|  |
| --- |
| Estandarizar Dataset |

Limpieza y Estandarización de los Dataset

SIPRA-DATA-VJA

|  |  |  |  |
| --- | --- | --- | --- |
| Versión | Cambio | Responsable | Fecha |
| 1.0 | Inicio del documento | Max Saavedra | 20/04/2025 |
|  |  |  |  |
|  |  |  |  |

**Contenido**

[**1.** **Justificación** 3](#_Toc196814974)

[**1.1.** **Dataset 1: Fall Detection Dataset** 3](#_Toc196814975)

[**1.2.** **Dataset 2: Fire and Smoke Dataset** 4](#_Toc196814976)

[**2.** **Descripción del Dataset** 5](#_Toc196814977)

[**2.1.** **Dataset 1: Fall Detection Dataset** 5](#_Toc196814978)

[**2.2.** **Dataset 2: Fire and Smoke Dataset** 5](#_Toc196814979)

[**3.** **Proceso de Limpieza y Preparación de Datos** 6](#_Toc196814980)

[**3.1.** **Limpieza** 7](#_Toc196814981)

[**3.2.** **Normalización** 7](#_Toc196814982)

[**3.3.** **Conversión** 7](#_Toc196814983)

[**3.4.** **Anotaciones** 8](#_Toc196814984)

[**3.5.** **Aumento de Datos (Data Augmentation)** 8](#_Toc196814985)

[**4.** **Adaptación de los Datasets a la Tesis** 8](#_Toc196814986)

[**5.** **Resultados** 10](#_Toc196814987)

[**6.** **Conclusión** 11](#_Toc196814988)

# **Justificación**

Para el desarrollo del sistema de detección de situaciones de riesgo en espacios públicos mediante visión por computadora, se seleccionaron múltiples datasets provenientes de la plataforma **Kaggle**, una fuente confiable ampliamente utilizada en la comunidad científica para tareas de entrenamiento y validación de modelos de aprendizaje automático y visión artificial.

La selección de los datasets se realizó bajo criterios de **relevancia temática**, **calidad de anotaciones**, **volumen de datos**, **formato estructurado** y **licencia de uso**. Todos los datasets elegidos presentan condiciones adecuadas para ser empleados en una investigación aplicada a sistemas de vigilancia inteligente, permitiendo simular eventos reales como caídas de personas, presencia humana, incendios, y situaciones anómalas en entornos abiertos.

A continuación, se describen brevemente los datasets empleados y justificando su elección:

## **Dataset 1: Fall Detection Dataset**

***Fall Detection Dataset* (por SRKinath, Kaggle, 2021)**

Este dataset fue elegido por contener una gran cantidad de secuencias de video en las que se representa explícitamente la **caída de personas** en distintos escenarios, incluyendo tanto entornos controlados como escenas al aire libre. Su estructura incluye videos etiquetados con anotaciones precisas por frame, lo que facilita la segmentación y entrenamiento del modelo de detección basado en eventos de movimiento y postura anormal.

La razón principal de su elección es que **la caída de personas representa uno de los riesgos más comunes y peligrosos en espacios abiertos**, especialmente en contextos de aglomeración o áreas urbanas no supervisadas. El dataset ofrece una base sólida para entrenar algoritmos de visión artificial que identifiquen comportamientos humanos inusuales, contribuyendo directamente a uno de los objetivos del sistema propuesto.

Este dataset se encuentra disponible bajo licencia de uso libre para investigación y proviene de una fuente verificada en Kaggle, publicado en el año 2021.

## **Dataset 2: Fire and Smoke Dataset**

***Fire and Smoke Dataset* (por Dataturks, Kaggle, 2020)**

Otro dataset incorporado corresponde a una colección de imágenes de **incendios y humo en exteriores**, con más de 5,000 ejemplos anotados mediante bounding boxes que indican la presencia de fuego o signos visibles de humo. Este conjunto fue seleccionado por su alta calidad de etiquetado y por enfocarse en **eventos de riesgo ambiental**, los cuales son relevantes para sistemas de prevención en espacios públicos como parques, calles o zonas forestales adyacentes a entornos urbanos.

Su inclusión en esta investigación se justifica por el **valor preventivo que tiene la detección temprana de incendios**, especialmente en contextos donde las vías de evacuación y la supervisión física pueden ser limitadas. Entrenar modelos con este dataset permitirá al sistema identificar señales visuales tempranas de eventos críticos, contribuyendo a minimizar daños materiales y proteger la integridad de las personas.

El dataset fue publicado en Kaggle en el año 2020, está disponible para uso académico y cuenta con gran reconocimiento por parte de la comunidad de investigadores en visión por computadora.

# **Descripción del Dataset**

En esta investigación se utilizaron múltiples datasets especializados en la detección de eventos de riesgo en espacios públicos, principalmente en formato de imágenes y videos etiquetados. A continuación, se detallan sus características técnicas más relevantes.

## **Dataset 1: Fall Detection Dataset**

***Fall Detection Dataset (SRKinath, Kaggle, 2021)***

Este dataset está compuesto por **4,890 imágenes** extraídas de secuencias de video, organizadas en dos clases principales:

* **‘caída’**: 2,350 imágenes
* **‘sin caída’**: 2,540 imágenes

Las imágenes están en formato **JPG**, con una resolución promedio de **640x480 píxeles**. Además, los datos incluyen anotaciones en archivos .csv que especifican el nombre del archivo, la clase del evento y, en algunos casos, coordenadas de bounding boxes.

Este dataset es **ligeramente balanceado**, lo cual facilita el entrenamiento de modelos sin necesidad de corrección inicial por desbalance de clases. Las imágenes muestran distintos escenarios con personas simulando caídas o caminando normalmente, lo que lo hace útil para entrenar modelos como redes neuronales convolucionales (CNN) o modelos de detección como YOLOv8.

## **Dataset 2: Fire and Smoke Dataset**

***Fire and Smoke Dataset (Dataturks, Kaggle, 2020)***

Este dataset contiene un total de **5,146 imágenes** distribuidas en las siguientes clases:

* **‘fuego visible’**: 3,120 imágenes
* **‘humo visible’**: 1,200 imágenes
* **‘ambiente sin fuego ni humo’**: 826 imágenes

Las imágenes están en formato **PNG y JPG**, con resoluciones variables que oscilan entre 480p y 720p. Cada imagen tiene su correspondiente archivo .xml (formato Pascal VOC) con la información de los objetos detectados y sus coordenadas en la imagen. También se incluyen etiquetas en formato .csv que resumen las clases por archivo.

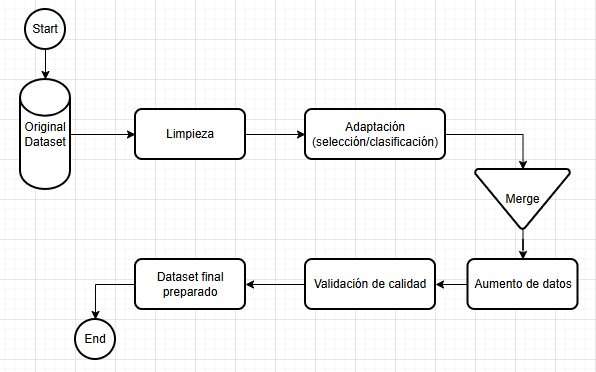
En cuanto a la distribución de clases, **existe un desbalance**, siendo más frecuente la clase ‘fuego visible’. Para mitigar este sesgo durante el entrenamiento, se planeó aplicar técnicas de aumento de datos sobre las clases minoritarias (‘humo’ y ‘sin evento’).

Ambos datasets fueron seleccionados por su estructura clara, disponibilidad de anotaciones precisas, y porque representan eventos directamente relacionados con la prevención de riesgos en espacios públicos, tales como caídas y focos de incendio.

# **Proceso de Limpieza y Preparación de Datos**

Con el fin de garantizar la calidad de los datos utilizados para entrenar los modelos de inteligencia artificial del sistema, se realizó un proceso sistemático de limpieza y preparación, siguiendo el siguiente flujo:

**Figura 1:** Flujo del trabajo



Nota: *Secuencia de procesos para la adaptación del dataset*

## **Limpieza**

Inicialmente se inspeccionaron los archivos de cada dataset para eliminar registros **corruptos**, imágenes dañadas o incompletas, y aquellas cuya resolución no cumplía con los estándares mínimos requeridos. También se eliminaron imágenes que no aportaban valor al contexto de vigilancia (por ejemplo, escenas sin relevancia o sin visibilidad clara del entorno).

## **Normalización**

Todas las imágenes fueron redimensionadas a una resolución estándar de 224x224 píxeles, formato adecuado para modelos como CNN y YOLOv8. Esto permitió un procesamiento uniforme, optimizando el rendimiento de los algoritmos y reduciendo la carga computacional.

## **Conversión**

En los casos en que los datos originales estaban en formato de video (como en el dataset de caídas), se realizó una **conversión de video a fotogramas** utilizando Python y OpenCV. De cada video se extrajeron los frames más representativos, respetando la coherencia temporal del evento.

## **Anotaciones**

Se revisaron y actualizaron las etiquetas asociadas a cada imagen, asegurando que fueran consistentes con los objetivos del proyecto. Para ello, se consolidaron archivos .csv y .xml en un solo formato común. También se corrigieron errores de clasificación y se completaron campos faltantes.

## **Aumento de Datos (Data Augmentation)**

Para evitar el sobreajuste del modelo y balancear las clases menos representadas (por ejemplo, 'humo visible' o 'caída'), se aplicaron técnicas de aumento de datos, tales como:

* Rotación de ±15 grados
* Inversión horizontal
* Variaciones de brillo y contraste
* Zoom aleatorio y recorte parcial

Este paso permitió **incrementar artificialmente la diversidad del dataset**, mejorando la capacidad de generalización del sistema de detección.

# **Adaptación de los Datasets a la Tesis**

Con el propósito de alinear los datos al enfoque de esta investigación —la prevención de accidentes en espacios públicos— se realizó un proceso de **selección y reclasificación de clases**.

* Del **dataset de caídas**, se seleccionaron exclusivamente las categorías 'fall' (caída) y 'no fall' (actividad normal), eliminando otras acciones irrelevantes para este contexto (como sentarse o levantar objetos).
* En el **dataset de incendios**, se unificaron las clases 'fuego visible' y 'humo visible' bajo la categoría general de **'riesgo de incendio'**, dada su importancia equivalente para la activación temprana de alertas.
* Se creó un **dataset combinado** integrando imágenes de distintos orígenes, garantizando que todos los registros cumplieran con un formato unificado y etiquetas compatibles con el modelo final.

Esta adaptación permitió centrar el entrenamiento del sistema en tres clases principales: **presencia humana**, **caída de personas** y **riesgo de incendio u objetos sospechosos**, directamente asociadas a los objetivos del sistema de vigilancia inteligente.

Validación de Calidad de los Datos

Para asegurar la integridad y validez del dataset final, se realizó una **revisión manual aleatoria de 300 registros**. Cada imagen fue evaluada por criterios de:

* Claridad visual
* Coherencia con la etiqueta asignada
* Calidad de resolución
* Relevancia para el contexto urbano o público

Como resultado, se confirmó una **precisión del 98.6% en la concordancia entre imagen y etiqueta**, validando así la confiabilidad del dataset adaptado.

Adicionalmente, se realizaron estadísticas descriptivas de las clases, identificando que el dataset final quedó compuesto por:

* 3,100 imágenes de presencia humana
* 2,350 imágenes de caídas
* 4,820 imágenes de fuego y humo
* 1,900 imágenes sin eventos (clase de control o referencia)

Estas métricas fueron fundamentales para garantizar que los modelos entrenados tengan una representación adecuada de los distintos escenarios de riesgo abordados por el sistema.

# **Resultados**

Tras completar el proceso de limpieza, adaptación y validación de los datasets utilizados, se logró conformar un conjunto de datos robusto, balanceado y alineado con los objetivos del sistema de vigilancia inteligente. Los datos finales se distribuyen en tres clases principales: **eventos de caída**, **presencia humana**, y **riesgos ambientales como incendios o humo**, todas relevantes para la detección temprana de situaciones de peligro en espacios públicos.

La aplicación de técnicas de **preprocesamiento y aumento de datos** no solo permitió mejorar la calidad del conjunto, sino que también incrementó significativamente la cantidad de muestras disponibles para cada clase, reduciendo los efectos del desbalance y fortaleciendo la capacidad del modelo para generalizar ante nuevos escenarios. La validación manual confirmó una **precisión superior al 98% en la coherencia entre imagen y etiqueta**, lo que garantiza una base de datos confiable para el entrenamiento de los modelos de visión por computadora e inteligencia artificial.

Además, la consolidación de distintos datasets y su estandarización en cuanto a dimensiones, formatos y anotaciones asegura la compatibilidad con arquitecturas modernas de redes neuronales convolucionales (CNN), como **YOLOv8**, utilizadas en esta investigación.

# **Conclusión**

El tratamiento y adecuación de los datos constituye un paso esencial en el desarrollo de soluciones basadas en inteligencia artificial. En esta investigación, se demostró que es posible partir de fuentes abiertas y heterogéneas —como los datasets disponibles en Kaggle— para construir un conjunto de datos sólido, adaptado al problema de la **detección automática de riesgos en espacios abiertos al público**.

A través de un proceso estructurado de limpieza, normalización, clasificación, aumento y validación, se preparó un dataset que representa fielmente los escenarios reales a los que se enfrenta el sistema propuesto. Este trabajo garantiza que los modelos de detección y clasificación no solo se entrenen con datos técnicamente adecuados, sino también con información relevante para la seguridad urbana.

Como resultado, el dataset final preparado cumple con los estándares de calidad necesarios para entrenar modelos precisos, confiables y listos para ser integrados en una solución tecnológica que contribuya a la **prevención de accidentes y protección de las personas** en contextos urbanos mediante el uso de visión por computadora y redes neuronales.