas a resolver problemas operativos como la planificación de recursos y de seguridad en partidos de alta demanda.

Podemos mencionar a la plataforma *Transfermarkt*, que recopila y publica la asistencia promedio anual a partidos de distintas ligas, incluyendo la argentina. Si bien esta información resulta útil como referencia estadística general, no contempla la variabilidad de cada encuentro en particular. En cambio, una solución de predicción partido a partido permitiría anticipar con mayor precisión la demanda específica de cada evento, ofreciendo datos de mayor valor para la planificación logística y operativa.

2.2.4. Nivel Local

En el ámbito de la Ciudad de Buenos Aires y alrededores, no existen investigaciones aplicadas al uso de Inteligencia Artificial para predecir asistencia. Sin embargo,

la accesibilidad a datos históricos, climáticos y otros factores permite pensar en un modelo local que integre estas variables. La regularidad de los partidos, las multitudes de hinchas y los patrones repetitivos de comportamiento generan condiciones ideales para el desarrollo de modelos predictivos adaptados a este contexto.

2.2.5. Conclusión

Como se observa, hay un desarrollo importante a nivel internacional en el uso de técnicas de machine learning para predecir la asistencia a eventos deportivos, con experiencias concretas que sirven como referencia. En América del Sur y especialmente en Argentina, el campo aún está poco explorado, lo que representa tanto un desafío como una gran oportunidad.

TABLA I: comparativa de los modelos analizados

Autores (Año)	**Contexto**	**País**	**Variables Utilizadas**	**Aplicación Principal**
Al-Buenain, Haouari y Jacob (2024)	Copa Mundial de la FIFA 2022	Qatar	Nacionalidad, distancia, contexto sociocultural	Organización y planificación de los encuentros
Pang y Wang (2024)	NFL (National Football League)	EE. UU.	Popularidad del equipo, redes sociales, día, historial	Optimización de entradas y recursos
Peruzzo y Asani (2024)	Serie A	Italia	Preferencias del público, atributos del partido	Mejora de experiencia e ingresos
Yamashita et al. (2021)	Distintas ligas regionales	Brasil	Clima, día, rendimiento reciente, tipo de rival	Planeamiento de seguridad y logística

Fuente: (PANG y WANG, 2024); (AL-BUENAIN et al., 2024); (PERUZZO y ASANI, 2024); (YAMASHITA et al., 2021)

En base a lo ya existente sobre estas experiencias en ligas de otros países, proponemos adaptarlo al contexto del fútbol argentino y construir un enfoque de predicción de asistencia específico para la LPF, planteándose de una manera en la que, en una versión futura, pueda adaptarse sin cambios sustanciales a cualquier otra liga del país o de la región. Además, si bien existen plataformas como *Transfermarkt*, que publica estadísticas históricas de asistencia en forma anual, la solución incorpora la capacidad de realizar predicciones y poner a disposición la información partido a partido, lo que permite anticiparse en la planificación de recursos y de seguridad en cada encuentro.

2.3. User Research

Con el objetivo de investigar si la solución propuesta tiene sentido en la práctica, realizamos entrevistas con dos personas que trabajan directamente en la planificación de este tipo de eventos, una que forma parte de la organización y otra que pertenece a un club en particular. Buscamos entender cómo manejan hoy estas situaciones, qué herramientas usan, qué dificultades enfrentan y si vieran útil contar con una solución de este estilo.

Por otra parte, también realizamos una encuesta dirigida a personas que asisten habitualmente a estos encuentros futbolísticos, con el objetivo de relevar su experiencia respecto a la organización de estos eventos. Buscamos validar desde la perspectiva del público algunas de las problemáticas asociadas a la planificación, gestión y asignación de recursos.

2.3.1. Entrevista a Tomás Salomone (GCBA)

Tuvimos la oportunidad de entrevistar a Tomás Salomone, quien trabaja en el Gobierno de la Ciudad de Buenos Aires (GCBA) y está directamente involucrado en la organización de operativos para eventos futbolísticos. Poder conversar con alguien comprometido con el tema es clave para entender los procedimientos actuales, analizarlos y poder llegar a implementar soluciones con posibilidades de mejora.

Consultamos sobre qué tipo de información consideran más útil a la hora de planificar y realizar las estimaciones. Lo primero que mencionó fue el historial de partidos similares. Destacó el contexto del encuentro: si es un clásico, si hay una rivalidad fuerte, si el resultado define algo importante. A esto se le suma la comunicación directa y activa que tienen

con los clubes, que suelen tener un indicio del tipo de público que va a asistir. También cuentan con recursos como los anillos digitales y medidores de tránsito, que les permiten monitorear el movimiento en tiempo real. Con todo eso, planifican los operativos de seguridad.

La debilidad actual se ve reflejada cuando no tienen tantos datos concretos: lo que hacen es sobreestimar la cantidad de gente que puede llegar a ir. Prefieren exagerar en la estimación y estar preparados para un escenario más exigente, teniendo en cuenta la importancia en términos de seguridad, tránsito, salud, entre otros. Esto da como resultado una gran cantidad de recursos que se asignan sin ser utilizados y, como consecuencia, un mayor costo de los operativos.

Luego se consultó sobre la utilidad de una herramienta que, usando inteligencia artificial y distintas variables, pueda anticipar con mayor precisión la cantidad de público esperada. Expresó que las herramientas que se usan actualmente para prever la concurrencia son una combinación de datos obtenidos de diversas fuentes y que las mismas combinadas funcionan con éxito. Sin embargo, reconoció que, si se logra unificar dichas fuentes e integrar correctamente esa información en una única herramienta, se puede mejorar ampliamente lo que ya existe, generando un impacto positivo en la planificación.

Después conversamos sobre las áreas del gobierno donde este tipo de predicciones resultaría útil. Comentó que las áreas de transporte, seguridad, salud, limpieza ya cuentan con sus propios métodos para monitorear distintos factores. Pero que, si se pudieran combinar y usar en conjunto, podrían tomar mejores decisiones y anticiparse aún más.

Por último, preguntamos qué debería tener una herramienta de estas características para poder ser usada en el ámbito público. Respondió que una herramienta que sirva para medir el comportamiento de las multitudes puede ser útil, siempre y cuando cumpla con los estándares del gobierno. Además, expresó que estaría dispuesto a participar de una prueba piloto si el sistema está alineado con los requisitos técnicos y legales que manejan.

Esta entrevista sirve para entender cómo se toman decisiones hoy en día y qué lugar podría ocupar una herramienta como la propuesta. Aunque ya hay prácticas instaladas que funcionan, también hay un interés en seguir mejorando.

Se adjunta la transcripción completa de la entrevista en el Anexo A.

2.3.2. Entrevista a Diego Filippi (AAAJ)

En el marco de la investigación, también realizamos una entrevista con Diego Filippi, quién es uno de los responsables del ingreso y posterior registro del público que asiste a los partidos en la Asociación Atlética Argentinos Juniors (AAAJ), con el fin de comprender en detalle los procesos actuales de organización y gestión de recursos en dicho club. A continuación, se presenta el desarrollo de la conversación en formato cómo fue tomando lugar:

Inicialmente, consultamos por cómo se toman hoy las decisiones sobre organización logística antes de un partido y qué información se usa. Respondió que la organización de cada evento deportivo se rige, en líneas generales, por un protocolo preestablecido, aunque este puede adaptarse en función del equipo visitante o de definiciones del comité de seguridad de la ciudad de Buenos Aires. Explicó que, previo al encuentro, el Comité de Seguridad de AAAJ, en conjunto con UTEDYC (responsables del control en los accesos), revisa el protocolo específico a seguir para el evento, asegurando que todas las disposiciones estén alineadas con el contexto particular del partido. En esta instancia se planifican las necesidades puntuales en materia de accesos, considerando variables como el tipo de rival y la etapa del torneo en la que se encuentra el club local. Estas condiciones pueden derivar en la necesidad de reforzar los accesos mediante la incorporación de tecnología, con el objetivo de mejorar la fluidez del ingreso del público. Además, indicó que el líder de seguridad de AAAJ es quien se encarga de coordinar estas acciones con los responsables del operativo policial y con el Comité de Seguridad de la Liga, garantizando una planificación y ejecución integral del dispositivo de seguridad.

En relación con las dificultades que aparecen cuando se organizan partidos con mucha asistencia de público, comentó que una de las principales se presenta en el control de accesos, específicamente en los molinetes. Ante el más mínimo inconveniente técnico en la lectura de carnets o códigos QR, el sistema se ve afectado, generando demoras que, en cuestión de minutos, pueden derivar en la congestión o incluso en el colapso de los accesos. Agregó que otra dificultad está vinculada al control de aforo en las tribunas, ya que en eventos de gran convocatoria o alta relevancia alcanzar el aforo máximo en determinados sectores puede generar complicaciones operativas, lo que en algunos casos obliga al cierre temporal de accesos para garantizar la seguridad.

Cuando consultamos qué suele suceder internamente si hay una diferencia significativa entre la cantidad de gente esperada y la que finalmente asiste, respondió que no ocurre nada en particular, dado que los protocolos establecidos y la cantidad de personal de seguridad privada y policial están planificados para soportar el máximo aforo que permite el estadio, aunque eso signifique una sobreestimación.

Respecto a situaciones recientes en las que la concurrencia no fue la esperada, comentó que por lo general la baja asistencia está estrictamente relacionada con el rendimiento del equipo en el torneo, junto con el rival del encuentro.

En cuanto a los aspectos del proceso organizativo que generan más incertidumbre o preocupación antes de un partido, explicó que organizar un evento de estas características siempre implica un cierto grado de incertidumbre, especialmente en lo relacionado con el factor humano. No obstante, aclaró que en AAAJ ese riesgo se minimiza considerablemente: se trata de un club con un fuerte espíritu familiar, donde todos los actores involucrados conocen su rol y colaboran con compromiso, lo que permite que cada operativo se desarrolle con orden y previsibilidad.

Sobre partidos en los que la organización funcionó especialmente bien o mal, recordó con claridad varios al inicio de las temporadas en los que se presentaron inconvenientes en la lectura de los molinetes, lo que generó dificultades en algunos accesos para el ingreso del público. Explicó que, a raíz de estos episodios, se realizó un análisis técnico del evento que permitió identificar y corregir la causa raíz del problema. Como resultado, se modificó el proceso de sincronización entre el sistema de socios y el sistema de control de accesos del estadio, lo que mejoró notablemente la estabilidad y confiabilidad del ingreso.

Consultamos sobre cómo afecta la asistencia a las decisiones en otras áreas, como seguridad, personal o entradas. Señaló que este aspecto es clave, y por ello el estadio cuenta con un sistema de cámaras estratégicamente ubicadas. Desde el centro de monitoreo, donde trabajan en conjunto la Policía Federal, el equipo de seguridad de AAAJ y personal de UTEDYC, se supervisa en tiempo real cada uno de los accesos, así como el aforo de las tribunas. Este monitoreo online permite una rápida detección de incidentes. En caso de producirse algún inconveniente, tanto la Policía Federal como el equipo de seguridad del club y el personal de

UTEDYC actúan de inmediato, coordinando acciones para restablecer el orden o mitigar problemas tecnológicos que impidan el acceso.

Al indagar dónde se nota primero cuando la planificación no alcanza, respondió que, si llegara a suceder, la sobreocupación se hace evidente a simple vista en el aforo de las tribunas.

También consultamos si consideraba útil contar con una solución tecnológica que realice predicciones confiables de asistencia a los partidos. Respondió que sí, sería de gran utilidad, ya que una herramienta de este tipo permitiría anticipar escenarios con mayor precisión y planificar en consecuencia aspectos clave como accesos, seguridad, disposición del personal y servicios. Recalcó que contar con predicciones confiables ayudaría a optimizar recursos y reducir imprevistos durante el evento.

Finalmente, al preguntarle qué es lo más difícil de organizar un partido, afirmó que, sin dudas, lo más complejo es coordinar las necesidades de último minuto del comité de seguridad y de la Policía Federal.

En conclusión, la entrevista permite comprender que la organización logística de un partido de fútbol combina protocolos preestablecidos con cierta flexibilidad necesaria para adaptarse a factores como el rival, la instancia del torneo y las disposiciones de los organismos de seguridad. Si bien existen desafíos vinculados principalmente al control de accesos y al manejo del aforo, en un club como Argentinos Juniors, los operativos se diseñan siempre para soportar la máxima capacidad del estadio, lo que otorga un marco de previsibilidad. Además, se destaca que la asistencia del público depende en gran medida del rendimiento deportivo, y que la incorporación de herramientas tecnológicas, como un software que permita realizar predicciones confiables de la posible asistencia a un encuentro, sería clave para optimizar la gestión y reducir la incertidumbre en la planificación.

Se adjunta la transcripción completa de la entrevista en el Anexo B.

2.3.3. Conclusión de las entrevistas realizadas

Como conclusión de las entrevistas realizadas, detectamos que la organización de un partido se basa en protocolos y experiencia acumulada (como el historial de encuentros o el contexto del rival), con comunicación activa con los clubes y ayuda de sensores en la ciudad

(como anillos digitales o el mismo tránsito) para planificar. Existe una relación entre la demanda y el rendimiento deportivo, el rival y contexto del encuentro. En los casos donde no existen datos suficientes, se tiende a sobreestimar la asistencia para garantizar seguridad, lo que eleva los costos por el aumento de recursos. Tanto desde el GCBA como desde los clubes surge la necesidad de contar con una herramienta de predicción para anticipar los recursos a desplegar en el día de un partido.

En base a estas conclusiones, construimos un modelo de *User Persona* que representa al perfil típico de quienes se encargan de las tareas de planificación y gestión de la asistencia a los partidos:

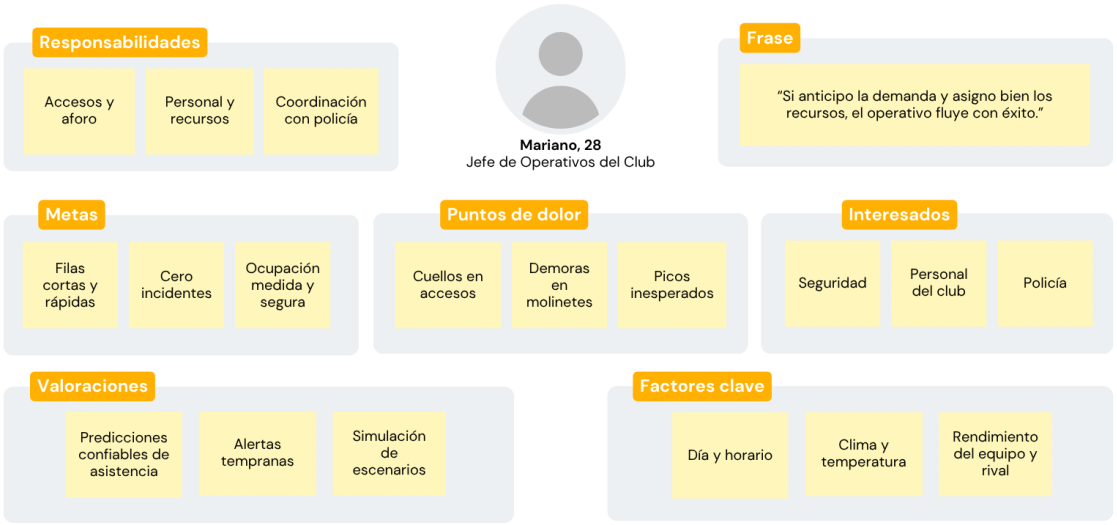


Figura 1. User Persona. Fuente: Elaboración propia, 2025.

Como complemento, diseñamos un mapa de empatía que resume las percepciones, necesidades y dificultades que tienen los responsables de la organización de los partidos. Esto nos permite visualizar no solo qué información manejan hoy y cómo toman decisiones, sino también cuáles son los puntos de dolor que más les afectan. De esta forma, podemos detectar oportunidades de mejora y validar la necesidad de la solución propuesta.

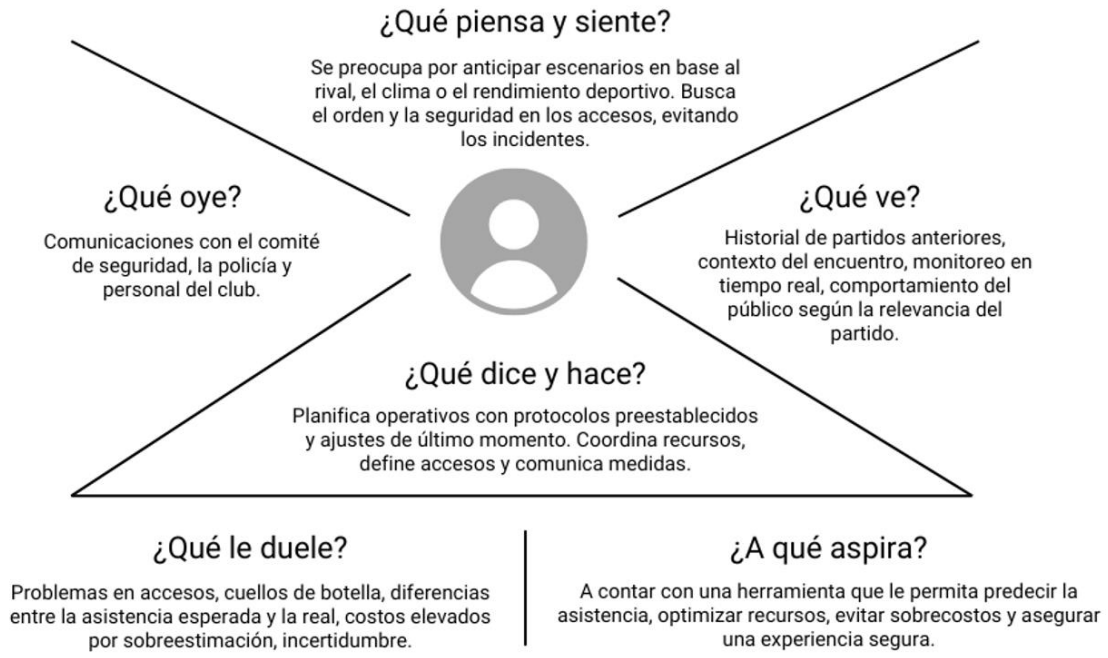


Figura 2. Mapa de empatía. Fuente: Elaboración propia, 2025.

2.3.4. Encuesta a asistentes habituales

La encuesta estuvo orientada a personas que asisten habitualmente a partidos de fútbol de la Liga Profesional de Fútbol (LPF), con el objetivo de conocer con qué frecuencia concurren, qué factores consideran determinantes al momento de decidir su asistencia, si han enfrentado demoras en los accesos, y cómo perciben la disponibilidad de recursos y personal operativo. Esto nos permitió validar y aportar evidencia empírica sobre los problemas que se presentan en el desarrollo de estos eventos.

Con una muestra de 298 personas encuestadas, cuyos perfiles corresponden mayoritariamente a asistentes habituales (según la frecuencia con la que asisten a los encuentros, cómo se puede ver en la Figura 3), los resultados reflejan una visión crítica sobre la organización de los partidos de la Liga Profesional de Fútbol, que no se limita a espectadores ocasionales, sino que impacta en quienes asisten semana tras semana.

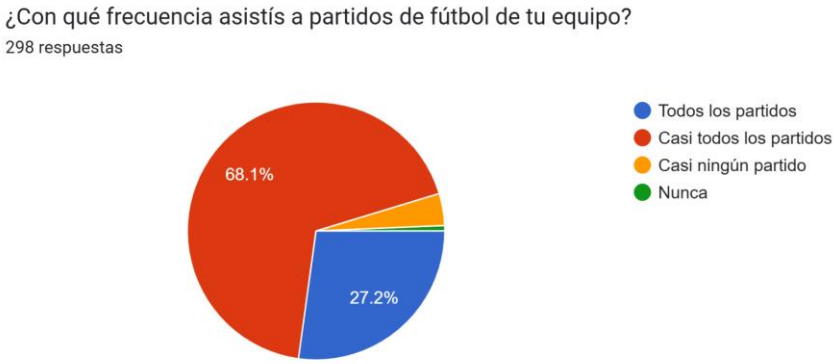


Figura 3. Frecuencia de asistencia a partidos de los encuestados. Fuente: Elaboración propia, 2025.

Ante la consulta sobre el nivel de satisfacción general (Figura 4), el 65% manifestó estar insatisfecho con la forma en que se gestionan estos eventos. Este dato revela un alto grado de disconformidad en la experiencia que tiene el espectador respecto de la organización (previa, durante y posterior) de los partidos, lo que refuerza la existencia de una problemática.

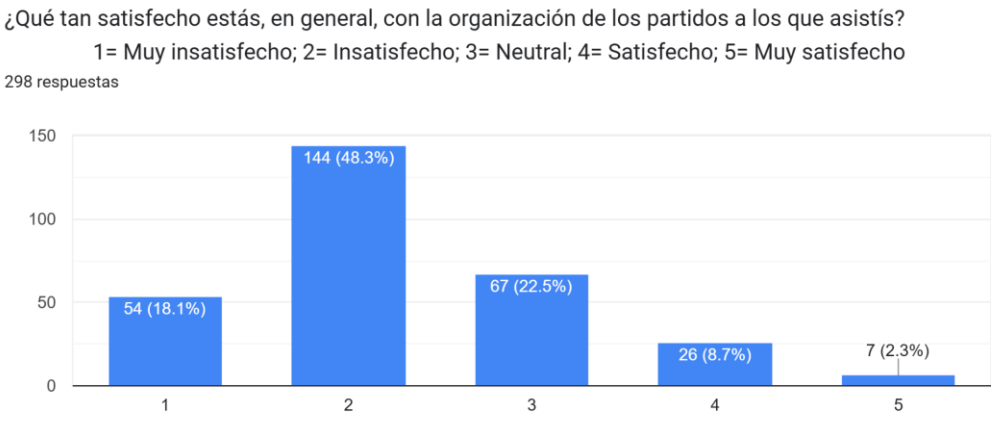


Figura 4. Satisfacción de los encuestados sobre la organización de los partidos. Fuente: Elaboración propia, 2025.

Tal como se observa en la Figura 5 en relación con los accesos al estadio, el 75% de los encuestados indicó haber experimentado demoras o inconvenientes al momento de

ingresar. Sin diferenciar características como el estadio o equipos del partido, la cuestión de los accesos se muestra como una de las principales fuentes de conflicto, afectando a la experiencia de una gran parte de los hinchas.

Figura 5. Encuestados afectados por demoras en los accesos a estadios. Fuente: Elaboración propia, 2025.

En cuanto a la disponibilidad de personal, el 57% considera que la cantidad asignada no es suficiente para cubrir la demanda en ciertos encuentros. Esto se observa en la Figura 6 y en la Figura 7, donde se refleja la necesidad de ajustar la asignación de recursos y de planificación, y donde se evidencia que la falta de personal humano impacta directamente en el desarrollo de los operativos.

Figura 6. Encuestados que creen que la cantidad de personal es suficiente. Fuente: Elaboración propia, 2025.

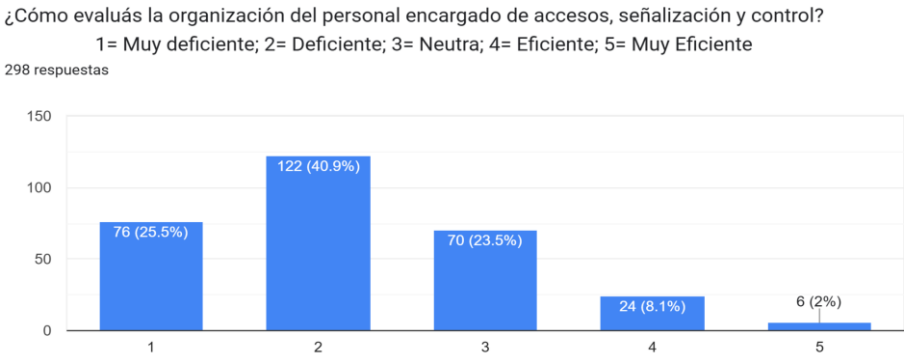


Figura 7. Evaluación de los encuestados sobre la organización del personal. Fuente: Elaboración propia, 2025.

Por otro lado, los factores que influyen en la decisión de asistir confirman que el comportamiento de los hinchas tiene en cuenta variables externas e internas (Figura 8). El día y horario del partido (69%) y las condiciones climáticas (55%) aparecen como determinantes, lo cual refleja que la asistencia está condicionada por cuestiones que exceden al partido en sí. En cambio, el tipo de rival (41%) y el rendimiento actual del equipo (38%) muestran el lado de la decisión que se relaciona con la pasión y la motivación deportiva. Este equilibrio entre factores contextuales y deportivos ayuda a comprender la complejidad que conlleva anticipar la demanda.

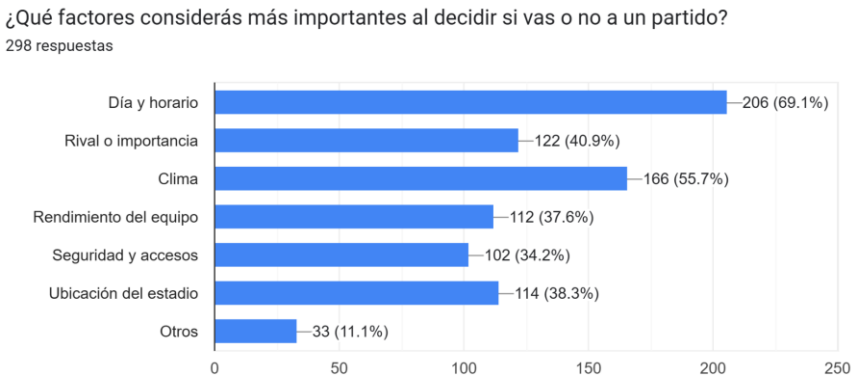


Figura 8. Factores influyentes en la asistencia a partidos según los encuestados. Fuente: Elaboración propia, 2025.

En resumen, los resultados de la encuesta dejan en claro que hay aspectos de la organización de los partidos que no funcionan correctamente. La cantidad de respuestas obtenidas (cerca de 300 encuestados) permite sentar una base de opiniones, asegurando que no se limiten a casos aislados, sino que representen al público que asiste con frecuencia.

Los problemas con los accesos, la falta de personal y las complicaciones para planificar la asistencia aparecen de forma constante en las respuestas. Esto demuestra que la experiencia del público se ve afectada por temas relacionados con la cantidad de recursos que se asignan, la cual resulta ineficiente cuando se compara con la demanda real en los partidos.

Además, deja en evidencia que la asistencia depende de múltiples factores que se relacionan tanto con el aspecto deportivo, como con el propio contexto. Estas variables resultan importantes a la hora de realizar una estimación y cada una tiene influencia directa, en distinto grado, en la concurrencia del público.

Frente a esto, contar con una herramienta que ayude a anticipar cuánta gente va a asistir resultaría útil y además necesaria para mejorar la organización y evitar los problemas que hoy se repiten partido tras partido.

A partir de las conclusiones surgidas a partir de la encuesta y las entrevistas, elaboramos un análisis FODA que muestra las fortalezas y debilidades de nuestra propuesta, así como también las oportunidades y amenazas que pueden surgir en su implementación en el contexto real.

Fortalezas:

- Respuesta frente a un problema existente, validado a través de entrevistas y una encuesta a asistentes habituales.
- Innovación en el contexto del fútbol argentino, donde no existen antecedentes.
- Optimización de recursos y reducción de costos operativos a través de estimaciones precisas de asistencia.

Oportunidades:

- Escalabilidad futura hacia otras ligas del país o de la región, pensada con el diseño de una arquitectura flexible y adaptable.

- Interés existente en los clubes y organizadores sobre la herramienta de predicción y en la utilización de la misma.
- Colaboración con los actores involucrados para facilitar la implementación y la adopción del uso de la plataforma.

Debilidades:

- Dependencia de la calidad y disponibilidad de los datos históricos, los cuales son fundamentales para el entrenamiento adecuado de los modelos.
- Necesidad de ajustes en los modelos para los factores contextuales, los cuales pueden variar en la importancia con la que influyen en la asistencia.

Amenazas:

- Resistencia al cambio por parte de ciertas instituciones que estén atadas a sus protocolos tradicionales.
- Incertidumbre en la disponibilidad futura de datos confiables y estandarizados, necesarios para continuar alimentando a los modelos luego de la implementación.
- Posibles cambios en las regulaciones de organismos de seguridad que afecten la implementación de la solución.

3. Descripción

El proyecto tiene como objetivo principal desarrollar una plataforma digital que permita a los tomadores de decisión de los clubes de la LPF anticipar la demanda de asistencia a los partidos para planificar con mayor eficiencia los recursos necesarios. Conocer de antemano el posible aforo habilita una mejor organización de seguridad, transporte, accesos y cantidad de personal.

Para lograr este objetivo, en la presente sección se detallan de manera estructurada las etapas desarrolladas, las cuales comprenden un análisis funcional, la recolección y tratado de datos, el entrenamiento de los distintos modelos, la arquitectura conceptual de la solución y sus atributos de calidad.

3.1. Análisis funcional

La presente sección tiene como finalidad identificar y describir las funcionalidades que debe cumplir la solución propuesta, así como las interacciones esperadas entre los actores y el sistema. En esta sección se detallan los requerimientos funcionales y no funcionales, y se presentan los principales casos de uso.

3.1.1. Requerimientos

En esta sección, se describen las necesidades y condiciones que el sistema debe satisfacer para cumplir con los objetivos propuestos. Dentro de la misma, se distinguen requerimientos funcionales y no funcionales.

Requerimientos funcionales

Los requerimientos funcionales definen las capacidades que el sistema debe garantizar para cumplir con los objetivos y alcance del proyecto. Los mismos se centran en las acciones y servicios que aseguran la utilidad de la solución en la planificación y gestión de recursos de un encuentro de fútbol, basándose en la asistencia estimada.

- RF-01: El sistema debe ser capaz de predecir con precisión la cantidad estimada de asistentes y el porcentaje de ocupación de los estadios para los encuentros a disputarse.

- RF-02: El sistema debe permitir la simulación de escenarios hipotéticos modificando las variables de entrada.
- RF-03: El sistema debe permitir la exportación de reportes, incluyendo información y estadísticas asociadas a los encuentros seleccionados.
- RF-04: El sistema debe ofrecer una interfaz web que permita a los usuarios institucionales interactuar con la solución, desde navegadores convencionales.
- RF-05: El sistema debe gestionar usuarios y permisos dentro de la aplicación, limitando el acceso exclusivamente a organizadores autorizados.
- RF-06: El sistema debe permitir la carga de asistencia real posterior al evento para mejorar la precisión de los modelos y calcular métricas de error.
- RF-07: El sistema debe mostrar un tablero interactivo con predicciones e información útil relacionada para dar soporte a la toma de decisiones.
- RF-08: El sistema debe permitir la búsqueda y el filtrado de partidos por equipo, fecha y competencia.
- RF-09: El sistema debe basar sus predicciones en modelos de Machine Learning entrenados con datos históricos y contextuales de los equipos.
- RF-10: El sistema debe permitir solicitar reportes específicos desde la interfaz de partidos, incorporando métricas descriptivas.
- RF-11: El sistema debe ofrecer vistas personalizadas según el rol del usuario (seguridad, logística o salud), mostrando únicamente la información relevante para su función.
- RF-12: El sistema debe permitir la solicitud de servicios de consultoría personalizados, facilitando la integración de datos históricos propios de las instituciones para el ajuste del modelo predictivo.

Requerimientos no funcionales

Los requerimientos no funcionales definen las propiedades que el sistema debe cumplir para operar de manera confiable, segura y eficiente, ante las necesidades del usuario final.

- RNF-01: El sistema debe ofrecer tiempos de respuesta menores o iguales a 3 segundos, asegurando un rendimiento óptimo en la consulta de predicciones y simulaciones.

- RNF-02: El sistema debe cumplir con criterios básicos de accesibilidad, permitiendo el uso por parte de distintos perfiles de usuario.
- RNF-03: El acceso a la plataforma debe realizarse bajo protocolos de seguridad que garanticen integridad y confidencialidad de los datos.
- RNF-04: El sistema debe ser escalable para adaptarse a diferentes volúmenes de consultas sin comprometer la calidad del servicio.
- RNF-05: El sistema debe ser mantenible, permitiendo actualizaciones y mejoras con mínima afectación al servicio.
- RNF-06: El sistema debe mantener una disponibilidad que garantice el acceso continuo a las funcionalidades principales, permitiendo la consulta y generación de predicciones.
- RNF-07: El sistema debe contar con mecanismos que permitan trazar las acciones de los usuarios dentro de la solución.

3.1.2. Casos de uso

En la siguiente sección se especifican los diferentes casos de uso del sistema, que definen las interacciones con los usuarios para lograr un objetivo específico.

Caso de Uso 1: Registro de asistencia real

- Actor involucrado: Analista de datos o administrador del club.
- Descripción: Una vez finalizado el partido, el usuario ingresa en el sistema la asistencia real registrada con el propósito de evaluar la precisión del modelo y contribuir a la mejora de futuras iteraciones del sistema.
- Flujo de interacción:
    - El sistema presenta los encuentros disputados pendientes de validación.
    - El usuario autorizado elige el partido correspondiente y selecciona la función “Registrar asistencia real”.
    - El usuario ingresa por teclado el número de asistencia y guarda el registro.

- La información queda almacenada para su posterior análisis.

Caso de Uso 2: Predicción de asistencia a un partido

- Actor involucrado: Planificador operativo de un club o institución gubernamental.
- Descripción: El sistema permite al usuario acceder a la aplicación web para obtener la predicción de asistencia correspondiente a un partido específico. La estimación incluye la cantidad proyectada de espectadores y el porcentaje esperado de ocupación, calculados a partir de información histórica y variables contextuales relevantes.
- Relaciones:
  - «include»: Predicción de asistencia.
- Flujo de interacción:
  - El usuario autorizado ingresa a la plataforma mediante credenciales válidas.
  - Una vez validado el acceso, desde el menú principal accede al apartado de Predicciones.
  - El usuario aplica filtros por competición.
  - El usuario selecciona un partido y procede a “simular asistencia”
  - El sistema procesa la solicitud y obtiene la predicción de asistencia correspondiente.
  - Se visualizan en pantalla los valores estimados de asistentes, el porcentaje de ocupación y sugerencias relacionadas con la planificación operativa.

Caso de Uso 3: Simulación de escenarios

- Actor involucrado: Planificador operativo de un club o institución gubernamental.
- Descripción: El usuario emplea la funcionalidad de simulación para analizar cómo posibles cambios en factores contextuales modifican la

proyección de asistencia a un encuentro deportivo. Esta función permite evaluar escenarios alternativos ingresando las variables a predecir.

- Relaciones:
  - «include»: Predicción de asistencia.
- Flujo de interacción:
  - El usuario autorizado ingresa a la plataforma mediante credenciales válidas.
  - Una vez validado el acceso, desde el menú principal accede al apartado de Escenarios.
  - El usuario ingresa los parámetros a incluir en la predicción y procede a “Simular escenario”.
  - El sistema calcula la predicción en función de los parámetros ingresados.
  - Se muestran por pantalla los resultados de la simulación.

Caso de Uso 4: Consulta de vistas personalizadas por rol

- Actor involucrado: Personal de seguridad, logística o salud pública.
- Descripción: El sistema muestra vistas personalizadas según el rol, con métricas y simulaciones específicas para optimizar la planificación de recursos, a partir de la autentificación del usuario.
- Flujo de interacción:
  - El usuario autorizado accede a la plataforma.
  - La plataforma detecta el rol del usuario y muestra las vistas personalizadas del mismo.
  - El usuario puede simular escenarios alternativos.
  - Las simulaciones muestran métricas específicas según el rol del usuario.
  - El sistema actualiza las recomendaciones de despliegue de recursos y permite exportar los resultados.

Caso de Uso 5: Solicitud de servicio de consultoría personalizada

- Actor involucrado: Representante institucional de un club o entidad gubernamental.
- Descripción: El sistema permite solicitar consultorías para personalizar el modelo con datos propios y obtener predicciones ajustadas al contexto de cada organización.
- Flujo de interacción:
    - El usuario autorizado ingresa a la plataforma mediante credenciales válidas.
    - Una vez validado el acceso, desde el menú principal accede al apartado de Consultoría.
    - Completa un formulario con los objetivos y el tipo de datos disponibles.
    - El sistema registra la solicitud y se comunica con el usuario vía mail para continuar el proceso.

Caso de Uso 6: Generación de reporte de información

- Actor involucrado: Planificador operativo de un club o institución gubernamental.
- Descripción: El usuario requiere generar un archivo con los resultados de predicciones y simulaciones realizadas, con el fin de contar con información estructurada para el análisis y la toma de decisiones operativas. Opcionalmente, tiene la posibilidad de acceder a un reporte personalizado del partido.
- Relaciones:
    - «include»: Predicción de asistencia.
    - «extend»: Solicitud de reporte específico por partido.
- Flujo de interacción:
    - El usuario autorizado ingresa a la plataforma mediante credenciales válidas.

- Una vez validado el acceso, desde el menú principal accede al apartado de Predicciones o Escenarios.
- El usuario simula la asistencia de un encuentro o escenario.
- El sistema ofrece opciones de reportes PDF o reporte personalizado.
- El usuario selecciona la opción deseada.
- El sistema genera un archivo para su descarga.

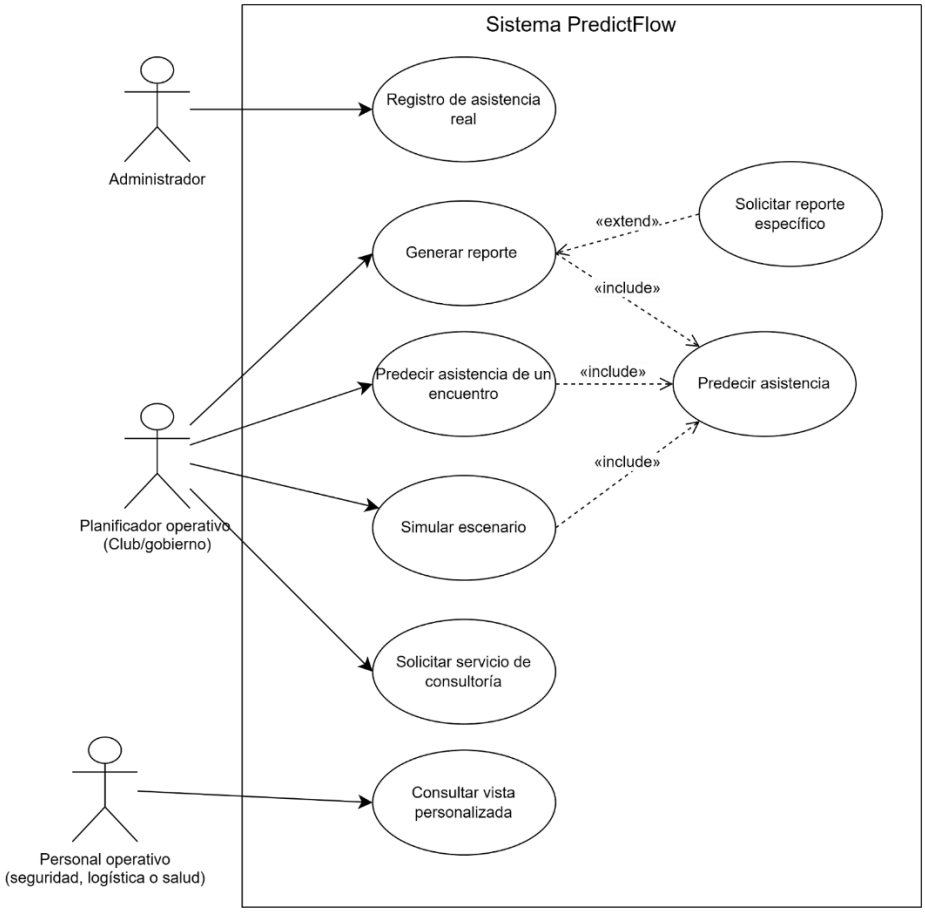


Figura 9. Diagrama de casos de uso - Fuente: Elaboración propia, 2025.

3.2. Atributos de Calidad

La calidad de un producto se compone de "*determinados atributos, los cuales proporcionan un modelo de referencia para que ésta sea especificada, medida y evaluada*"

(ISO/IEC, 2023). A continuación, se detallan los atributos principales que se encuentran presentes en PredictFlow.

3.2.1. Usabilidad

La usabilidad es el "*grado en el que un producto o sistema puede ser utilizado por usuarios específicos para alcanzar objetivos específicos con eficacia, eficiencia y satisfacción en un contexto de uso determinado*" (ISO/IEC, 2023). En este caso, la meta es que los usuarios puedan consultar predicciones sin necesidad de conocimientos técnicos.

La solución posee una interfaz clara y accesible, que presenta los datos de forma visual y sencilla. A su vez, se incorpora una sección destinada a la simulación de escenarios mediante la utilización de filtros desplegables (jornada futbolística, fecha, hora, equipo local y visitante), gráficos intuitivos para interpretar tendencias y explicaciones breves acerca de cada predicción.

3.2.2. Disponibilidad

La disponibilidad es la "*capacidad de un producto para estar operativo y accesible cuando se requiere su uso*" (ISO/IEC, 2023). En este sentido, un predictor de asistencia debe poder consultarse en todo momento que sea de necesidad para el usuario.

Para garantizar la disponibilidad del sistema, la infraestructura se implementa en la nube de AWS (Amazon Web Services), con instancias replicadas que aseguran la continuidad del servicio frente a fallos o posibles sobrecargas. Además, se incorpora un respaldo donde se almacenan predicciones recientes, para que, en caso de una interrupción en el servicio principal, los usuarios puedan acceder a información previamente generada.

3.2.3. Escalabilidad

La escalabilidad se refiere a la "*capacidad de un producto para manejar cargas de trabajo crecientes o decrecientes, o para adaptar su capacidad a fin de gestionar la variabilidad*" (ISO/IEC, 2023). Esto hace referencia a que la solución no solo debe funcionar en escenarios pequeños, sino también al crecer en tamaño y demanda.

En PredictFlow, la escalabilidad se sustenta a partir de una infraestructura en la nube que permite la asignación dinámica de recursos en función de la demanda. Bajo este enfoque, en escenarios con un alto volumen de usuarios realizando consultas de manera simultánea, el sistema tiene la capacidad de distribuir las peticiones entre distintos servidores para reducir demoras y disminuir el riesgo de colapsos.

3.2.4. Reusabilidad

La reusabilidad es la "*capacidad de un producto para ser utilizado como activo en más de un sistema, o en la construcción de otros activos*" (ISO/IEC, 2023). En el caso de PredictFlow, se implementa una estructura de datos que permite incorporar nuevas temporadas, variables adicionales o incluso extender el sistema a otras ligas y contextos deportivos sin necesidad de rediseñar toda la arquitectura. De esta manera, los componentes desarrollados no solo sirven para el escenario actual, sino que pueden adaptarse a nuevos entornos.

3.3. Arquitectura de la solución

El objetivo de la presente sección es describir en detalle la composición técnica de la solución propuesta, con el fin de explicar cómo cada uno de sus componentes contribuye al cumplimiento de los objetivos del proyecto. Se analizan los elementos que la integran y la forma en que interactúan entre sí, abarcando tanto la estructura general como los flujos de información que la sostienen.

Para ello, se adopta un enfoque progresivo que permite comprender la solución desde una perspectiva global y, posteriormente, profundizar en sus aspectos internos. En primera instancia, se expone la arquitectura de software completa, donde se describen los módulos principales, sus interrelaciones y las tecnologías que intervienen. Luego, se presenta una arquitectura basada en el modelo C4, que complementa la anterior al ofrecer una representación jerárquica y visual del sistema, facilitando su interpretación en distintos niveles de abstracción.

3.3.1. Arquitectura de Software

La arquitectura de PredictFlow se estructura en dos entornos diferenciados: el entorno de desarrollo y el entorno productivo, los cuales permiten mantener la independencia entre las etapas de experimentación, validación y operación del sistema.

En el entorno de desarrollo se realizan las tareas de preparación de datos, construcción del modelo predictivo y pruebas funcionales. Este espacio comprende tres componentes principales:

- **Fuentes de datos**: conformadas por los registros históricos de asistencia, estadísticas futbolísticas y datos meteorológicos obtenidos de servicios como *Transfermarkt*, Liga Profesional de Fútbol (LPF) y *Meteostat*.
- **Proceso ETL (Extracción, Transformación y Carga)**: encargado de la integración y depuración de los datos provenientes de múltiples fuentes. Mediante el uso de librerías como *Pandas* y *NumPy*, se ejecutan tareas de limpieza, transformación y generación de variables derivadas. Seleccionamos estas herramientas por su eficiencia en el manejo de grandes volúmenes de datos y su adopción dentro del ecosistema de *Python*.
- **Entrenamiento de modelos de Machine Learning**: donde se lleva a cabo el entrenamiento, validación y selección de modelos predictivos. Se utilizan algoritmos implementados en *scikit-learn*, con soporte del entorno *Python*, comparando métricas de desempeño para seleccionar el modelo con mejor capacidad predictiva. La elección de *Python* se relaciona con la gran variedad de librerías enfocadas en el análisis estadístico y aprendizaje automático.

Una vez completada esta etapa, los modelos entrenados se almacenan y se despliegan en el entorno productivo. El entorno productivo, implementado sobre infraestructura en la nube, está compuesto por cuatro bloques funcionales: frontend, backend, modelo de inteligencia artificial y base de datos.

- El frontend, desarrollado en *React*, actúa como interfaz principal para los usuarios, permitiendo visualizar predicciones, simular escenarios y acceder a reportes. Lo elegimos por su capacidad para construir interfaces interactivas y reactivas, optimizando la experiencia del usuario y simplificando el manejo de estado de toda de la aplicación.

- El backend, construido sobre *FastAPI*, expone endpoints seguros que gestionan la autenticación, el enrutamiento de solicitudes y la comunicación con el modelo predictivo y la base de datos. Lo elegimos como framework por su alto rendimiento y su integración con modelos de machine learning en *Python*. Dentro de este bloque se destacan tres subcomponentes:
    - Gestor de solicitudes, encargado de recibir y validar las peticiones del frontend.
    - Conversor de features, que adapta las variables ingresadas al formato requerido por el modelo.
    - Motor de inferencia, responsable de ejecutar el modelo y generar las predicciones.
- El modelo de IA se despliega como un servicio independiente, optimizado para recibir las características del partido y devolver las estimaciones de asistencia y ocupación en tiempo real.
- La base de datos *MySQL* centraliza la información histórica, los resultados de predicciones y los registros de uso de la aplicación, posibilitando el seguimiento y la trazabilidad de las operaciones. La seleccionamos por ser estable, robusta y fácil de integrar con entornos en la nube.

Para garantizar la continuidad operativa y el rendimiento del sistema, se implementan mecanismos de monitoreo y escalabilidad que permiten observar el consumo de recursos, latencias y tiempos de respuesta. En escenarios de alta demanda, tanto el backend como el servicio contenedor del modelo pueden escalar horizontalmente para mantener la disponibilidad del servicio.

La separación entre entornos de desarrollo y producción asegura un flujo de trabajo controlado, evitando que las modificaciones experimentales afecten la operación y garantizando la calidad en la entrega de nuevas versiones del sistema.

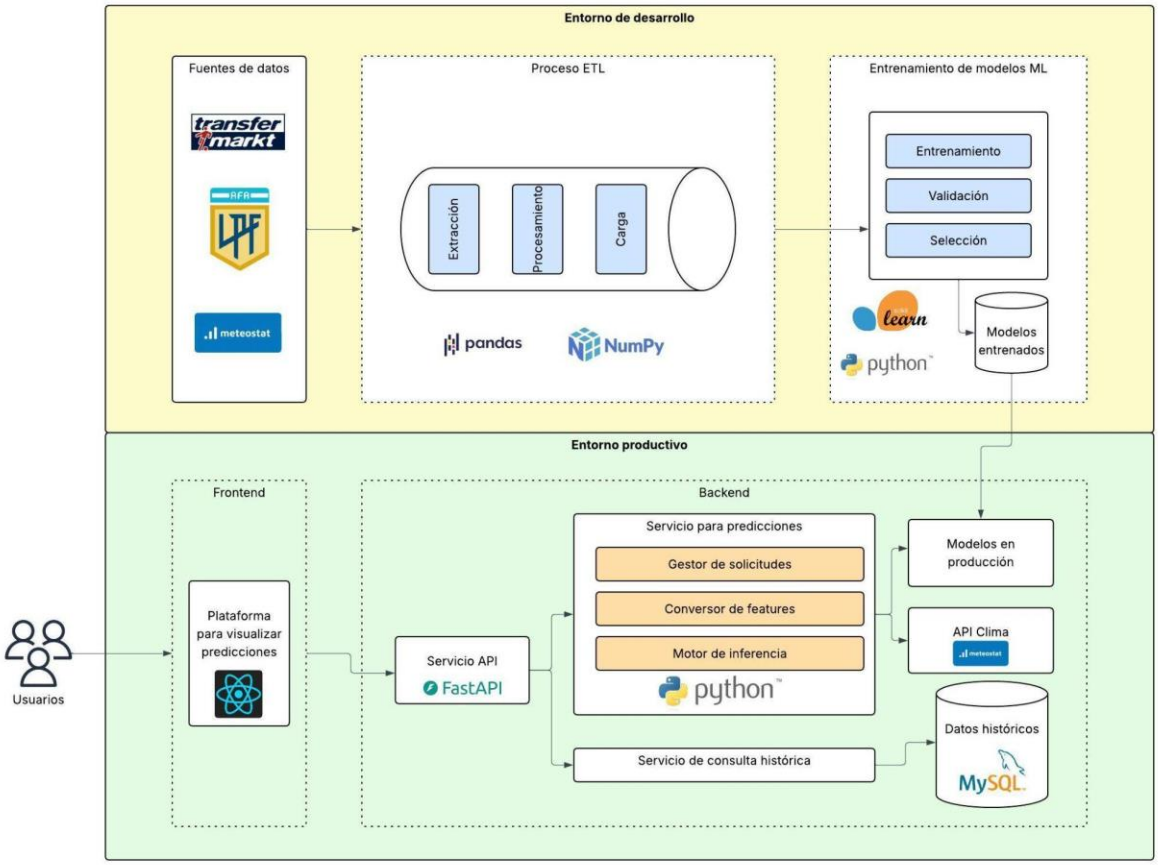


Figura 10. Arquitectura de software de PredictFlow - Fuente: Elaboración propia, 2025.

3.2.2. Modelo C4

El modelo **C4** constituye una metodología de representación visual orientada a la descripción estructurada de la arquitectura de software. Su finalidad es proporcionar una comprensión integral del sistema, permitiendo observar tanto su funcionamiento general como los detalles específicos de su implementación (Brown, 2024).

A diferencia de otros enfoques, el modelo C4 se fundamenta en una descomposición jerárquica que permite analizar el sistema desde su visión global hasta sus elementos más específicos, favoreciendo la comunicación entre equipos y la coherencia entre diseño e implementación. El modelo se estructura en cuatro niveles:

- **Contexto del sistema**: expone la posición del sistema dentro de su entorno operativo, identificando los actores externos y los sistemas con los que interactúa.
- **Contenedores**: describe los principales bloques funcionales del sistema, como aplicaciones, servicios y bases de datos. Define las responsabilidades y los flujos de información entre ellos.
- **Componentes**: detalla la estructura interna de cada contenedor, especificando los módulos y servicios que ejecutan funciones particulares. Explica cómo se distribuyen las responsabilidades dentro del sistema.
- **Código**: representa la arquitectura a nivel de código fuente, incluyendo clases, métodos y dependencias.

En el marco de este trabajo, el modelo C4 se aplicó para documentar la solución PredictFlow, elaborando los diagramas correspondientes a cada nivel.

Diagrama de Contexto

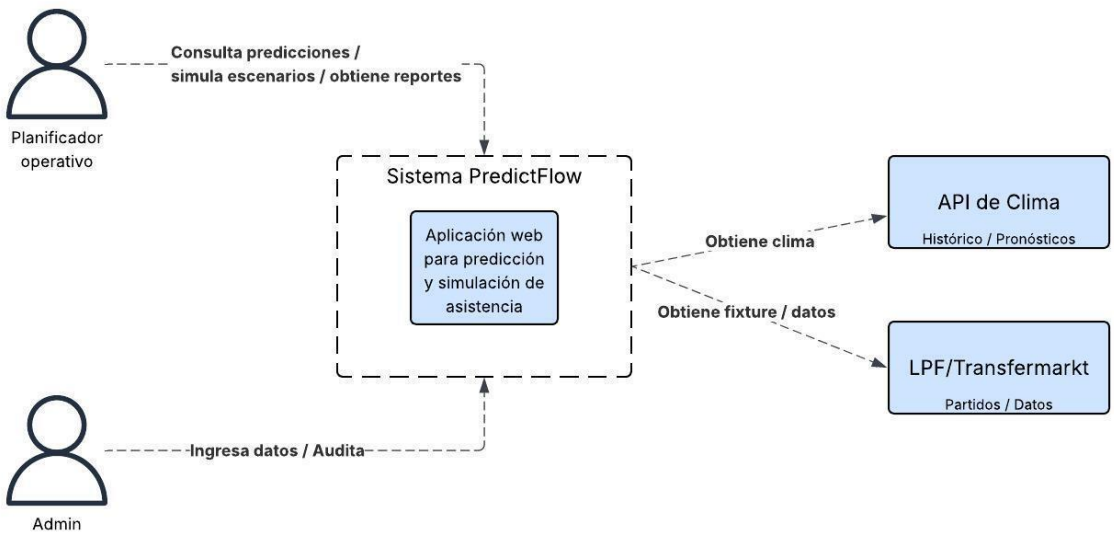


Figura 11. Diagrama de Contexto - Fuente: Elaboración propia, 2025.

- **Planificador operativo:** usuario que consulta predicciones, ejecuta simulaciones de escenarios y obtiene reportes para planificar la asignación de recursos en los partidos.

- **Administrador:** actor responsable de ingresar información, auditar los datos registrados y garantizar la calidad de los datos utilizada por el sistema.
- **Sistema PredictFlow:** aplicación web para predicción y simulación de asistencia, que actúa como núcleo y permite la interacción entre usuarios y servicios externos.
- **API de Clima:** servicio externo encargado de proveer información meteorológica histórica y pronósticos, utilizada para enriquecer las predicciones.
- **LPF/Transfermarkt:** fuente de datos externos que aporta información de partidos, fixture y estadísticas históricas necesarias para la estimación de asistencia.

Diagrama de Contenedores

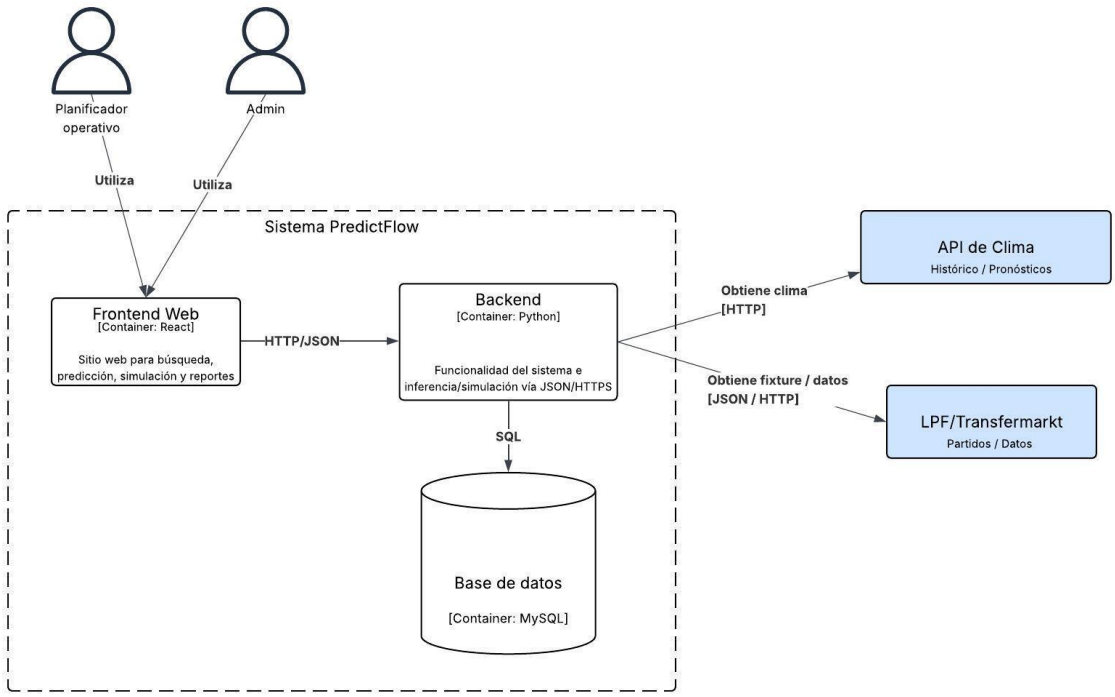


Figura 12. Diagrama de Contenedores - Fuente: Elaboración propia, 2025.

En este nivel se omiten los elementos Planificador operativo, Administrador, API de Clima y LPF/Transfermarkt, dado que ya fueron descritos en el diagrama de contexto. A continuación, se presentan únicamente los contenedores internos del sistema:

- **Frontend Web (React)**: aplicación web que constituye la interfaz principal para la interacción de los usuarios con el sistema. Permite realizar búsquedas, ejecutar predicciones, simular escenarios y acceder a reportes de asistencia.
- **Backend (Python)**: contenedor que maneja la lógica de PredictFlow. Gestiona la inferencia y simulación de datos, coordinando la comunicación entre el frontend, la base de datos y los servicios externos. Expone su funcionalidad a través de endpoints en formato HTTP/JSON.
- **Base de datos (MySQL)**: repositorio relacional utilizado para almacenar la información de equipos, partidos, fixtures, usuarios y registros de asistencia. Soporta operaciones de consulta y actualización que son ejecutadas por el backend.

Diagrama de Componentes

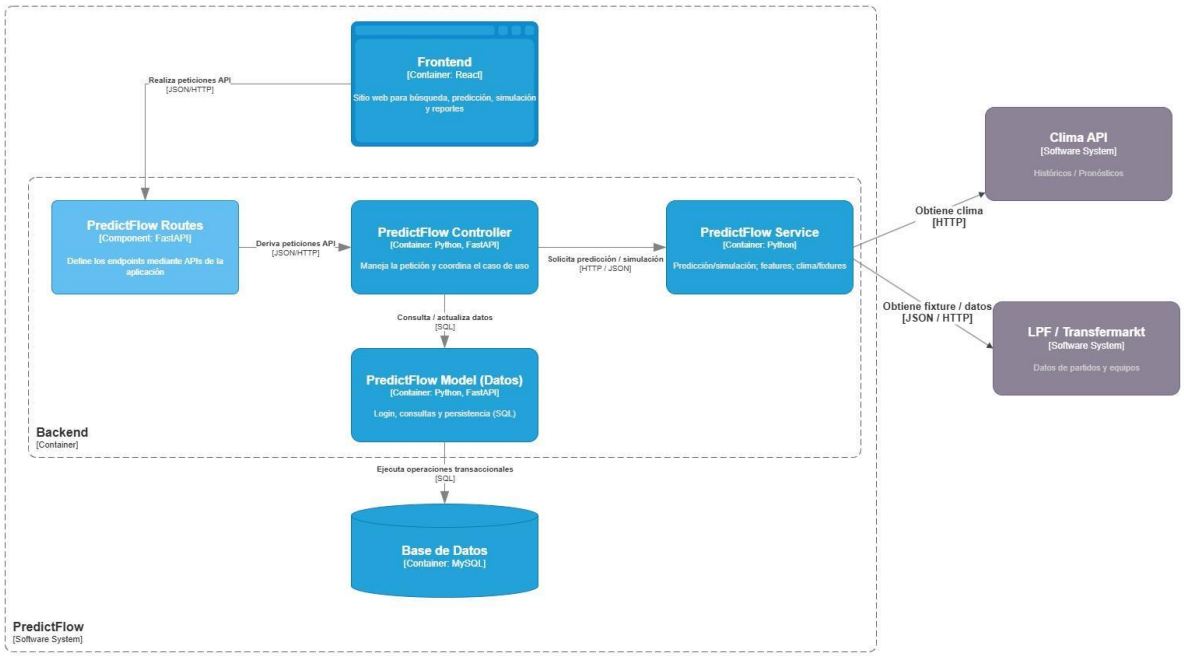


Figura 13. Diagrama de Componentes - Fuente: Elaboración propia, 2025.

En este nivel se omiten los elementos Planificador operativo, Administrador, API de Clima, LPF/Transfermarkt, Frontend Web y Base de Datos, ya explicados en los niveles anteriores. Se detallan únicamente los componentes internos del backend:

- **PredictFlow Routes (FastAPI)**: componente encargado de definir los endpoints de la aplicación y recibir solicitudes externas en formato HTTP/JSON. Actúa como punto de entrada para las interacciones del frontend.
- **PredictFlow Controller (Python)**: módulo que coordina los casos de uso del sistema. Recibe las solicitudes derivadas por las rutas, valida la lógica de negocio y distribuye las peticiones hacia los servicios de predicción o hacia la capa de datos.
- **PredictFlow Service (Predicción/Simulación)**: componente dedicado al procesamiento de predicciones y simulaciones. Construye las características necesarias (features) y ejecuta los modelos de machine learning que generan las estimaciones de asistencia y ocupación.
- **PredictFlow Model (Datos)**: capa de persistencia encargada de interactuar con la base de datos. Se ocupa de las operaciones de consulta y actualización en SQL, centralizando la gestión de datos históricos y actuales.

Diagrama de Secuencia

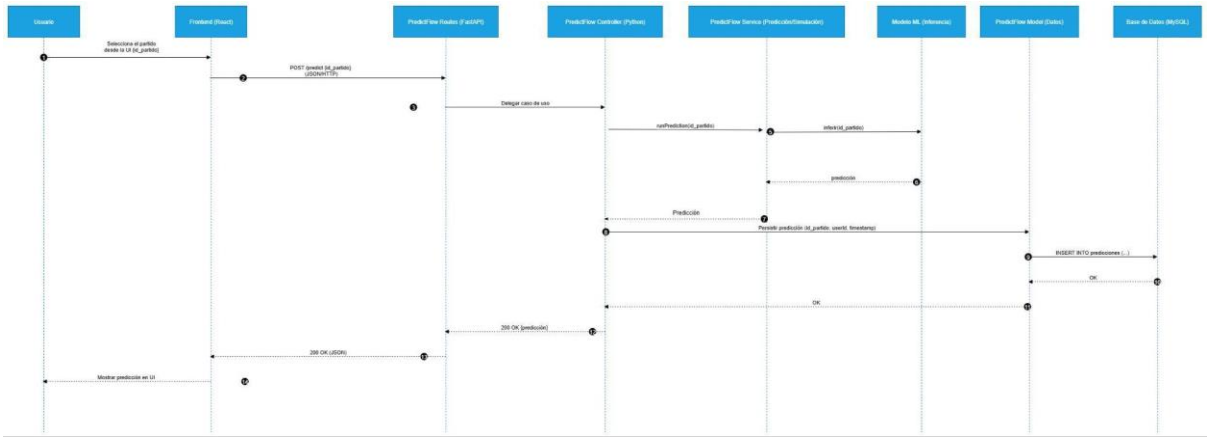


Figura 14. Diagrama de Secuencia - Fuente: Elaboración propia, 2025.

A continuación, se detallan los pasos específicos comprendidos en la Figura 14:

1. El usuario ingresa desde la interfaz a la sección de predicciones y selecciona el partido que desea predecir {id_partido}, requerido para la inferencia.
2. El Frontend (React) serializa número de partido y envía una solicitud HTTP POST a /predict (JSON/HTTP) hacia PredictFlow Routes.
3. PredictFlow Routes (FastAPI) recibe la petición y deriva el caso de uso al PredictFlow Controller.
4. El PredictFlow Controller valida la entrada e invoca al PredictFlow Service con runPrediction(id_partido).
5. El PredictFlow Service prepara las features y solicita la inferencia al Módulo ML (inferir(id_partido)).
6. El Módulo ML ejecuta el modelo entrenado y retorna la predicción al Service.
7. El PredictFlow Service persiste la predicción (id_partido, features, id_usuario, timestamp) llamando al PredictFlow Model (Datos).
8. En paralelo, el Controller queda a la espera de la respuesta de negocio (predicción) que viene del Service.
9. El PredictFlow Model (Datos) ejecuta el INSERT de la predicción en la Base de Datos (MySQL).
10. La Base de Datos confirma la operación (OK) al Model.
11. El PredictFlow Model propaga el OK al PredictFlow Service para cerrar la transacción lógica.
12. PredictFlow Routes responde al Frontend con 200 OK (JSON), incluyendo la predicción.
13. El Frontend presenta el resultado en la UI, actualizando la vista con la predicción obtenida.

3.3. Modelo de Datos

El sistema PredictFlow utiliza una base de datos relacional MySQL, la cual constituye el almacenamiento persistente de la información gestionada por la plataforma. Su diseño tiene como objetivo garantizar la integridad y trazabilidad de los datos.

El dominio de la solución abarca la información necesaria para analizar y estimar la asistencia a encuentros de la Liga Profesional de Fútbol Argentino. En particular, incluye datos vinculados a equipos, estadios, torneos y partidos, tanto disputados como programados, así como a los usuarios que acceden a la plataforma y a las predicciones y simulaciones de asistencia generadas por el sistema. Con el objetivo de reducir redundancias, evitar inconsistencias y facilitar el mantenimiento de la información, se consideró necesario definir un esquema de base de datos normalizado.

En una etapa inicial, la información de los partidos y sus predicciones se encontraba concentrada en estructuras no normalizadas, donde se encontraban datos de equipos, estadios, torneos y condiciones contextuales dentro de un mismo registro. Esta organización generaba redundancias de los atributos, dificultando el mantenimiento y la evolución del modelo de datos. A partir de esta situación, se llevó adelante un proceso de normalización hasta la Tercera Forma Normal.

En primer lugar, se garantizó el cumplimiento de la Primera Forma Normal, asegurando que cada atributo contenga valores atómicos y que cada tabla cuente con una clave primaria que identifique de manera única cada registro. Posteriormente, se abordó la Segunda Forma Normal mediante la eliminación de dependencias parciales, separando la información correspondiente a equipos, estadios y torneos en tablas independientes, evitando que atributos no dependientes de la clave principal permanecieran en la tabla de partidos.

Finalmente, el diseño se ajustó para cumplir con la Tercera Forma Normal, eliminando dependencias transitivas. En esta etapa se separaron conceptos como el tipo de usuario en una entidad específica, de modo que los atributos dependieran exclusivamente de la clave primaria de cada tabla y no de otros atributos no clave.

El modelo se compone de ocho tablas: Usuarios, Tipos_usuario, Equipos, Estadios, Torneos, Partidos, Predicciones y Escenarios_simulados. En la Figura 15 se presenta el diagrama entidad-relación que describe su estructura y vínculos.

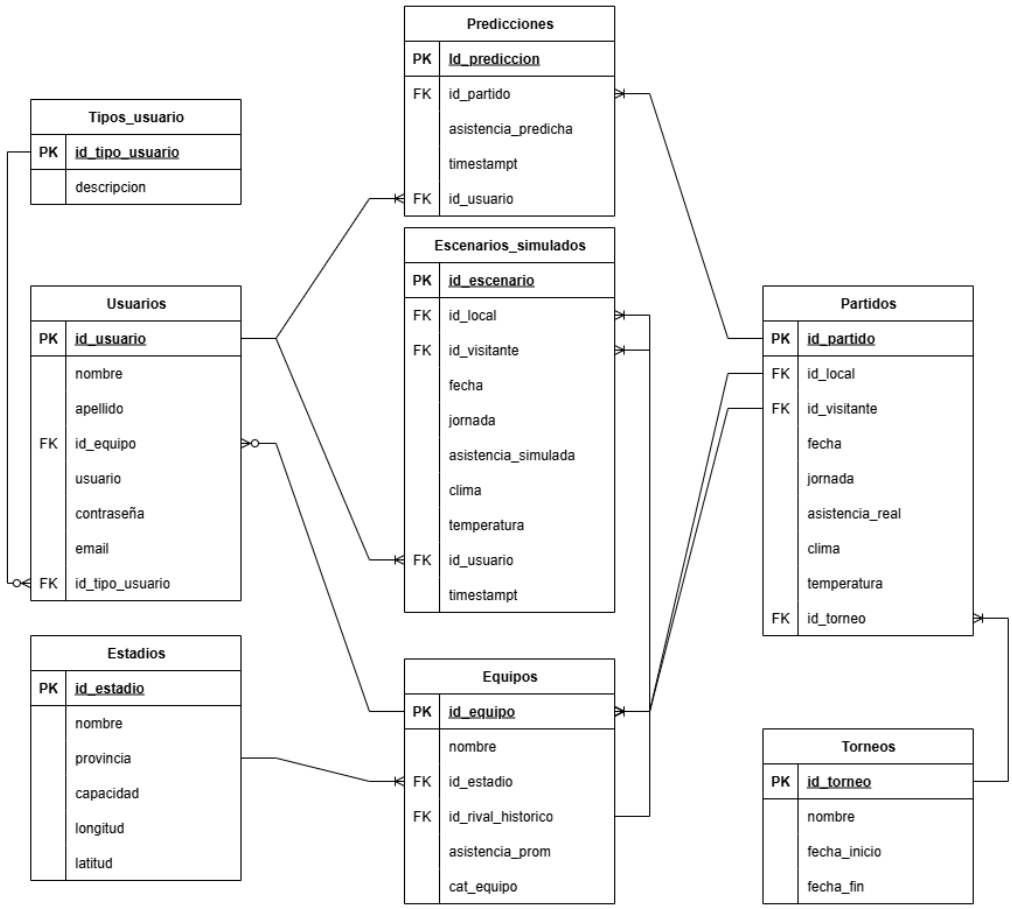


Figura 15. Modelo de datos relacional - Fuente: Elaboración propia, 2025.

A continuación, se describe el contenido y propósito de cada tabla:

- **Tabla Usuarios**: Contiene la información de las personas que interactúan con el sistema. Cada registro representa un usuario único, identificado por id_usuario. Incluye los campos nombre, apellido, usuario, contraseña y email. Cada usuario puede estar asociado a un equipo específico mediante la clave foránea id_equipo. Además, se le asocia un tipo de usuario para gestionar los roles y permisos en la aplicación.
- **Tabla Tipos_usuario**: Permite almacenar los distintos tipos de usuarios que tienen acceso al sistema.

- **Tabla Equipos**: Registra los clubes participantes del torneo. Se identifica mediante id_equipo y almacena su nombre, asistencia promedio (asistencia_prom), categoría (cat_equipo) y su vínculo con el estadio donde disputa sus partidos (id_estadio). También incluye una referencia a su rival histórico a través del campo id_rival_historico.
- **Tabla Estadios**: Contiene los datos de los escenarios donde se desarrollan los encuentros. Los campos principales son nombre, provincia, capacidad, longitud y latitud. Cada estadio puede estar vinculado a uno o varios equipos, lo que permite reflejar el uso compartido de instalaciones.
- **Tabla Torneos**: Agrupa los campeonatos o competiciones registradas en el sistema. Se identifica mediante id_torneo e incluye los campos nombre, fecha_inicio y fecha_fin, que determinan el período de duración de cada torneo.
- **Tabla Partidos**: Almacena los datos de los encuentros disputados o programados. Cada partido se identifica mediante id_partido y se relaciona con los equipos local y visitante (id_local, id_visitante). Además, registra información contextual como la fecha, jornada, condiciones climáticas (clima, temperatura), asistencia real (asistencia_real) y el torneo correspondiente (id_torneo).
- **Tabla Predicciones**: Registra los resultados generados por el modelo predictivo. Cada predicción está asociada a un partido (id_partido) y a un usuario (id_usuario). Incluye el campo asistencia_predicha, que almacena el valor estimado de asistencia, y un timestamp que indica el momento en que se generó la predicción.
- **Tabla Escenarios_simulados**: Permite almacenar los resultados de simulaciones creadas por los usuarios. Cada registro contiene la combinación de equipos (id_local, id_visitante), la fecha, jornada, variables contextuales (clima y temperatura) y la asistencia proyectada (asistencia_simulada). A su vez, incluye una referencia al usuario que ejecutó la simulación (id_usuario) y el momento en que se generó la simulación (timestamp).

3.4. Pipeline de Datos

El pipeline de datos tiene como objetivo integrar y transformar la información proveniente de distintas fuentes relacionadas con los partidos de la Liga Profesional, incorporando tanto los registros de asistencia como los atributos contextuales de equipos y estadios, para lograr un conjunto de datos listo para el entrenamiento de los modelos.

El proceso se estructuró siguiendo el enfoque ETL (Extract, Transform, Load), que organiza las tareas de extracción, transformación y carga de datos dentro de un flujo secuencial y modular. Cada fase fue definida de manera independiente, estableciendo un orden lógico de ejecución y un esquema de validaciones básicas sobre los datos procesados.

El diseño separa las etapas de extracción, transformación y carga. En la fase de extracción, se reúnen los datos provenientes de distintas fuentes relacionadas con los partidos, los equipos y los estadios, junto con información complementaria de tipo climático y temporal. Durante la fase de transformación, se ejecutan tareas de limpieza, normalización y enriquecimiento de los registros, que incluyen la depuración de duplicados, la estandarización de formatos y la incorporación de variables derivadas a partir de cálculos estadísticos y temporales. Finalmente, en la fase de carga, se integran todos los conjuntos procesados en una estructura tabular consolidada, donde cada registro representa un encuentro y cada columna una variable asociada, obteniendo así el conjunto de datos que se utiliza en el entrenamiento del modelo.

3.4.1. Extracción

La fase de extracción constituye el punto inicial del pipeline de datos y tiene por finalidad reunir, organizar y almacenar la información proveniente de diversas fuentes relacionadas con los encuentros disputados en la Liga Profesional de Fútbol Argentino. En esta instancia se recopilan los registros originales que posteriormente serán sometidos a procesos de transformación y análisis, preservando su estructura y formato sin aplicar modificaciones ni filtros previos.

El conjunto principal de datos está conformado por los registros históricos de partidos oficiales correspondientes al período 2022–2025, que incluyen información sobre los equipos participantes, el torneo, la jornada, la fecha, el estadio, las condiciones climáticas y la

cantidad de asistentes. Las temporadas 2020 y 2021 no fueron incorporadas debido a las restricciones sanitarias implementadas durante la pandemia de COVID-19, que afectaron la normalidad en la competencia y la disponibilidad de datos sobre asistencia. El período seleccionado dispone de registros continuos y comparables entre temporadas, lo que permite conformar una base homogénea de observaciones.

Junto con los datos de partidos, se incorporaron fuentes complementarias destinadas a enriquecer el contexto del análisis. Entre ellas se incluyen los registros de equipos y estadios, que aportan información sobre la capacidad habilitada de las instalaciones, la localización geográfica, la provincia y las rivalidades históricas entre clubes. También se integraron variables meteorológicas y temporales asociadas a la fecha y hora de los encuentros, obtenidas de registros climáticos históricos en función de la ubicación del estadio. Estas variables permiten disponer de datos contextuales relativos a las condiciones ambientales en el momento de cada evento.

Durante la fase de extracción, todas las fuentes fueron consolidadas en un entorno unificado de almacenamiento que reúne los registros en su formato original. En esta etapa se definieron las estructuras base que contienen los datos crudos, sobre los cuales se aplicarán las operaciones de limpieza, normalización y generación de variables en las fases siguientes. La organización de los conjuntos en este nivel inicial facilita la trazabilidad de los datos y su posterior actualización cuando se incorporan nuevas temporadas o fuentes de información.

En la Tabla II se presentan los campos principales que conforman el conjunto de partidos, mientras que la Tabla III detalla las variables complementarias correspondientes a los equipos, estadios y torneos.

Tabla II. Datos provenientes del dataset de partidos.

Nombre	Descripción	Tipo	Formato
Torneo	Nombre del torneo o campeonato correspondiente	Texto	Cadena de caracteres

Jornada	Número de la jornada del torneo	Número	Entero
Local	Equipo que actúa como local	Texto	Cadena de caracteres
Visitante	Equipo que actúa como visitante	Texto	Cadena de caracteres
Fecha	Fecha y hora del partido	Fecha	DD-MM-AAAA HH:MM
Estadio	Nombre del estadio donde se disputó el encuentro	Texto	Cadena de caracteres
Asistentes	Cantidad de público presente en el estadio	Número	Entero
Temperatura	Temperatura promedio al momento del encuentro	Número	Decimal
Lluvia	Condición climática (0 = no llovió, 1 = lluvia)	Binaria	0/1

Fuente: Elaboración propia, 2025

Tabla III. Fuentes de datos complementarias.

Fuente de datos	Descripción	Variables principales	Formato
Equipos	Información institucional de los clubes participantes en la Liga Profesional.	Nombre del equipo, estadio principal, rival histórico, asistencia promedio, categoría del club	CSV
Estadios	Registro de la capacidad, localización y denominación oficial de los estadios utilizados en los torneos.	Nombre del estadio, provincia, capacidad habilitada, coordenadas geográficas (latitud y longitud)	CSV
Torneos	Información sobre las competiciones disputadas entre 2022 y 2024.	Nombre del torneo, fecha de inicio, fecha de finalización	CSV

Fuente: Elaboración propia, 2025.

3.4.2. Transformación

La fase de transformación comprende el conjunto de operaciones destinadas a depurar, normalizar y organizar los datos extraídos en la etapa anterior, con el objetivo de obtener una estructura coherente y compatible con las fases posteriores del pipeline. En este punto se aplica el tratamiento necesario para convertir los registros crudos en un formato analítico uniforme, garantizando que las distintas fuentes se integren bajo una misma estructura tabular.

Las tareas realizadas incluyen la homogeneización de formatos y nomenclaturas, la corrección de inconsistencias detectadas durante la integración y la normalización de los valores numéricos y categóricos. Los nombres de equipos, estadios, torneos y provincias fueron estandarizados para evitar duplicaciones o discrepancias generadas por diferencias de escritura o codificación.

Las variables de fecha se transformaron a un formato cronológico uniforme, a partir del cual se derivaron atributos temporales de referencia, tales como año, mes, día de la semana y franja horaria. Se generó además un indicador binario que identifica los partidos

disputados durante el fin de semana, con el propósito de conservar esta información en las etapas posteriores.

Las variables numéricas vinculadas con la asistencia, la capacidad de los estadios y las condiciones climáticas fueron convertidas a un formato decimal estandarizado. Se aplicaron controles de rango y consistencia entre campos, verificando la correspondencia entre la asistencia y la capacidad del estadio, y descartando registros que no cumplían con los criterios básicos de integridad. En los casos en los que se detectaron valores faltantes o atípicos, se aplicaron imputaciones controladas cuando fue posible, empleando promedios históricos del equipo local o valores climatológicos representativos.

Durante esta etapa también se llevó a cabo el enriquecimiento contextual de los datos mediante la incorporación de variables de origen climático, obtenidas a partir de registros meteorológicos históricos en función de la ubicación geográfica de los estadios y de la fecha del encuentro. Este procedimiento permitió disponer de valores aproximados de temperatura promedio y condiciones atmosféricas asociadas a cada partido.

Finalizado el proceso de depuración y enriquecimiento, los registros fueron ordenados cronológicamente por equipo local, de modo que las observaciones mantuvieran la secuencia temporal de los partidos. Esta organización permitió asegurar que las variables dependientes del tiempo pudieran calcularse posteriormente sin recurrir a información de encuentros futuros.

El resultado de la fase de transformación es un conjunto de datos estandarizado, libre de duplicados y estructurado en un formato tabular en el que cada fila representa un encuentro y cada columna una variable asociada. Este conjunto constituye la base operativa sobre la que se desarrolla la etapa de generación de variables derivadas o feature engineering.

3.4.3. Generación de variables

La generación de variables constituye una sub fase de la etapa de transformación del pipeline ETL. En esta instancia se amplía el conjunto de datos previamente depurado mediante la creación de atributos derivados, construidos a partir de operaciones estadísticas, temporales y contextuales sobre los registros originales. Su propósito es representar de manera

estructurada la información histórica y las condiciones asociadas a cada partido, para que el conjunto resultante pueda ser utilizado en el modelado predictivo.

Las variables derivadas se elaboraron utilizando funciones de agregación, cálculos de promedio, medidas de tendencia y combinaciones entre atributos existentes. Cada transformación se aplicó respetando la secuencia temporal de los datos, de modo que la información utilizada para el cálculo de un registro correspondiera exclusivamente a partidos anteriores del mismo equipo. Este criterio permitió mantener la coherencia cronológica del conjunto de entrenamiento.

Las variables generadas se agruparon según su naturaleza en tres categorías principales, que organizan el origen y el tipo de información aportada: históricas del equipo local, contextuales del partido, y temporales y climáticas. Esta clasificación facilita la trazabilidad de las variables dentro del proceso y permite identificar el conjunto de factores incluidos en la modelización.

Históricas del equipo local: se calculan a partir de los registros previos del mismo club en condición de local. Incluyen el promedio ponderado de asistencia, la ocupación promedio del estadio, la tendencia de ocupación en los últimos tres partidos, la desviación estándar de asistencia en los últimos cinco encuentros, la variación de ocupación reciente y los días transcurridos desde el último partido. Estas variables representan de forma cuantitativa la evolución de la convocatoria del equipo local.

Contextuales del partido: incorporan información cualitativa sobre las características del encuentro y su relevancia deportiva. Entre ellas se encuentran las variables binarias que indican si el rival pertenece al grupo de equipos de alta convocatoria, si el enfrentamiento corresponde a un clásico y si dicho clásico se disputa durante el fin de semana. Estas variables se derivan de la relación entre los equipos participantes y del calendario de competencia.

Temporales y climáticas: describen las condiciones del entorno al momento del partido. Se incluyen variables como el mes, el día de la semana, la hora, la franja horaria (mañana, tarde o noche), la temperatura promedio y los indicadores de condiciones extremas, tales como temperaturas inferiores a 10 °C o partidos vespertinos con temperaturas elevadas.

Estas variables se obtienen a partir de la combinación entre la información temporal del calendario y los registros climáticos históricos asociados a la fecha y ubicación del estadio.

En esta sub fase también se prepararon las variables categóricas para su posterior utilización en los algoritmos de aprendizaje automático. Se aplicó la técnica de codificación One-Hot Encoding, que convierte cada categoría en un conjunto de variables binarias independientes, permitiendo que los algoritmos procesen información nominal sin introducir relaciones ordinales entre los valores (GERÓN, 2022).

El conjunto final resultante de esta etapa quedó conformado por 27 variables, integradas en un formato tabular que combina atributos históricos, contextuales, temporales y climáticos. Entre ellas se incluyen indicadores derivados de asistencia y ocupación del equipo local, variables cualitativas relacionadas con la relevancia del encuentro (clásico, rival de alta convocatoria) y factores ambientales y temporales asociados a la fecha y hora del partido.

La estructura completa de las variables que conforman el conjunto de entrenamiento se presenta en el Anexo D donde se documentan los nombres, descripciones, tipos de dato y categorías correspondientes.

3.4.4. Valoración

La valoración de los datos tuvo por finalidad examinar la composición y coherencia del conjunto final de entrenamiento, describiendo la distribución de las variables y las relaciones internas entre los factores históricos, contextuales, temporales y climáticos incluidos.

El dataset se conformó por 750 registros correspondientes a partidos disputados entre 2022 y 2024 inclusive, con 27 variables estructuradas en formato tabular. La variable objetivo, definida como porcentaje de ocupación del estadio, presentó un promedio cercano al 50 %, reflejando la diversidad de escenarios y niveles de convocatoria entre equipos.

El análisis descriptivo mostró valores mínimos de asistencia próximos a 450 espectadores y máximos en torno a 83 000, con una capacidad promedio de 37 000 localidades. Debido a esta variabilidad estructural, se empleó la ocupación relativa como medida de referencia, en lugar del número absoluto de asistentes.

Las variables de calendario presentaron una distribución equilibrada: aproximadamente el 55 % de los encuentros se disputó en fines de semana y un 65 % en horario nocturno. La temperatura media fue de 19 °C, con un 6 % de partidos bajo condiciones frías y un 10 % en situaciones de calor vespertino, lo que permitió conservar la variable continua de temperatura junto con los indicadores binarios de condiciones extremas.

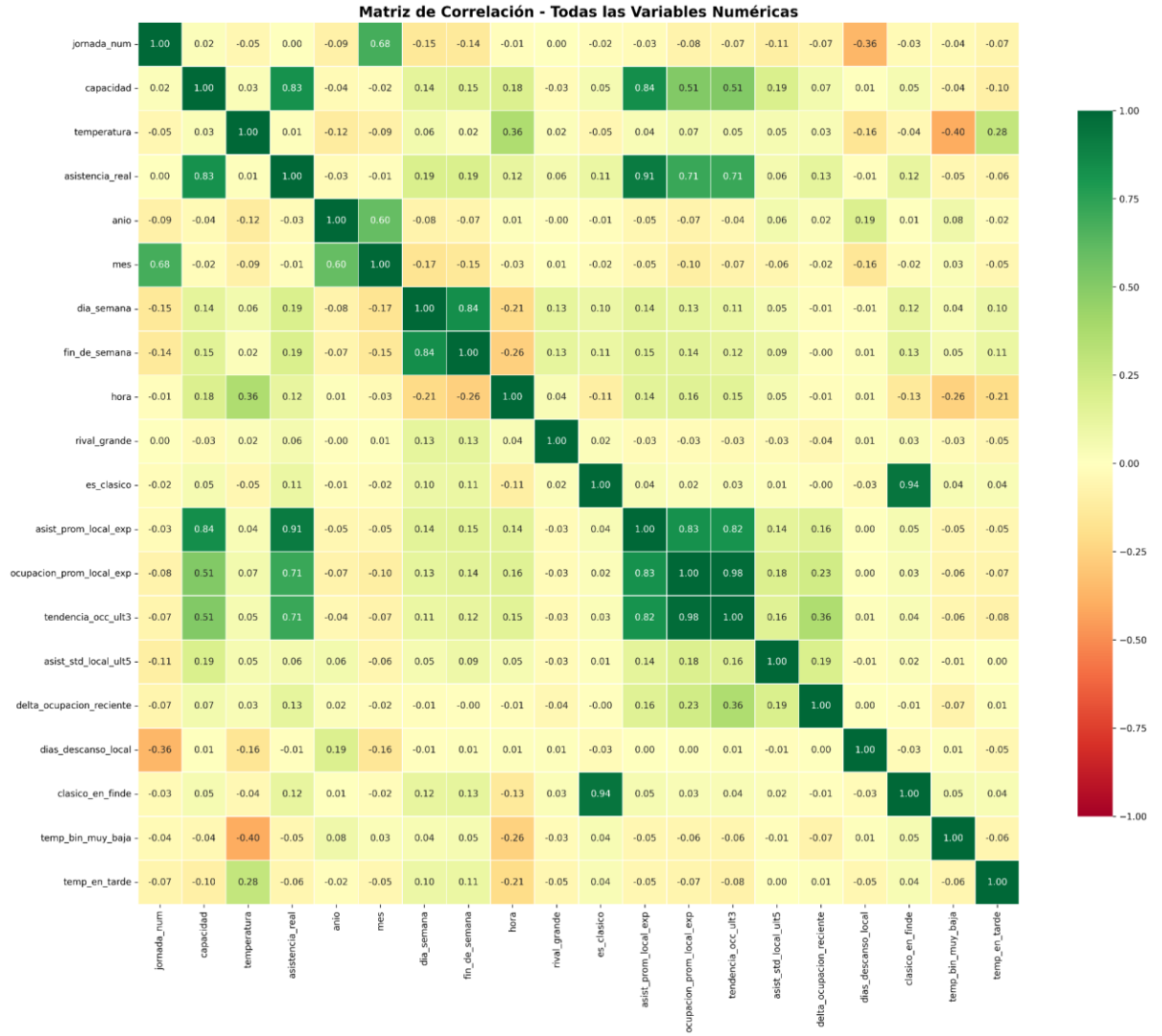
En el plano contextual, los partidos considerados como clásicos representaron el 3 % del total y los enfrentamientos con rivales de alta convocatoria, el 18 %. Estas proporciones fueron suficientes para mantener ambas variables sin afectar el equilibrio del conjunto.

Los indicadores históricos del público local mostraron correlaciones directas con la variable objetivo, por lo que se conservaron como principales predictores. Otras variables, como la variabilidad o el cambio reciente en la ocupación, presentaron menor relación, pero aportaron información complementaria sobre la estabilidad del comportamiento del público.

El análisis de correlaciones permitió identificar redundancias entre indicadores de naturaleza similar, priorizándose la conservación de aquellos más representativos. El conjunto final quedó compuesto por atributos numéricos, categóricos y binarios, sin registros duplicados ni valores faltantes.

La Figura 16 muestra la matriz de correlación elaborada sobre las variables numéricas del conjunto final de entrenamiento, utilizada para documentar las relaciones internas del dataset y verificar la coherencia del proceso de preparación de datos.

Figura 16. Matriz de correlación entre las variables numéricas del dataset.



Fuente: Elaboración propia, 2025.

3.5. Entrenamiento de modelos

La predicción de la asistencia de público a partidos de fútbol mediante técnicas de aprendizaje automático se plantea en este trabajo como un problema de regresión, dado que el objetivo consiste en estimar la cantidad de espectadores y el porcentaje de ocupación de los estadios. Este enfoque permite obtener valores continuos que representan cuantitativamente el nivel de concurrencia esperado en cada encuentro.

El conjunto de datos procesado en la etapa anterior constituyó el insumo principal para el entrenamiento de los modelos. Su estructura final, validada durante la valoración de los datos, presentó registros consistentes y relaciones verificables entre las variables históricas, contextuales, temporales y climáticas. La variable objetivo presentó una alta dispersión y las variables predictoras mostraron relaciones no lineales entre sí. En consecuencia, se incorporaron al proceso de entrenamiento algoritmos basados en árboles de decisión, que permiten operar con datos heterogéneos y modelar interacciones entre variables sin requerir transformaciones de linealidad.

Durante las etapas exploratorias del trabajo se realizaron pruebas con diversos modelos de regresión, incluyendo enfoques lineales regularizados, tales como Ridge y Lasso, y a su vez métodos basados en árboles de decisión, cómo lo son XGBoost, Random Forest y CatBoost, con el fin de analizar su comportamiento bajo un mismo esquema de validación temporal.

El proceso de entrenamiento se estructuró bajo un criterio temporal. El conjunto de datos se dividió de forma cronológica, asignando el 80 % inicial de los registros al entrenamiento y el 20 % restante a la evaluación. La división se realizó siguiendo la secuencia real de los encuentros, conservando el orden histórico y evitando la incorporación de información de períodos posteriores en la etapa de ajuste.

La variable objetivo utilizada corresponde al porcentaje de ocupación del estadio, obtenido mediante el cociente entre la asistencia registrada y la capacidad habilitada de cada sede. Los valores resultantes se expresaron en una escala relativa comprendida entre 0 y 1.

En la fase de preprocesamiento se aplicó codificación *One-Hot Encoding* a las variables categóricas, generando representaciones binarias independientes (GÉRON, 2022). Las variables numéricas se mantuvieron en su escala original, dado que los algoritmos empleados no requieren normalización.

El entrenamiento se realizó bajo un pipeline unificado, aplicando idénticas condiciones de preprocesamiento, validación y evaluación en los tres modelos seleccionados. Se registraron los parámetros de configuración, las métricas obtenidas y las versiones del entorno de ejecución con el fin de garantizar la trazabilidad y la reproducibilidad del proceso.

El ajuste de los hiperparámetros se efectuó mediante búsqueda aleatoria combinada con validación cruzada temporal, respetando el orden cronológico de los encuentros en cada iteración. La configuración final de los modelos se detalla en la Tabla IV.

Tabla IV. Configuración de hiperparámetros de los modelos seleccionados.

Modelo	Principales hiperparámetros
Random Forest	n_estimators = 800; max_features = sqrt; max_depth = None; min_samples_leaf = 2
XGBoost	n_estimators = 800; learning_rate = 0.1; max_depth = 10; reg_lambda = 2.0; subsample = 0.8; colsample_bytree = 0.8
CatBoost	iterations = 800; learning_rate = 0.05; depth = 8; loss_function = RMSE; l2_leaf_reg = 3.0
Ridge	alpha = 100.0 random_state = 42 tol = 1e-4
Lasso	alpha = 10.0 max_iter = 10000 random_state = 42

Fuente: Elaboración propia, 2025.

Todos los modelos se entrenaron bajo las mismas condiciones de validación, utilizando un conjunto común de métricas (RMSE, MAE, R^2 y MAPE) y la misma división temporal del conjunto de datos. Los resultados obtenidos durante la evaluación y la comparación de desempeño se presentan en el Capítulo 6 – Pruebas, donde se analiza el rendimiento de cada modelo y se explica la elección del modelo final.

Con base en los resultados del entrenamiento y la validación, se seleccionó el XGBoost Regressor como modelo final de predicción. Este modelo fue integrado al sistema PredictFlow por su buen nivel de precisión, estabilidad y tiempos de respuesta adecuados para su ejecución en producción.

3.6. Interfaz gráfica de usuario

En esta sección se detalla la interfaz gráfica del sistema PredictFlow, diseñada para facilitar la interacción de los usuarios responsables de la planificación y organización de partidos de la Liga Profesional de Fútbol. El sistema permite acceder a información histórica y proyectada, así como a herramientas de simulación para la estimación de asistencia y la gestión de recursos.

3.6.1. Panel de control

El Panel de Control representa una visión general del sistema y muestra los indicadores más relevantes de la temporada. Su propósito es ofrecer a quienes toman las decisiones información de valor que facilite la planificación estratégica y el seguimiento de la asistencia en los partidos. En la parte superior se presentan indicadores generales de referencia, mientras que en el centro se despliegan visualizaciones que permiten interpretar de manera rápida la evolución de la concurrencia y la comparación entre distintos encuentros. Además, se incluye un apartado que anticipa los próximos partidos con sus estimaciones proyectadas, lo cual permite a los organizadores anticipar escenarios y planificar la asignación de recursos necesarios. Finalmente, en la parte inferior se ofrece un acceso directo a las principales funcionalidades del sistema (Fixture, Predicciones y Escenarios).

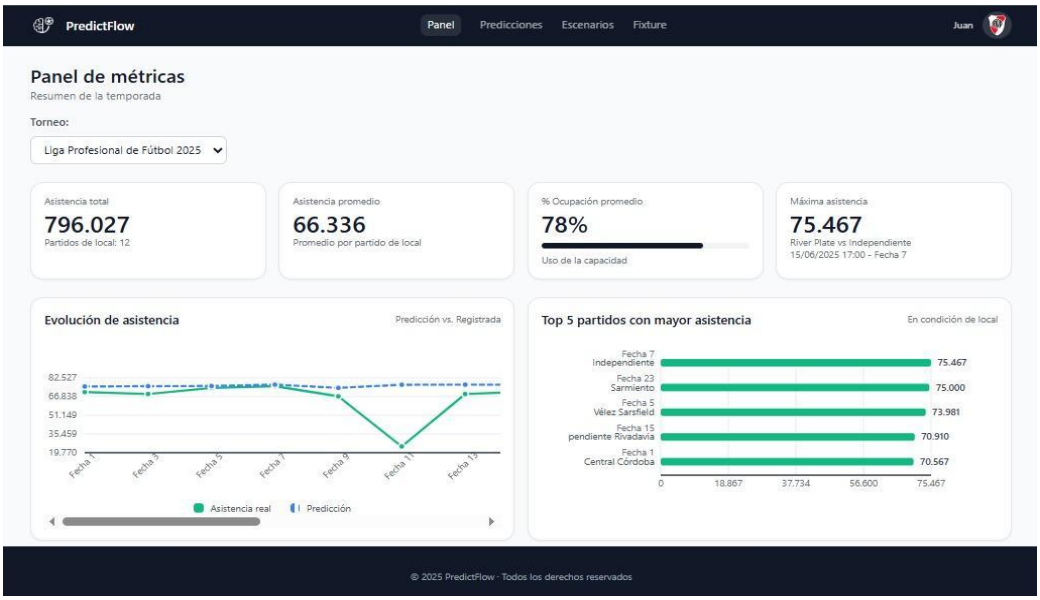


Figura 17. Panel de control de PredictFlow - Fuente: Elaboración propia, 2025.

3.6.2. Predicciones

La sección de Predicciones permite consultar estimaciones de asistencia para los partidos de local de cada club. Los resultados se presentan en un panel que muestra la proyección de ocupación del estadio, la cantidad estimada de asistentes y la variación respecto a la media histórica. Asimismo, el sistema incorpora un módulo de recomendaciones de recursos operativos, que incluye la demanda sugerida de efectivos de seguridad, puestos de acreditación, personal de orientación y personal médico. En la parte inferior se muestran tarjetas con los próximos partidos, cada una con acceso directo a la simulación correspondiente.

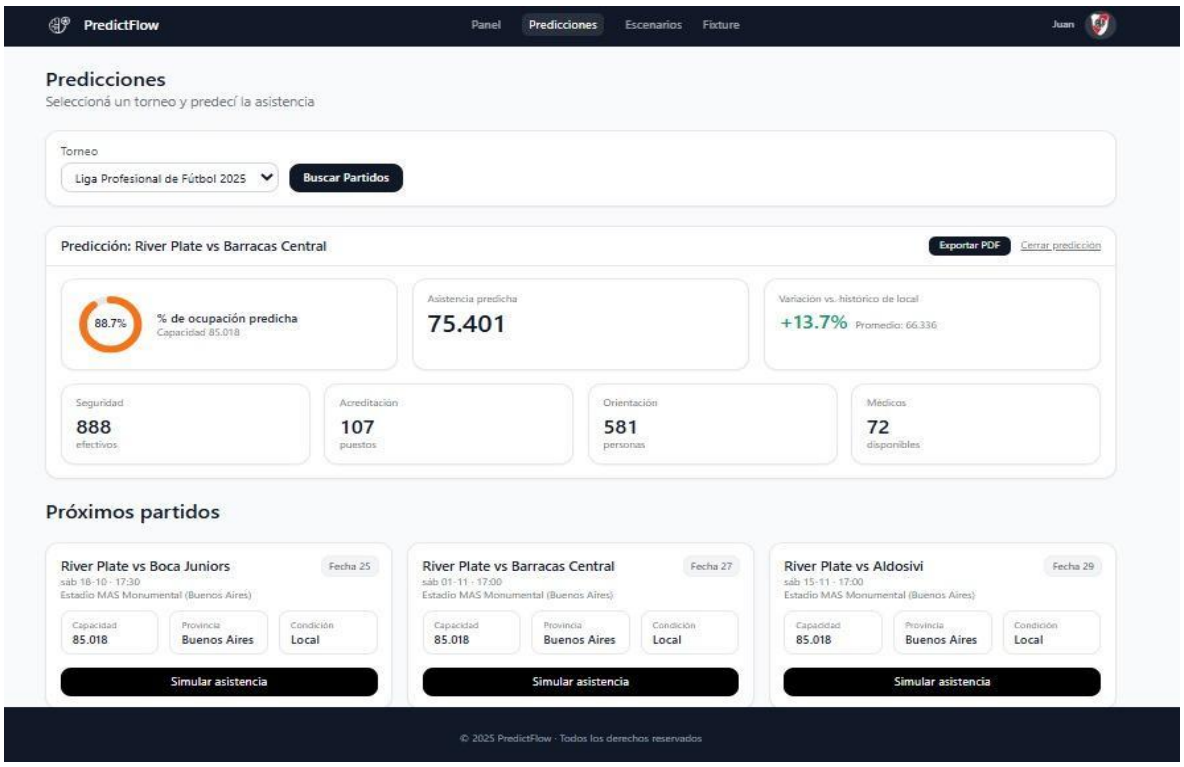


Figura 18. Módulo de predicciones de PredictFlow - Fuente: Elaboración propia, 2025.

3.6.3. Escenarios

La pantalla de Escenarios ofrece la posibilidad de simular diferentes condiciones para un partido específico, permitiendo modificar variables como los equipos participantes, la fecha, la hora y el clima. Una vez configurados los parámetros, el sistema proyecta los valores de asistencia, porcentaje de ocupación, variación respecto a antecedentes históricos y la demanda de recursos operativos.

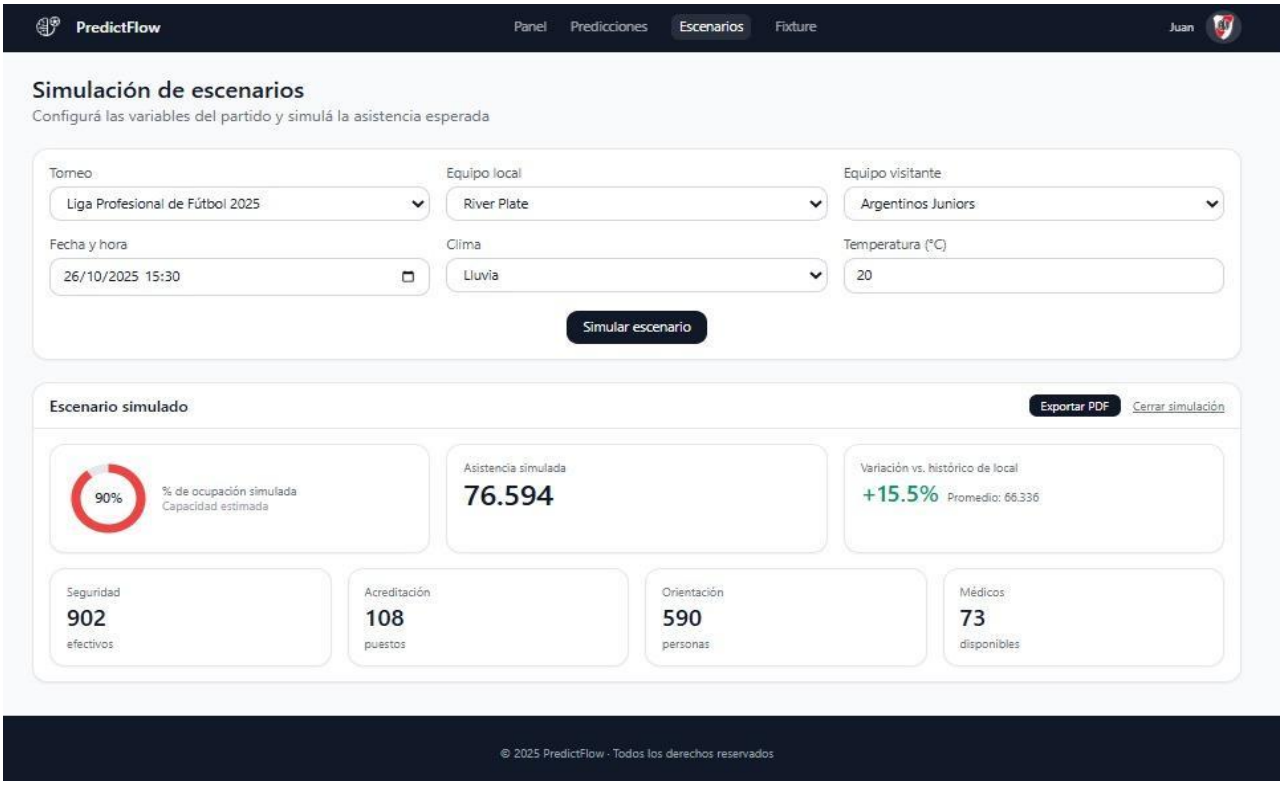


Figura 19. Módulo de escenarios de PredictFlow - Fuente: Elaboración propia, 2025.

3.6.4. Fixture

La pantalla de Fixture presenta un resumen de la temporada y el detalle de cada partido. En la parte superior se muestran indicadores generales, mientras que en la grilla se listan las jornadas con información correspondiente a fecha, hora, equipos, estadio, provincia y estado de la asistencia.

El sistema distingue entre partidos ya disputados, en los que se visualiza el número de asistentes registrados, partidos en los que aún no se cargó el dato, señalados como “Pendiente”, y partidos futuros, identificados como “Próximo”. Esta organización permite realizar un seguimiento histórico y operativo de la asistencia en cada jornada.

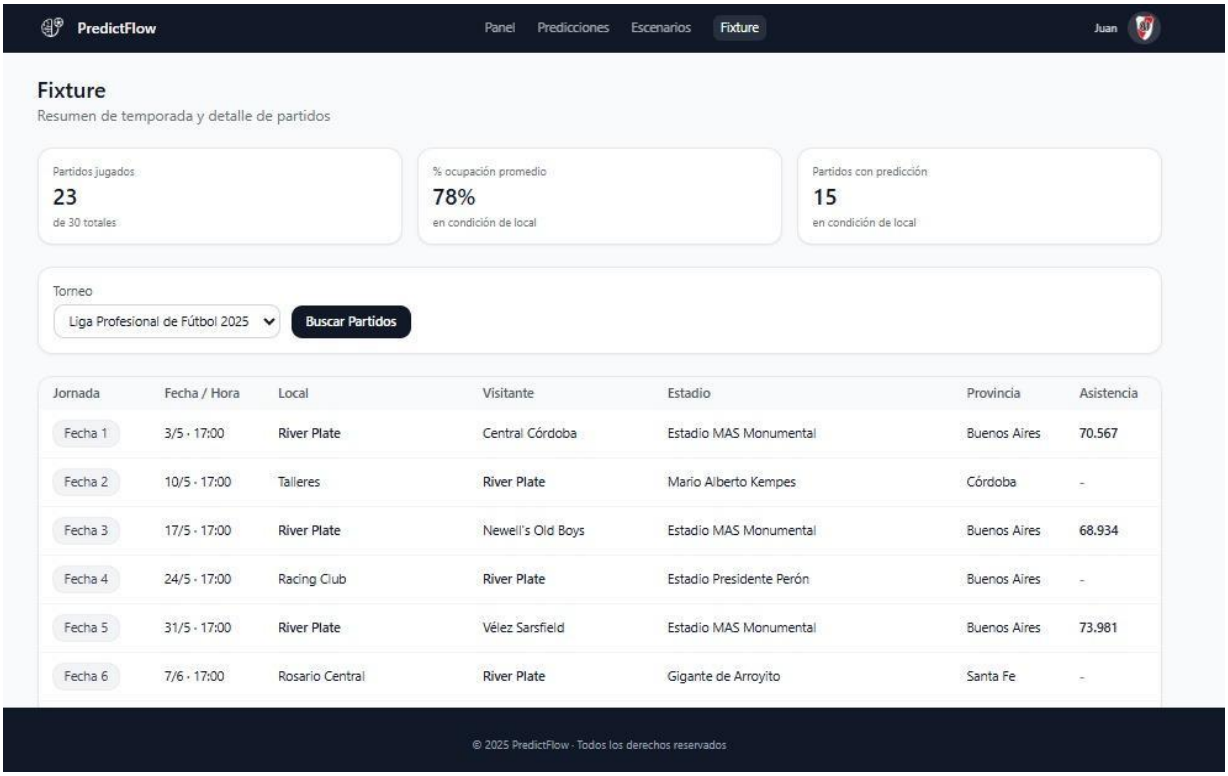


Figura 20. Módulo de fixture de PredictFlow - Fuente: Elaboración propia, 2025.

3.7. Plan de despliegue

El sistema se despliega sobre la infraestructura en la nube de *Amazon Web Services* (AWS), dentro de la región us-east-1. Todo se organiza dentro de una *Virtual Private Cloud* (VPC), que aísla los recursos y permite definir subredes públicas y privadas. En la subred pública se ubican los servicios que deben ser accesibles desde internet, mientras que la base de datos se aloja en una subred privada, protegida mediante reglas de seguridad y control de tráfico.

El frontend se aloja en *Amazon Simple Storage Service* (S3), donde se almacenan los archivos del frontend. Su distribución global se realiza a través de *Amazon CloudFront*, que asegura baja latencia y conexión segura mediante HTTPS. El dominio del sitio se administra con *Amazon Route 53*, que gestiona las solicitudes DNS de forma centralizada.

En la capa de aplicación, el backend se ejecuta sobre una instancia *Amazon Elastic Compute Cloud* (EC2), encargada de recibir las peticiones y coordinar el flujo de datos. Las interacciones con los modelos se procesan en *Amazon Elastic Container Service* (ECS),

donde se aloja el contenedor con el entorno de ejecución de los modelos de machine learning. Esta separación permite escalar de manera independiente el backend y los servicios de predicción, garantizando un uso eficiente de los recursos. Las peticiones externas se manejan a través de *Amazon API Gateway*, que expone los endpoints HTTPS y enruta las solicitudes hacia el backend.

Para el almacenamiento, se utiliza *Amazon Aurora MySQL Serverless v2*, ya que es una base de datos relacional que ajusta automáticamente su capacidad según la carga de trabajo. Está desplegada en la subred privada de la VPC y solo puede ser accedida por el backend, lo que garantiza la integridad y protección de la información.

Por último, *Amazon CloudWatch* se emplea para monitorear el rendimiento, detectar errores y optimizar los recursos utilizados.

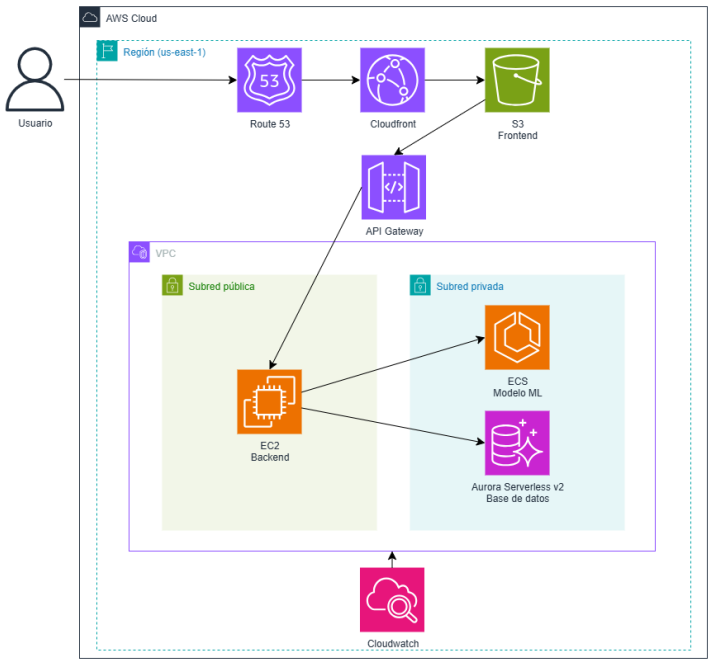


Figura 21. Arquitectura de PredictFlow en AWS - Fuente: Elaboración propia, 2025.

3.8. Marca

La identidad de marca de PredictFlow busca reflejar de manera coherente los valores y la propuesta del proyecto.

3.8.1. Logotipo

Figura 22. Logotipo de PredictFlow - Fuente: Elaboración propia, 2025.

El diseño del logotipo busca representar la unión entre inteligencia artificial y fútbol, los dos ejes centrales del proyecto. La imagen combina una mitad de cerebro, que simboliza la capacidad analítica, el procesamiento de datos y la tecnológica, con una pelota de fútbol haciendo referencia al propio deporte.

Figura 23. Paleta de colores de PredictFlow - Fuente: Elaboración propia, 2025.

El tono azul oscuro en la paleta de colores fue elegido porque transmite confianza, modernidad y se asocia con el enfoque tecnológico del proyecto. En resumen, el logotipo logra una imagen moderna, simple y reconocible dentro del ámbito tecnológico y deportivo.

3.9. Marco Legal

PredictFlow fue desarrollado como un sistema de análisis y predicción basado en información agregada, orientado a estimar la asistencia a encuentros deportivos a partir de datos históricos y variables contextuales. Desde el punto de vista legal, el encuadre del sistema se encuentra determinado por la naturaleza de los datos utilizados y por el alcance funcional definido para la solución.

Durante el desarrollo del proyecto se trabajó exclusivamente con registros de asistencia a nivel de partido, información deportiva de carácter público, características de los estadios y variables contextuales y climáticas. El sistema no procesa información que permita identificar personas físicas ni incorpora datos sensibles vinculados a espectadores, empleados u otros actores individuales.

Asimismo, PredictFlow no realiza cruces de información con bases de datos externas que administren datos personales, ni realiza integraciones con plataformas transaccionales, cómo la gestión de entradas, control de accesos o identificación de asistentes. Su alcance se limita a la generación de estimaciones estadísticas de asistencia, utilizadas como insumo para la planificación operativa y el análisis institucional.

Las predicciones generadas constituyen valores estimativos basados en patrones históricos y variables contextuales, cuya interpretación y utilización quedan bajo la responsabilidad de los usuarios que acceden a la plataforma. En este sentido, la solución se posiciona como una herramienta de apoyo analítico, sin intervenir directamente en procesos de control o ejecución.

Finalmente, el software desarrollado y los modelos predictivos asociados constituyen una producción de carácter técnico, científico y académico. En caso de una futura expansión funcional que implique el tratamiento de datos personales o la integración con sistemas que administren información que permita la identificación de individuos, será necesario adoptar políticas de gestión, seguridad y uso de la información alineadas con la normativa vigente en materia de protección de datos personales.

4. Metodologías de desarrollo

Para el desarrollo del proyecto se adoptó un enfoque basado en metodologías ágiles, con el objetivo de organizar el trabajo de manera iterativa e incremental. Este tipo de metodologías permitió planificar y ejecutar ciclos cortos de trabajo, teniendo en cuenta la adaptación a ciertos cambios resultantes y la generación progresiva de resultados. En particular, se aplicó el marco de trabajo *Scrum*, el cual estructura el desarrollo en etapas breves denominadas *sprints*, con roles definidos, reuniones periódicas y un seguimiento continuo de las tareas. Este marco permite "*generar valor a través de soluciones adaptativas para problemas complejos*" (SCHWABER y SUTHERLAND, 2020).

Dado que el equipo se conforma por dos integrantes, se adoptó una versión simplificada de Scrum, manteniendo los principios fundamentales, pero adaptando los roles al contexto del Proyecto Final de Ingeniería. La participación fue equitativa en todas las etapas distribuyendo las responsabilidades, tanto técnicas como documentales, priorizando la coordinación para cumplir con las entregas preliminares definidas en el cronograma provisto.

4.1. Fases del desarrollo

El desarrollo del proyecto se organizó en tres entregas preliminares y una entrega final, cada una con objetivos definidos y entregables concretos que reflejan el avance progresivo de la solución. Esta división permitió mantener una estructura de trabajo clara y verificar resultados en cada etapa:

- **Entrega Preliminar I:**

  En la etapa inicial se elaboró el marco teórico del proyecto, incorporando fundamentos sobre inteligencia artificial, predicción de asistencia y gestión de recursos en eventos deportivos. En paralelo, se llevó a cabo el estado del arte, analizando investigaciones previas y soluciones existentes en otros contextos. También se realizó el User Research, incluyendo entrevistas con involucrados en la organización de partidos con el fin de identificar necesidades reales y variables relevantes para la predicción, así como también una encuesta a asistentes frecuentes para obtener una validación propia de los actores principales, en este caso los espectadores. Los

resultados obtenidos resultaron fundamentales para definir los requerimientos iniciales del sistema.

- **Entrega Preliminar II:**

  En esta fase se diseñó la arquitectura técnica de la solución, estableciendo la relación entre las capas de datos, la lógica de negocio, los modelos predictivos y la visualización por parte de los usuarios. Se desarrollaron métodos para la recolección, limpieza y normalización de datos, integrando registros históricos de asistencia, condiciones climáticas y estadísticas deportivas. A partir de la base consolidada, se entrenaron modelos predictivos utilizando diferentes técnicas de machine learning para luego compararlos, evaluar sus desempeños y ajustar parámetros para optimizar los resultados.

- **Entrega Preliminar III:**

  Una vez validados los modelos, se implementaron los componentes del sistema. El backend, encargado de exponer los servicios de predicción mediante una API entre la base de datos, los modelos y la interfaz gráfica. En conjunto, se construyó el frontend, diseñado para la visualización de los resultados y la simulación de escenarios. Este módulo fue integrándose gradualmente con el backend durante los sprints. Al finalizar esta etapa, se consolidó un flujo funcional completo, partiendo desde la consulta de asistencia a un partido hasta la visualización de las predicciones.

- **Entrega Final:**

  En la fase final se espera refinar los modelos predictivos, optimizando hiperparámetros y variables de entrada para mejorar la precisión, para luego realizar pruebas para consolidar el rendimiento del modelo final. Asimismo, se realizará el despliegue completo sobre la infraestructura en AWS, configurando los entornos de desarrollo y producción. Este proceso va a permitir disponer de una versión funcional del sistema accesible en la nube, garantizando siempre la escalabilidad y disponibilidad tanto de la plataforma como de los resultados de las predicciones.

4.2. Aplicación del enfoque ágil

Durante el desarrollo del proyecto, la aplicación del enfoque ágil evidenció ciertos ajustes necesarios respecto al cronograma tentativo realizado previamente. Varias tareas modificaron su duración o momento de ejecución en relación con lo planificado al comienzo, principalmente porque en las primeras etapas la organización del trabajo se basó en estimaciones generales y un entendimiento parcial sobre el alcance técnico del sistema. Con el avance del proyecto y la experiencia adquirida en cada fase, fue posible identificar con mayor claridad las actividades prioritarias y así reorganizar el trabajo en función de ellas.

Este proceso de reajuste resultó clave para mantener la continuidad del desarrollo y optimizar los tiempos disponibles. La metodología ágil permitió adaptar la secuencia de tareas sin perder de vista los objetivos generales y específicos, a medida que se comprendía mejor el alcance y la relación entre los distintos componentes del proyecto. De este modo, el enfoque adoptado favoreció un ritmo de trabajo realista y ordenado, ajustado a las necesidades concretas que fueron surgiendo en cada etapa del proyecto.

5. Análisis Económico

El análisis económico del proyecto abarca tanto el modelo de negocio como el análisis financiero con el objetivo de evaluar su viabilidad, así como también la proyección a futuro.

5.1. Modelo de negocio

El modelo de negocio de PredictFlow se basa en la prestación de un servicio dirigido a aquellas organizaciones que participan directamente en la gestión y planificación de los encuentros deportivos en la Liga Profesional de Fútbol. La propuesta busca adoptar un esquema basado en suscripciones fijas (mensuales o anuales), complementado con servicios adicionales que pueden contratarse según las necesidades y requerimientos de cada entidad.

Además de la suscripción principal, el modelo propone las siguientes fuentes de ingresos:

- **Reportes específicos por partido**: posibilidad de generar análisis detallados y estadísticas avanzadas para encuentros en particular.
- **Vistas personalizadas por rol**: orientadas a diferentes áreas de seguridad, logísticas y médicas.
- **Servicios de consultoría**: permiten adaptar e implementar modelos personalizados con datos históricos propios de un club.

De esta manera, las suscripciones permiten contar con una base estable de ingresos, mientras que los servicios complementarios amplían las oportunidades de ganancias. Así, se propone una estructura sustentable que permitirá, en un futuro, escalar hacia distintas ligas futbolísticas tanto en el ámbito nacional como en el internacional.

El sistema está diseñado para ser utilizado por tres tipos principales de usuarios:

- **Clubes de la LPF:** utilizan la plataforma para estimar, a través de predicciones y simulaciones de escenarios, la cantidad de público esperada en cada partido, visualizar métricas y tendencias en el contexto de la competencia y registrar la asistencia real a los encuentros.
- **Organismos gubernamentales:** acceden a los reportes y simulaciones para definir la cantidad de efectivos policiales, personal de salud y recursos

logísticos necesarios en cada evento. También pueden observar el tablero principal con indicadores y gráficos en base a partidos próximos.

- **Fuerzas de seguridad, logística y personal médico:** utilizan las vistas personalizadas según su rol para simular la asistencia en escenarios concretos y determinar cuántos recursos propios se recomienda desplegar según las condiciones del partido.

Con el objetivo de complementar la propuesta previamente desarrollada, se elaboró un Business Model Canvas que permite identificar los segmentos de clientes, la propuesta de valor, los canales de distribución, las fuentes de ingresos, la estructura de costos, los recursos y actividades clave, así como las asociaciones estratégicas necesarias. De este modo, se refuerza la relación entre la estrategia de generación de valor y la viabilidad económica y operativa del proyecto:

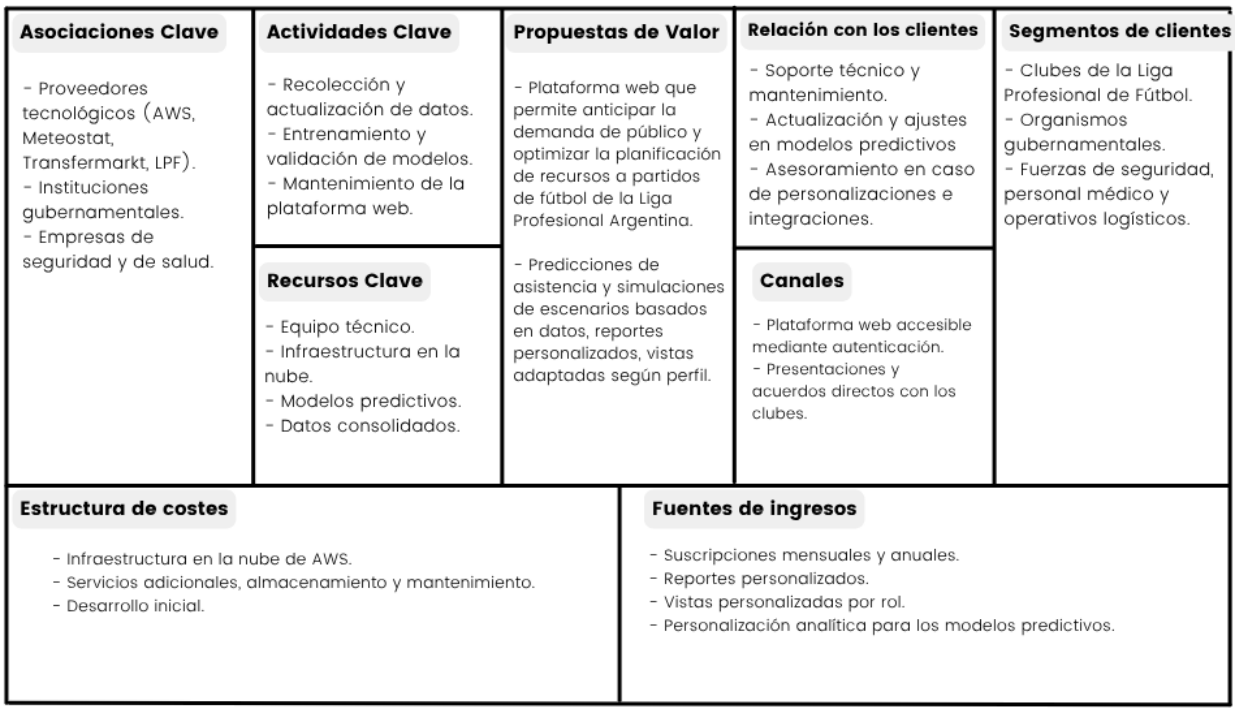


Figura 24. Business Model Canvas - Fuente: Elaboración propia, 2025.

5.2. Análisis financiero

Con el objetivo de evaluar la viabilidad económica del proyecto, se realizó un análisis financiero que contempla la inversión inicial, costos operativos, costos variables y fuentes de ingresos. Esto permite analizar la rentabilidad del modelo de negocio y proyectar distintos escenarios de crecimiento en relación con el nivel de adopción de la solución.

En la Tabla V se detallan los costos asociados a la inversión inicial, que contempla los recursos necesarios para el desarrollo del proyecto:

TABLA V: recursos para el desarrollo de PredictFlow.

Recurso	Descripción	Cantidad	Costo
Personal Humano	Desarrollador Frontend	50 horas	USD 625
Personal Humano	Desarrollador Backend	250 horas	USD 3.125
Computadora	Macbook Air M4	2 unidades	USD 2.000
Infraestructura	Servicios Cloud AWS	N/A	USD 240

Fuente: Elaboración propia, 2025.

Siguiendo con el análisis, los costos operativos permiten dimensionar los recursos necesarios para garantizar el correcto funcionamiento de la plataforma. En la Tabla VI, se detallan los costos fijos vinculados a los servicios de AWS:

TABLA VI: esquema de costos fijos de la infraestructura en AWS.

Servicio	Descripción	Frecuencia	Costo
Amazon S3	Aloja los archivos estáticos del sitio web	Mensual	USD 2
Amazon CloudFront	Distribuye el contenido del sitio de forma segura mediante una red CDN con HTTPS global	Mensual	USD 4
Amazon Route 53	Administra el dominio y las solicitudes DNS hacia el sitio y la API	Mensual	USD 1
Amazon API Gateway	Expone los endpoints del backend y enruta las peticiones hacia las funciones Lambda	Mensual	USD 3
Amazon EC2	Ejecuta el backend de la aplicación	Mensual	USD 2
Amazon ECS	Aloja y ejecuta los modelos de machine learning en contenedores	Mensual	USD 2
Amazon Aurora MySQL Serverless v2	Base de datos relacional principal del sistema	Mensual	USD 70
Amazon CloudWatch	Registra logs y métricas, permitiendo monitoreo y diagnóstico	Mensual	USD 2

Fuente: AWS Pricing Calculator, 2025.

Por otro lado, en la Tabla VII se pueden observar los distintos costos relacionados con el mantenimiento, almacenamiento y actualización de datos, así como también las altas de los nuevos clientes. Estos costos se consideran variables porque cambian según el nivel de actividad del proyecto. Cuando se dan de alta nuevos clientes o el tráfico de la aplicación supera los límites establecidos, estos costos aumentan. En cambio, cuando la demanda decrece, los mismos disminuyen:

TABLA VII: esquema de costos variables.

Descripción	Relación	Frecuencia	Costo
Margen operativo	Por cliente	Anual	USD 24
Alta cliente	Cantidad de altas	Única	USD 300

Fuente: Elaboración propia, 2025.

A continuación, se detallan los precios de los distintos tipos de ingresos que plantea el modelo de negocio. Estos valores se establecen en función de la magnitud económica que caracteriza al sector futbolístico y del margen de oportunidad existente, dado que la implementación de la solución propuesta permitiría generar ahorros significativos en comparación con los costos operativos actuales:

TABLA VIII: esquema de ingresos de PredictFlow.

Ingreso	Cliente	Frecuencia	Precio (USD)
Suscripción	Clubes y entidades gubernamentales	Mensual / Anual	USD 40 / USD 400
Reportes específicos por partido	Clubes y entidades gubernamentales	Única	USD 100
Vistas por rol	Áreas operativas	Manual	USD 50
Consultoría y personalización	Clubes y entidades gubernamentales	Única	USD 1.000

Fuente: Elaboración propia, 2025.

Se elaboraron tres escenarios posibles (pesimista, neutro y optimista) que permiten proyectar los ingresos, costos y flujos de fondos en un horizonte de evaluación de tres años. En cada caso se consideraron distintos niveles de adopción y crecimiento del proyecto dentro del ámbito futbolístico, donde el escenario pesimista muestra una incorporación más lenta y limitada por parte de algunos clubes y organismos, el neutro contempla una adopción estable y sostenida por una porción significativa de los equipos y el optimista plantea una expansión completa, en la que logra implementarse en todos los clubes de la Liga Profesional de Fútbol, consolidándose como una herramienta reconocida dentro del sector:

TABLA IX: escenario de la adopción pesimista.

Ingreso	**Año 1**	**Año 2**	**Año 3**
Suscripción	2 anuales	5 anuales	8 anuales
Reportes específicos por partido	2 totales	4 totales	7 totales
Vistas por rol	1 anual	3 anuales	4 anuales
Consultoría y personalización	0 totales	0 totales	1 total
Total	**USD 1.050**	**USD 2.550**	**USD 5.100**

Fuente: Elaboración propia, 2025.

TABLA X: escenario de la adopción neutra.

Ingreso	**Año 1**	**Año 2**	**Año 3**
Suscripción	5 anuales	10 anuales	15 anuales
Reportes específicos por partido	10 totales	20 totales	30 totales
Vistas por rol	3 anuales	6 anuales	8 anuales
Consultoría y personalización	1 total	2 totales	3 totales
Total	**USD 4.150**	**USD 8.300**	**USD 12.400**

Fuente: Elaboración propia, 2025.

TABLA XI: escenario de la adopción optimista.

Ingreso	**Año 1**	**Año 2**	**Año 3**
Suscripción	10 anuales	20 anuales	30 anuales
Reportes específicos por partido	20 totales	40 totales	60 totales
Vistas por rol	5 anuales	10 anuales	15 anuales
Consultoría y personalización	1 total	3 totales	5 totales
Total	**USD 7.250**	**USD 15.500**	**USD 23.750**

Fuente: Elaboración propia, 2025.

En base a los tres escenarios planteados, se realiza para cada caso un flujo de fondos y, con los resultados obtenidos, se calculan tres variables para cada uno de los escenarios: por un lado, el VAN, que representa la diferencia entre el valor actual de los flujos futuros y la inversión inicial, reflejando el valor que el proyecto agrega al capital invertido. Por otro lado, la TIR, que corresponde a la tasa de descuento que iguala esos flujos con la inversión, indicando el rendimiento porcentual esperado. Por último, el Payback, el cuál expresa el tiempo requerido para recuperar la inversión inicial a partir de los flujos netos de caja (GITMAN y ZUTTER, 2012). Esto puede observarse con detalle en el Anexo C.

Para el cálculo del VAN se adoptó una tasa de descuento del 10% anual, valor que supera el rendimiento de instrumentos considerados de bajo riesgo, como las Notas del Tesoro de los Estados Unidos a 10 años, cuya rentabilidad anual ronda el 4,13% (UNITED STATES DEPARTMENT OF THE TREASURY, 2025). Esta diferencia representa un valor promedio razonable para proyectos tecnológicos emergentes, equilibrando el riesgo operativo y el costo de oportunidad del capital en el contexto local.

TABLA XII: análisis del VAN, TIR y Payback sobre los escenarios.

Escenario	VAN	TIR	Payback
Pesimista	USD -3.915	-21%	No se recupera
Neutro	USD 7.084	53%	22 meses
Optimista	USD 19.441	108%	16 meses

Fuente: Elaboración propia, 2025.

A partir del análisis financiero realizado, se observa que el proyecto presenta una proyección económica favorable en los escenarios optimista y neutro, alcanzando un VAN estimado de USD 19.441, una TIR del 108% y un período de repago aproximado de 16 meses. Por lo contrario, bajo un escenario pesimista, el VAN se reduce a USD -3.915, la TIR es negativa (-21%) y no se alcanza el período de repago, lo que evidencia que la rentabilidad del proyecto se vería comprometida en contextos de baja adopción por parte de los usuarios. En función de esto, resulta fundamental implementar políticas y estrategias que fomenten la adopción temprana de la solución y fortalezcan su posicionamiento en el mercado.

Los resultados obtenidos en el análisis financiero evidencian que PredictFlow resulta un proyecto rentable en la mayoría de los escenarios, sustentable en el tiempo y cuenta con la capacidad de generar retornos significativos sobre la inversión en el mediano plazo.

6. Pruebas

El presente capítulo describe el conjunto de pruebas realizadas para validar el funcionamiento y la efectividad del sistema desarrollado. En primer lugar, se presentan los resultados del entrenamiento y evaluación de los modelos predictivos utilizados para estimar la asistencia. A continuación, se detallan las pruebas funcionales e integrales aplicadas sobre la plataforma PredictFlow, orientadas a verificar la usabilidad, la navegabilidad y la correcta integración entre los módulos de la aplicación.

6.1. Modelos entrenados y evaluación de resultados

El proceso de entrenamiento tuvo como objetivo evaluar el desempeño de distintos modelos de regresión aplicados a la estimación de asistencia en partidos de la Liga Profesional de Fútbol Argentino.

Se empleó un conjunto de 750 registros correspondientes a los torneos disputados entre los años 2022 y 2024, utilizando como variable objetivo el porcentaje de ocupación del estadio, posteriormente traducido a cantidad de asistentes estimados.

La división de los datos se realizó respetando la secuencia temporal de los encuentros, asignando el 80% inicial al entrenamiento y el 20% restante a la evaluación, con el fin de preservar la coherencia cronológica y evitar la fuga de información futura.

En la etapa experimental se probaron cinco modelos de regresión: Ridge, Lasso, Random Forest, XGBoost y CatBoost. Todos fueron entrenados con un mismo pipeline de preprocesamiento y validados mediante métricas estándar de regresión, como el error cuadrático medio (RMSE), error absoluto medio (MAE), coeficiente de determinación (R^2) y error porcentual medio (MAPE). Los hiperparámetros finales de cada modelo se definieron mediante una búsqueda aleatoria combinada con validación cruzada temporal, seleccionando las configuraciones que mostraron el mejor desempeño dentro del proceso de optimización.

Tabla XIII. Resultados comparativos de desempeño de los modelos de regresión

Modelo	**R^2**	**MAE**	**RMSE**	**MAPE (%)**
XGBoost	0.9174	3.258	4.692	25.45
Random Forest	0.9009	3.623	5.140	27.46
CatBoost	0.8969	3.576	5.245	30.30
Ridge	0.8665	4.368	5.867	39.43
Lasso	0.8642	4.344	6.017	38.14

Fuente: Elaboración propia, 2025.

Los resultados muestran que los modelos basados en árboles presentaron mejor capacidad de ajuste y generalización que los enfoques lineales.

El modelo XGBoost alcanzó el mayor coeficiente de determinación (R^2 = 0.9174) y los menores errores absolutos y cuadráticos, demostrando mayor precisión y estabilidad ante escenarios variables. El Random Forest logró un desempeño cercano, aunque con mayor dispersión frente a valores inusuales, mientras que CatBoost mantuvo buena precisión, pero con un mayor costo computacional. Los modelos Ridge y Lasso fueron útiles como referencia base, pero no lograron capturar las relaciones no lineales entre los factores históricos, contextuales y climáticos.

La importancia de variables obtenidas para el modelo XGBoost destaca como más influyentes la ocupación promedio local, la capacidad del estadio y la tendencia reciente de asistencia, seguidas por las condiciones climáticas y si el partido es un clásico disputado durante sábado o domingo.

En función de los resultados obtenidos, se seleccionó XGBoost Regressor como modelo final para la realización de predicciones y su posterior despliegue en producción. El modelo combina un nivel de precisión adecuado con buena capacidad de generalización y tiempos de inferencia compatibles con las necesidades operativas del sistema.

6.2. Pruebas funcionales

Las pruebas funcionales tuvieron como propósito verificar el cumplimiento de los requerimientos principales del sistema, asegurando que los módulos implementados operen correctamente y produzcan los resultados esperados.

Durante este proceso se evaluó la interacción completa entre los componentes de la aplicación, asegurando la correcta integración entre los mismos y a su vez, se realizó la evaluación de usabilidad de la aplicación.

Las validaciones se efectuaron en el entorno de desarrollo de la aplicación PredictFlow, garantizando condiciones equivalentes a las de uso productivo, con el fin de documentar empíricamente el comportamiento del sistema.

Tabla XIV. Casos de prueba funcionales

ID	Caso de prueba	RF	Descripción	Resultado esperado	Resultado obtenido
CP-01	Consulta de predicción de asistencia	RF-01	Se selecciona un partido y se solicita la predicción de asistencia y ocupación.	El sistema muestra los valores estimados sin errores, con tiempo de respuesta <3 s.	Satisfactorio.
CP-02	Simulación de escenarios	RF-02	Se modifican variables contextuales (día, horario, clima) y se recalcula la predicción.	Los resultados se ajustan coherentemente a las condiciones seleccionadas.	Satisfactorio.
CP-03	Consulta de vistas personalizadas por rol	RF-11	Se accede con distintos roles (seguridad, logística, salud) y se verifican las métricas específicas de cada perfil.	El sistema presenta correctamente las vistas e indicadores correspondientes al rol activo.	Satisfactorio.

CP-04	Registro de asistencia real	RF-04	Se ingresa la asistencia final del partido y se calcula la diferencia con la predicción.	El registro se almacena correctamente para análisis posteriores.	Satisfactorio.
CP-05	Solicitud de reporte específico por partido	RF-05	Se genera un reporte detallado con información histórica y contextual.	El reporte se muestra y exporta sin errores.	Satisfactorio.

Fuente: Elaboración propia, 2025.

En conjunto, las pruebas permitieron comprobar el funcionamiento general del sistema y la coherencia de los resultados obtenidos en las distintas funcionalidades. El sistema mostró un comportamiento estable y tiempos de respuesta adecuados durante la ejecución de los casos de prueba. Como mejora posible, se identificó la conveniencia de agregar mensajes informativos y leyendas en las métricas mostradas, para facilitar la interpretación de los resultados por parte de los usuarios.

7. Discusión

En proyectos basados en técnicas de aprendizaje automático, la disponibilidad, calidad y consistencia de los datos condicionan de manera directa el desarrollo del sistema. En el caso de PredictFlow, las primeras dificultades estuvieron asociadas a la obtención de datos históricos de asistencia, los cuales no se encuentran centralizados ni estandarizados en una única fuente. La recopilación de los mismos requirió integrar registros provenientes de distintas fuentes, así como definir criterios de exclusión para aquellos datos que no representaban condiciones normales de concurrencia, lo que redujo la cantidad de registros disponibles para el entrenamiento del modelo.

Una vez conformado el conjunto inicial de datos, el principal desafío se trasladó al proceso de preparación y transformación, que incluyó tareas de limpieza, normalización y validación cruzada entre partidos, estadios y equipos. Durante esta etapa se identificaron inconsistencias y valores faltantes que debieron ser corregidos para garantizar la coherencia del dataset. Asimismo, fue necesario generar variables derivadas para capturar el comportamiento histórico de la asistencia, lo que evidenció la importancia del preprocesamiento en el desarrollo del sistema.

Durante el análisis de los datos, fue posible comprender la relación existente entre las variables explicativas y la variable objetivo, así como su influencia en la definición del problema. A partir de este análisis se estableció el enfoque adoptado y la forma de representar la asistencia, priorizando métricas relativas como el porcentaje de ocupación del estadio frente a valores absolutos, como la cantidad de asistentes, de acuerdo con las características del dominio y la información disponible.

En relación con el alcance del proyecto, si bien la propuesta original contemplaba la incorporación de fuentes adicionales y funcionalidades más avanzadas en etapas posteriores, el desarrollo efectivo del sistema se concentró en los componentes definidos para el primer release. Esta delimitación respondió a la decisión de priorizar el desarrollo en el núcleo funcional del sistema, postergando la incorporación de otras funcionalidades para etapas posteriores. En este sentido, futuras iteraciones podrían incorporar estas funcionalidades para enriquecer el sistema aumentando el valor agregado de la solución.

En conjunto, el desarrollo del proyecto mostró que la implementación de una solución predictiva en un contexto real requiere adaptar el alcance a las limitaciones técnicas y a la disponibilidad de información, priorizando la solidez del modelo y su aplicación práctica.

8. Conclusiones

La asistencia a los encuentros futbolísticos en Argentina representa un desafío constante para la organización, debido a que diversos factores como el rival, la importancia del partido o el clima generan variaciones que son difíciles de anticipar. Estas limitaciones impactan directamente en la planificación de recursos y en la experiencia del público.

El desarrollo del presente trabajo permitió relevar y comprender las principales dificultades en la organización de los partidos de la Liga Profesional de Fútbol, tanto desde la perspectiva de los organismos involucrados como desde la experiencia del espectador. De este modo, se reafirma la necesidad de una herramienta de predicción de asistencia que anticipe la demanda y optimice la asignación de recursos.

Frente a esta problemática, se diseñó una solución que integra la recolección y procesamiento de datos históricos y contextuales, el entrenamiento de un modelo de predicción basado en técnicas de aprendizaje automático, y una arquitectura de software que permite consultar escenarios, generar predicciones y visualizar los resultados de manera simple e intuitiva.

Desde una perspectiva técnica, el sistema demostró resultados acordes con los objetivos planteados, validando la eficacia del modelo predictivo y la solidez de la arquitectura desarrollada. A su vez, el análisis financiero evidenció que PredictFlow es un proyecto viable y sostenible, con potencial de rentabilidad y expansión dentro del ámbito futbolístico nacional. Su implementación permitiría reducir costos operativos, optimizar la asignación de recursos y mejorar la experiencia del público, contribuyendo al profesionalismo en la gestión de eventos deportivos.

Si bien el proyecto puede seguir perfeccionándose, los avances alcanzados constituyen una base sólida para futuras mejoras y adaptaciones a otras ligas del país o de la región. En definitiva, el trabajo realizado no solo busca cumplir los objetivos académicos propuestos, sino también aportar una solución tecnológica escalable y de impacto real, orientada a transformar la forma en que se planifican y gestionan los espectáculos futbolísticos en Argentina.

9. Bibliografía

AL-BUENAIN, Ahmad, HAOUARI, Mohamed y JACOB, Jithu Reji. *Predicting Fan Attendance at Mega Sports Events - A Machine Learning Approach: A Case Study of the FIFA World Cup Qatar 2022*. En: *Mathematics* [en línea]. 2024, vol. 12, n. 6 [consulta: 10 junio 2025]. Disponible en: https://www.mdpi.com/2227-7390/12/6/926

AMAZON WEB SERVICES. AWS Pricing Calculator [en línea]. Seattle: Amazon Web Services, 2025 [consulta: 9 octubre 2025]. Disponible en: https://calculator.aws/

BISHOP, Christopher M. *Pattern Recognition and Machine Learning*. New York: Springer, 2006. ISBN 978-0-387-31073-2.

BREIMAN, Leo, *et al*. *Classification and Regression Trees*. Belmont: Wadsworth International Group, 1984. ISBN 978-0412048418.

BROWN, Simon. *The C4 model for visualising software architecture*. [en línea]. 2024 [consulta: 1 de octubre de 2025]. Disponible en: https://c4model.com

GÉRON, Aurélien. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 2a ed. Sebastopol: O'Reilly Media, 2019. ISBN 978-1492032649.

GITMAN, Lawrence J. y ZUTTER, Chad J. *Principios de administración financiera.* 13a ed. México: Pearson Educación, 2012. ISBN 978-607-32-0432-8.

IBM. *What is Supervised Learning?* [en línea]. [s. l.]: IBM, 2024 [consulta: 15 junio 2025]. Disponible en: https://www.ibm.com/think/topics/supervised-learning

ISO/IEC 25010:2023, *Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Product quality model*. 2a ed.

ISO/IEC 25019:2023, *Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Quality-in-use model*. 1a ed.

MITCHELL, Tom M. *Machine Learning*. New York: McGraw-Hill, 1997. ISBN 978-0070428072.

PANG, Yu y WANG, Fengchen. *Forecasting Stadium Attendance Using Machine Learning Models: A Case of the National Football League*. En: *Studia Sportiva* [en línea].

2024 [consulta: 10 junio 2025].
Disponible en: https://journals.muni.cz/studiasportiva/article/view/38549

PERUZZO, E. y ASANI, A. *Predictive Stadium Attendance Using Machine Learning* [en línea]. Padua: Università degli Studi di Padova, 2024 [consulta: 10 junio 2025]. Disponible en: https://www.research.unipd.it/retrieve/47ccd238-5a27-4354-95e5-dd465da09313/unpaywall-bitstream-805365196.pdf

RUSSELL, Stuart y NORVIG, Peter. *Artificial Intelligence: A Modern Approach*. 4a ed. [s. l.]: Pearson, 2020. ISBN 978-0134610993.

SAMUEL, Arthur L. *Some Studies in Machine Learning Using the Game of Checkers*. IBM Journal of Research and Development, 1959, vol. 3, n. 3, p. 210-229. DOI 10.1147/rd.33.0210

SCHLOSSER, Tobias, *et al*. *A Consolidated Overview of Evaluation and Performance Metrics for Machine Learning and Computer Vision*. ResearchGate [en línea]. 2024 [consulta: 14 junio 2025]. Disponible en: https://www.researchgate.net/publication/374558675_A_Consolidated_Overview_of_Evaluation_and_Performance_Metrics_for_Machine_Learning_and_Computer_Vision

SCHWABER, Ken y SUTHERLAND, Jeff**.** *La Guía de Scrum: la guía definitiva del marco de trabajo Scrum, las reglas del juego* [en línea]. Scrum.org y ScrumGuides.org, 2020 [consulta: 1 octubre 2025]. Disponible en: https://scrumguides.org/docs/scrumguide/v2020/2020-Scrum-Guide-Spanish-Latin-South-American.pdf

UNITED STATES DEPARTMENT OF THE TREASURY [en línea]. Washington, D.C.: U.S. Department of the Treasury, 2025 [consulta: 9 octubre 2025]. Disponible en: https://home.treasury.gov/

YAMASHITA, Gabrielli, *et al*. *Customized prediction of attendance to soccer matches based on symbolic regression and genetic programming*. En: *Expert Systems with Applications* [en línea]. 2021 [consulta: 10 junio 2025]. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0957417421012677

ZHOU, Zhi-Hua. *Ensemble Methods: Foundations and Algorithms*. Boca Raton: Chapman and Hall/CRC, 2012. ISBN 978-1439830031.

10. Anexos

10.1. ANEXO A: Transcripción de Entrevista a Tomás Salomone

Juan Estarli: Hola Tomás ¿cómo estás?, nosotros somos Juan Estarli y Tomas Bussolini y somos estudiantes de Ingeniería en Informática en la Universidad Argentina de la Empresa. Actualmente nos encontramos desarrollando nuestro PFI y nos contactamos con vos para poder preguntarte acerca de temas que nos resultan útiles. Así que, desde ya, gracias por recibirnos.

Tomás Salomone: Gracias a ustedes, Juan y Tomas, un placer.

Juan Estarli: Bueno, ¿te parece que para comenzar nos comentes un poco acerca de tu rol relacionado al mundo del fútbol?

Tomás Salomone: Sí, cómo no. Mi nombre es Tomás Salomone, actualmente estoy trabajando en el Gobierno de la Ciudad de Buenos Aires, y soy un encargado directo en todo lo que es el proceso y organización de operativos para eventos futbolísticos.

Juan Estarli: Genial Tomás. La idea principal que tenemos es entender el proceso actual de planificación de un partido de fútbol, cómo se organiza y qué cosas se tienen en cuenta. Así que, en base a esto, ¿qué tipo de información les resulta más útil a la hora de planificar y hacer estimaciones de asistencia?

Tomás Salomone: Lo primero que miramos es el historial de partidos similares. Eso siempre nos sirve de referencia. Pero también influye mucho el contexto: si es un clásico, si hay rivalidad fuerte, si el resultado define algo importante. Todo eso cambia mucho el panorama.

Tomas Bussolini: Claro, tiene sentido.

Tomás Salomone: Además, hablamos bastante con los clubes. Ellos suelen tener un indicio del tipo de público que va a asistir, y eso suma. Y después están los recursos más técnicos: los anillos digitales y los medidores de tránsito, que nos dan información en tiempo real. Con todo eso armamos la planificación de los operativos de seguridad.

Tomas Bussolini: O sea que combinan datos históricos, contexto y lo que pueden llegar a medir en vivo.

Tomás Salomone: Exacto.

Juan Estarli: ¿y qué pasa cuando no tienen tantos datos?

Tomás Salomone: Bueno, en esos casos lo que hacemos es sobreestimar. Preferimos calcular de más y estar preparados para el escenario más exigente posible. Porque como hablamos de seguridad, tránsito y salud no nos podemos quedar cortos.

Juan Estarli: Claro, mejor pasarse que quedarse corto.

Tomás Salomone: Exactamente. Pero la consecuencia es que, en la mayoría de estos casos, terminamos asignando recursos que después no se usan, y eso eleva bastante lo que son los costos de los operativos.

Tomas Bussolini: Se entiende. Y te pregunto: ¿cómo verías una herramienta que use inteligencia artificial y distintas variables para anticipar con más precisión la cantidad de público?

Tomás Salomone: La realidad es que hoy ya usamos varias fuentes y nos funcionan bien, pero si pudiéramos unificarlas e integrarlas en una sola plataforma, sería un salto de calidad. Podríamos mejorar lo que tenemos y planificar de manera mucho más precisa.

Tomas Bussolini: Y supongo que eso no solo les serviría a ustedes, sino que también a otros sectores.

Tomás Salomone: Tal cual. Áreas como transporte, seguridad, salud o limpieza ya tienen sus propios sistemas. Pero si todos esos datos se pudieran combinar, las decisiones serían mucho mejores y podríamos anticiparnos más a los problemas.

Juan Estarli: Claro, sería como centralizar todo en una misma plataforma.

Tomás Salomone: Sí, exactamente.

Juan Estarli: Buen Tomás y para ir cerrando, ¿qué requisitos debería cumplir una herramienta así para que pueda implementarse en el ámbito público?

Tomás Salomone: Tiene que cumplir con los estándares del gobierno, tanto técnicos como legales. Son temas muy sensibles. Pero si la herramienta es confiable y se ajusta a estos requisitos, sería muy útil. Incluso te diría que estaría dispuesto a participar en una prueba piloto.

Juan Estarli: Buenísimo Tomás, mil gracias por tu tiempo y por compartir toda esta información que es muy valiosa para nuestro proyecto.

Tomás Salomone: Gracias a ustedes y estamos en contacto para cualquier otra duda o consulta que necesiten.

10.2. ANEXO B: Transcripción de Entrevista a Diego Filippi

Juan Estarli: Buenas Diego ¿cómo va?, somos Juan Estarli y Tomás Bussolini, estudiantes de Ingeniería en Informática en la UADE. Estamos trabajando en nuestro PFI y queríamos hacerte algunas preguntas que nos sirven para la investigación. Gracias por recibirnos.

Diego Filippi: Qué tal, Juan, Tomás, gracias por invitarme. Soy Diego Filippi, y soy uno de los responsables del ingreso y del registro del público que asiste a los partidos en Argentinos Juniors. Me parece buenísimo charlar así poder resolver las dudas que tengan.

Tomas Bussolini: Perfecto Diego, te comento que nuestro objetivo actualmente es recopilar información acerca de cómo funciona la organización y planificación de partidos en la Liga Profesional de Fútbol, para poder luego identificar aquellos puntos de dolor claves a tratar y mejorar. Así que para comenzar te preguntamos ¿cómo se toman hoy las decisiones sobre la organización logística antes de un partido y qué información usan?

Diego Filippi: La organización de cada evento deportivo se rige, en líneas generales, por un protocolo preestablecido, aunque este puede adaptarse en función del equipo visitante o de definiciones del comité de seguridad de la Ciudad de Buenos Aires. Previo al encuentro, el Comité de Seguridad de AAAJ, en conjunto con UTEDYC, que son los responsables del control en los accesos, revisa el protocolo específico a seguir para el evento, asegurando que todas las disposiciones estén alineadas con el contexto particular del partido. En esta instancia se planifican las necesidades puntuales en materia de accesos, considerando variables como el tipo de rival y la etapa del torneo en la que se encuentra el club local. Estas condiciones son las que pueden derivar en la necesidad de reforzar los accesos mediante la incorporación de tecnología, con el objetivo de mejorar la fluidez del ingreso del público. Sumado a esto, también está el líder de seguridad de AAAJ es quien se encarga de coordinar estas acciones con los responsables del operativo policial y con el Comité de Seguridad de la Liga, garantizando una planificación y ejecución integral del dispositivo de seguridad.

Tomas Bussolini: Bien, súper claro Diego. O sea, lo principal es un protocolo de base, ajustes que se realizan según rival y contexto, y coordinación con todos los actores.

Juan Estarli: Y cuando el partido promete mucha asistencia, ¿dónde se complican más las cosas?

Diego Filippi: Una de las principales dificultades se presenta en el control de accesos, específicamente en los molinetes. Ante el más mínimo inconveniente técnico en la lectura de carnets o códigos QR, el sistema se ve afectado, generando demoras que, en cuestión de minutos, pueden derivar en la congestión o incluso en el colapso de los accesos. Otra dificultad se ve en el control de aforo en las tribunas. Sucede que, en eventos de gran convocatoria, alcanzar el aforo máximo en determinados sectores puede generar complicaciones operativas, lo que en algunos casos obliga a cerrar temporalmente algunos accesos para garantizar la seguridad.

Juan Estarli: Claro, un cuello de botella en los molinetes y después complicaciones en las tribunas.

Tomas Bussolini: ¿Y qué pasa internamente si hay una diferencia grande entre lo esperado y lo que finalmente asiste?

Diego Filippi: La realidad es que no ocurre nada en particular porque los protocolos establecidos y la cantidad de personal de seguridad privada y policial están planificados para soportar el máximo aforo que permite el estadio, aunque eso signifique una sobreestimación.

Tomas Bussolini: O sea, prefieren estimar al máximo para no quedarse cortos, aunque cueste más.

Diego Filippi: Exacto.

Juan Estarli: Y Diego, en casos donde la concurrencia haya estado por debajo de la esperada, ¿a qué crees que se debió?

Diego Filippi: Por lo general, la baja asistencia está directamente relacionada con el rendimiento del equipo en el torneo, junto con el rival del encuentro.

Juan Estarli: Claro, se entiende. Y en la previa de cada partido, ¿qué es lo que más incertidumbre o preocupación les genera?

Diego Filippi: Organizar un evento de estas características siempre implica un cierto grado de incertidumbre, especialmente en lo relacionado con el factor humano. Pero en AAAJ ese riesgo se minimiza considerablemente por el fuerte espíritu familiar que tiene el club, donde todos los actores involucrados conocen su rol y colaboran con compromiso, lo que permite que cada operativo se desarrolle con orden y previsibilidad.

Tomas Bussolini: ¿Tenes algún ejemplo de un partido en el que la organización haya funcionado especialmente bien o mal?

Diego Filippi: Recuerdo con claridad varios al inicio de las temporadas en los que se presentaron inconvenientes en la lectura de los molinetes, lo que generó dificultades en algunos accesos para el ingreso del público. A raíz de estos episodios, lo que se hizo fue un análisis técnico del evento que permitió identificar y corregir la causa raíz del problema. Como resultado, se modificó el proceso de sincronización entre el sistema de socios y el sistema de control de accesos del estadio, lo que mejoró notablemente la estabilidad y confiabilidad del ingreso.

Juan Estarli: Perfecto, una buena mejora en los procesos.

Tomas Bussolini: ¿Cómo impacta la asistencia en decisiones de seguridad, personal o entradas durante el partido?

Diego Filippi: Es un aspecto clave, y por ello el estadio cuenta con un sistema de cámaras estratégicamente ubicadas. Desde el centro de monitoreo, donde trabajan en conjunto la Policía Federal, el equipo de seguridad de AAAJ y personal de UTEDYC, se supervisa en tiempo real cada uno de los accesos, así como el aforo de las tribunas. Este monitoreo online permite una rápida detección de incidentes. En caso de producirse algún inconveniente, tanto la Policía Federal como el equipo de seguridad del club y el personal de UTEDYC actúan de inmediato, coordinando acciones para restablecer el orden o mitigar problemas tecnológicos que impidan el acceso.

Juan Estarli: Y si la planificación no alcanza, ¿dónde se nota primero?

Diego Filippi: Si llegara a suceder, la sobreocupación se hace evidente a simple vista en el aforo de las tribunas.

Tomas Bussolini: Diego te llevo un poco a lo que es la tecnología, si existiera una solución que pueda llegar a predecir con cierto grado de confiabilidad la asistencia, ¿te serviría?

Diego Filippi: Sí, sería de gran utilidad. Una herramienta de este tipo permitiría anticipar escenarios con mayor precisión y planificar en consecuencia aspectos clave como accesos, seguridad, disposición del personal y servicios. También contar con predicciones confiables nos ayudaría a optimizar recursos y reducir imprevistos durante el evento.

Juan Estarli: Perfecto. Y para cerrar, ¿qué es lo más difícil de organizar un partido?

Diego Filippi: Sin dudas, lo más complejo es coordinar las necesidades de último minuto del comité de seguridad y de la Policía Federal.

Tomas Bussolini: Listo. Clarísimo. Diego, muchísimas gracias por el tiempo y tu disposición, nos ayuda un montón para nuestro trabajo.

Diego Filippi: No hay de que, gracias a ustedes. Si tienen alguna otra consulta sobre el tema, no duden en contactarme.

10.3. ANEXO C: Flujos de fondo

TABLA XV: flujo de fondos de la adopción pesimista.

Concepto	**Año 0**	**Año 1**	**Año 2**	**Año 3**
Inversión inicial	USD -5.990	USD 0	USD 0	USD 0
Costos	USD 0	USD 1.680	USD 2.052	USD 2.124
Ingresos	USD 0	USD 1.050	USD 2.550	USD 5.100
Flujo neto anual	**USD -5.990**	**USD -630**	**USD 498**	**USD 2.976**
Flujo acumulado	**USD -5.990**	**USD -6.620**	**USD -6.122**	**USD -3.146**

Fuente: Elaboración propia, 2025.

TABLA XVI: flujo de fondos de la adopción neutra.

Concepto	**Año 0**	**Año 1**	**Año 2**	**Año 3**
Inversión inicial	USD -5.990	USD 0	USD 0	USD 0
Costos	USD 0	USD 2.652	USD 2.772	USD 2.892
Ingresos	USD 0	USD 4.150	USD 8.300	USD 12.400
Flujo neto anual	**USD -5.990**	**USD 1.498**	**USD 5.528**	**USD 9.508**
Flujo acumulado	**USD -5.990**	**USD -4.492**	**USD 1.036**	**USD 10.544**

Fuente: Elaboración propia, 2025.

TABLA XVII: flujo de fondos de la adopción optimista.

Concepto	**Año 0**	**Año 1**	**Año 2**	**Año 3**
Inversión inicial	USD -5.990	USD 0	USD 0	USD 0
Costos	USD 0	USD 4.272	USD 4.512	USD 5.592
Ingresos	USD 0	USD 7.250	USD 15.500	USD 23.750
Flujo neto anual	**USD -5.990**	**USD 2.978**	**USD 10.988**	**USD 18.158**
Flujo acumulado	**USD -5.990**	**USD -3.012**	**USD 7.976**	**USD 26.134**

Fuente: Elaboración propia, 2025.

10.4. ANEXO D: Dataset de entrenamiento

Tabla XVIII. Dataset final para entrenamiento del modelo.

Variable	Descripción	Tipo de dato	Categoría
local	Equipo que actúa como local	Categórica	Identificación
visitante	Equipo que actúa como visitante	Categórica	Identificación
jornada_num	Número de jornada del torneo	Numérica	Temporal
provincia	Provincia donde se juega el partido	Categórica	Contextual
capacidad	Capacidad total del estadio	Numérica	Contextual
temperatura	Temperatura ambiente en °C	Numérica	Contextual
anio	Año del partido	Numérica	Temporal
mes	Mes del partido (1-12)	Numérica	Temporal
dia_semana	Día de la semana (0 = lunes / 6 = domingo)	Numérica	Temporal
fin_de_semana	Indicador si es fin de semana (0/1)	Binaria	Temporal
hora	Hora de inicio del partido	Numérica	Temporal
horario_cat	Categoría de horario (mañana/tarde/noche)	Categórica	Temporal

rival_grande	Indicador si el rival es equipo grande (0/1)	Binaria	Histórica
es_clasico	Indicador si es partido clásico (0/1)	Binaria	Histórica
asist_prom_local_exp	Promedio histórico de asistencia del local con decaimiento exponencial	Numérica	Histórica
ocupacion_prom_local_exp	Ocupación promedio histórica del estadio local con decaimiento exponencial	Numérica	Histórica
tendencia_occ_ult3	Tendencia de ocupación en los últimos 3 partidos del local	Numérica	Histórica
asist_std_local_ult5	Desviación estándar de asistencia en últimos 5 partidos del local	Numérica	Histórica
delta_ocupacion_reciente	Diferencia entre tendencia reciente y ocupación histórica	Numérica	Histórica
dias_descanso_local	Días transcurridos desde el último partido del local	Numérica	Temporal
clasico_en_finde	Indicador si es clásico jugado en fin de semana (0/1)	Binaria	Interacción
ocupacion_clasicos	Ocupación promedio histórica en partidos clásicos	Numérica	Histórica
ocupacion_no_clasicos	Ocupación promedio histórica en partidos no clásicos	Numérica	Histórica

ocupacion_historica_tipo	Ocupación histórica según tipo de partido (clásico o no)	Numérica	Histórica
temp_bin_muy_baja	Indicador de temperatura muy baja (< 10 °C) (0/1)	Binaria	Contextual
temp_en_tarde	Indicador de temperatura alta en horario de tarde (0/1)	Binaria	Interacción

Fuente: Elaboración propia, 2025.

10.5. ANEXO E: Cronograma

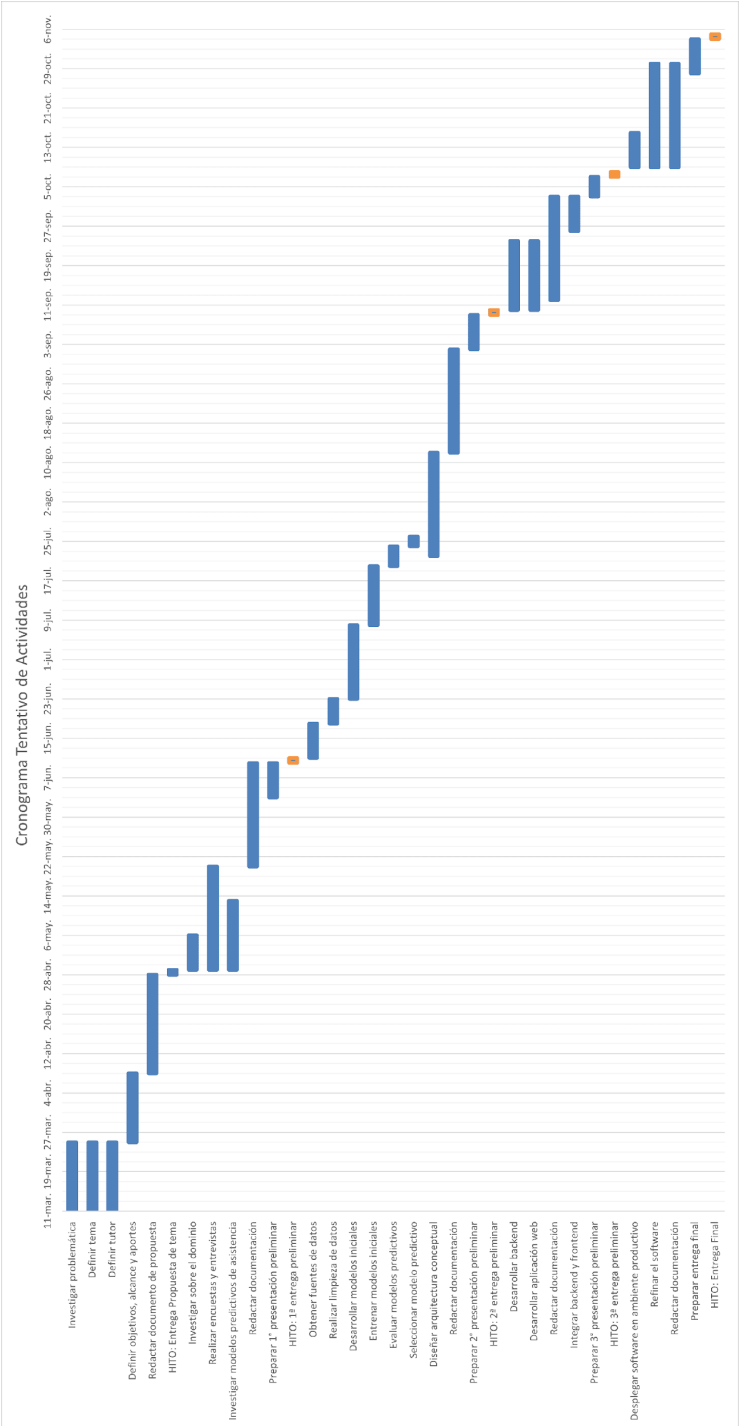


Figura 25. Cronograma tentativo de las actividades del proyecto. Fuente: Elaboración propia, 2025.