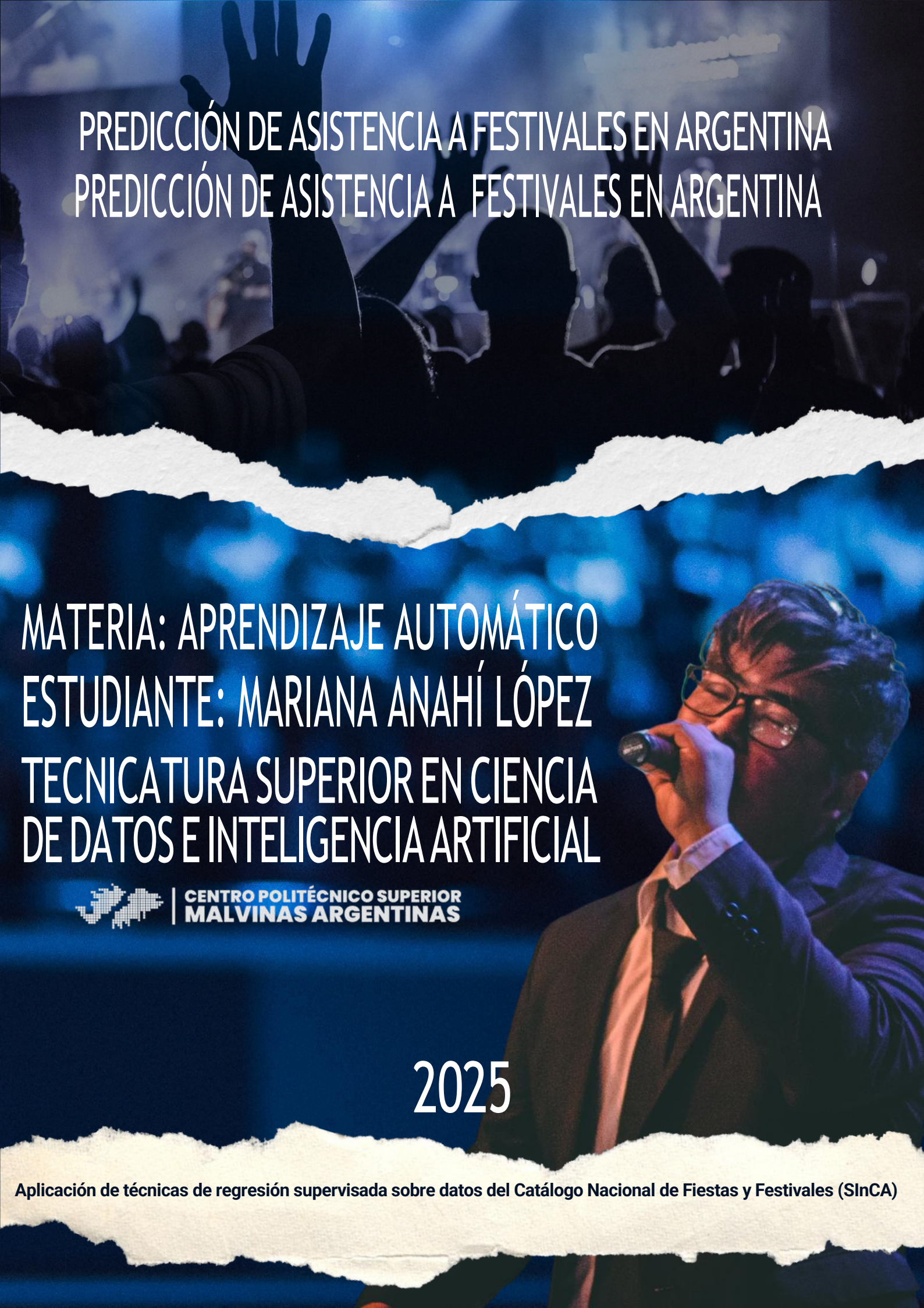
PREDICCIÓN DE ASISTENCIA A FESTIVALES EN ARGENTINA

PREDICCIÓN DE ASISTENCIA A FESTIVALES EN ARGENTINA



MATERIA: APRENDIZAJE AUTOMÁTICO

ESTUDIANTE: MARIANA ANAHÍ LÓPEZ

TECNICATURA SUPERIOR EN CIENCIA DE DATOS E INTELIGENCIA ARTIFICIAL

CENTRO POLITÉCNICO SUPERIOR
MALVINAS ARGENTINAS

2025

Aplicación de técnicas de regresión supervisada sobre datos del Catálogo Nacional de Fiestas y Festivales (SInCA)

Predicción de la cantidad de asistentes a festivales y fiestas populares en Argentina

Descripción del proyecto

El presente trabajo se desarrolla con el objetivo de analizar los festivales y fiestas populares realizadas en distintas provincias argentinas y construir un modelo predictivo que permita estimar la cantidad aproximada de asistentes a cada evento.

El estudio parte de un dataset público con información oficial y busca, además de aplicar técnicas de limpieza y transformación de datos, explorar relaciones entre variables sociales, culturales y geográficas que influyen en la convocatoria de estos eventos.

La elección de este dominio se basa en su relevancia económica y cultural: los festivales impactan de manera directa en el turismo, la ocupación hotelera, el comercio local y la difusión de las identidades regionales. Por ello, analizar sus patrones de asistencia puede aportar una herramienta valiosa para la planificación y la toma de decisiones tanto a nivel público como privado.

Descripción del dataset

El dataset original contiene información sobre 3255 festivales y fiestas populares registradas en todo el territorio argentino.

Cada fila representa un evento, con variables relacionadas a su localización, características temáticas, tipo de gestión, duración, fecha de realización y cantidad aproximada de asistentes.

Cantidad de instancias: 3255

Cantidad de variables originales: 27

Variables originales

id (o código interno del registro), provincia, cod_prov, departamento, cod_dep, localidad_paraje, nombre, tematica_principal, tematica_secundaria, tipo_de_gestion, tipo_de_gestion_privado, tipo_entrada, periodicidad, modalidad, duracion_(días), cantidad_aprox_de_asistentes, aniversario, año_primera_edicion, ediciones, ultimo_año_de_realizacion, semana_de_realizacion, mes_de_realizacion, nota, fuente, latitud, longitud, categoria.

Si bien todas aportaban información contextual, no todas eran útiles para el objetivo analítico del proyecto —estimar o modelar la cantidad aproximada de asistentes a cada evento. Para seleccionar las variables a analizar se aplicaron varios criterios:

• Relevancia predictiva: se conservaron variables que pudieran tener relación directa con la asistencia (por ejemplo, provincia o tipo de entrada).

• Completitud: se eliminaron columnas con más del 50% de valores nulos.

• Redundancia: se descartaron aquellas que duplicaban información o tenían variabilidad irrelevante.

Ejemplos eliminados: cod_prov, cod_dep, latitud, longitud, nota, fuente, tematica_secundaria, tipo_de_gestion_privado, modalidad, aniversario, semana_de_realizacion.

Estas variables se eliminaron en una primera etapa, ya que no aportaban valor predictivo y dificultaban el tratamiento posterior de los datos

Imputación de valores faltantes

Luego se decidio hacer una imputacion con los valores nulos para que el analisis del data set sea significativo, asi que se tomaron ciertas variables para trabajar:

- duracion_dias
- cantidad_aprox_de_asistentes

- mes_de_realizacion

Para no eliminar esos registros (dado que implicaria perder información valiosa), se aplicó una imputación basada en agrupamientos:

es decir, los valores faltantes se completaron calculando la media o la moda dentro de grupos con características similares.

Imputación de duracion_dias

- Variables de referencia utilizadas:

provincia, departamento, tematica_principal

- Lógica aplicada:

Se calculó la duración promedio de festivales que compartían la misma provincia y temática principal, dentro del mismo departamento cuando estaba disponible.

Si no existían coincidencias exactas, se utilizó el promedio general por provincia.

- Motivo:

La duración de un evento tiende a estar asociada con el tipo de celebración (por ejemplo, fiestas religiosas suelen durar varios días, mientras que ferias gastronómicas pueden ser de uno o dos).

Imputación de cantidad_aprox_de_asistentes

- Variables de referencia utilizadas:

provincia, departamento, tematica_principal

- Lógica aplicada:

Se completaron los valores faltantes con el promedio de asistentes correspondiente a festivales de la misma provincia y temática principal.

Este enfoque permitió respetar diferencias regionales (por ejemplo, mayor asistencia en Buenos Aires que en Tierra del Fuego) y por tipo de evento.

- Motivo:

La cantidad de público depende en gran medida del contexto geográfico y del tipo de actividad principal.

Imputación de mes_de_realizacion

- Variables de referencia utilizadas:

provincia, nombre (cuando el evento se repite anualmente)

- Lógica aplicada:

En casos donde el mismo evento se encontraba registrado en otros años o provincias, se recuperó el mes de realización observando coincidencias de nombre.

Cuando no fue posible, se aplicó la moda provincial (mes más frecuente para festivales en esa provincia).

- Motivo:

Los festivales suelen realizarse en fechas fijas o recurrentes cada año, por lo que la provincia y el nombre del evento son buenos predictores.

Variables finales seleccionadas para análisis y modelado

Finalmente las variables seleccionadas para el analisis y modelado fueron las siguientes:

Las variables conservadas fueron aquellas con valor explicativo potencial respecto de la asistencia a los eventos y que presentaban niveles aceptables de completitud:

- Provincia: define el contexto geográfico.

- Temática principal: indica el tipo de evento.

- Tipo de gestión: distingue entre público o privado.

- Tipo de entrada: influye directamente en la asistencia

- Mes de realización: refleja la estacionalidad.

- Duración (días): representa la magnitud del evento.

- Último año de realización: mide la vigencia del festival.

- Cantidad aprox. de asistentes: variable objetivo (target).

Proceso de Limpieza y procesamiento:

1. Normalización de nombres de columnas: eliminación de tildes y espacios.

2. Reemplazo de valores no disponibles (“s/d”, “Sin dato”, etc.) por NaN.

3. Eliminación de variables irrelevantes o redundantes.

4. Imputación de valores faltantes: duración y cantidad de asistentes según provincia, departamento y temática principal.

5. Conversión de tipos de datos a formato numérico

Transformaciones para modelado (One-Hot Encoding)

Para el entrenamiento de modelos de aprendizaje automático, se generó una versión completamente numérica del dataset mediante:

- Conversión binaria de tipo_entrada (0 = gratuita, 1 = paga).

- Codificación One-Hot de variables categóricas: provincia, tematica_principal, tipo_de_gestion, mes_de_realizacion.

- Eliminación de texto residual (nombre, departamento, tipo_entrada).

- Verificación de tipos de datos: todas las columnas fueron convertidas a enteros (int) para mantener homogeneidad.

Archivos generados:

Se generaron archivos distintos por cada etapa

Primero : DATA SET ORIGINAL. festivales_original.csv — descargado desde datos.gob.ar.

Segundo : DATA SET IMPUTADO. festivales_imputado.csv — Versión limpia e imputada, con datos corregidos.

Tercero : DATA SET CODIFICADO. festivales_codificado_modelo.csv — Versión numérica final lista para modelar.

Con este proceso se logró construir una base de datos coherente, estructurada y preparada para modelado predictivo, conservando el mayor volumen de información posible.

El objetivo final será identificar los factores que más influyen en la cantidad de asistentes a los festivales y construir un modelo que pueda estimar, de forma aproximada, la convocatoria esperada para eventos futuros.