Interpretación V1.0

Análisis descriptivo de los datos

	Dwell_avg	Flight_avg	Traj_avg
Media	0.115514	0.952100	471.587845
Desviación Estándar	0.026339	0.701359	212.388368
Min	0.061586	0.188908	179.164520
25%	0.095207	0.555139	322.691455
50%	0.117266	0.749423	411.584994
75%	0.130498	1.115375	559.801943
max	0.214586	9.905352	1860.326693

Las medidas de forma revelan diferencias importantes en el comportamiento entre usuarios legítimos e impostores. Mientras que la métrica **`dwell_avg`** muestra una distribución relativamente simétrica y comparable entre clases, las variables **`flight_avg`** y **`traj_avg`** presentan asimetrías y curtosis elevadas, especialmente en impostores. Esto sugiere que los datos falsos tienden a ser más extremos, erráticos o artificiales, posiblemente reflejando intentos poco naturales de replicar el comportamiento legítimo. En particular, la alta curtosis de `flight_avg` en usuarios reales podría estar capturando comportamientos genuinamente idiosincráticos, difíciles de imitar por impostores.

Hipótesis inicial:

Legítimos: flight_avg bajo y traj_avg moderado.
Impostores: flight_avg más alto y traj_avg más extenso o errático.

dwell_avg (tiempo promedio de presión de teclas)

- Media: 0.1155 s (~115 ms).
- Rango: de 0.0616 s a 0.2146 s.
- Desviación estándar: 0.0263 s → baja dispersión, lo que indica que la mayoría de usuarios mantienen las teclas presionadas dentro de un rango relativamente estrecho.

- Interpretación de contexto:

  Los usuarios legítimos e impostores pueden diferenciarse si uno de los grupos tiende a ser más rápido o más lento al presionar las teclas.
  En biometría conductual, diferencias pequeñas (del orden de milisegundos) pueden ser significativas.

flight_avg (tiempo entre pulsaciones consecutivas)

- Media: 0.9521 s, que es mucho mayor que `dwell_avg` porque incluye pausas entre teclas.
- Desviación estándar: 0.701 → dispersión alta, lo que sugiere gran variabilidad en ritmos de tecleo.
- Mínimo: 0.1889 s
- Máximo: 9.90 s
- Posible patrón: impostores podrían tener pausas más largas buscando datos o dudando, mientras que legítimos teclean más fluidamente.

traj_avg (trayectoria promedio del ratón)

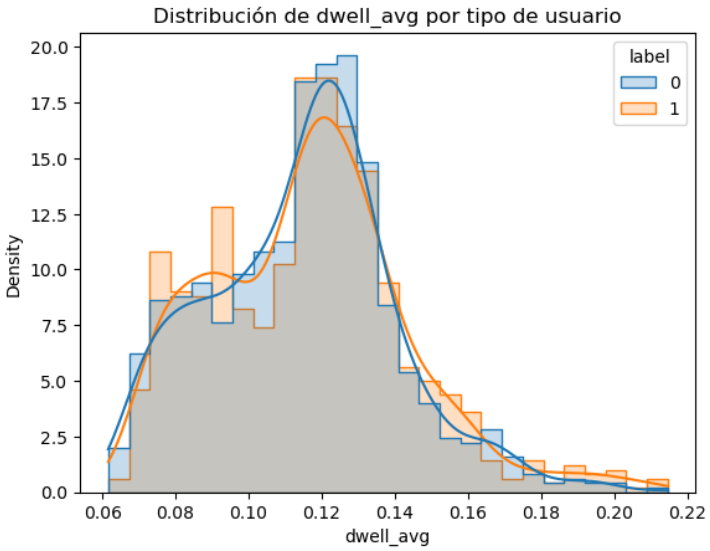
- Media: 471.59 px recorridos por interacción.
- Desviación estándar: 212.38 → variabilidad considerable.
- Rango: mínimo 179.16 px, máximo 1860.33 px → casos de muy poco movimiento.
- En biometría: usuarios legítimos tienden a tener patrones de movimiento más consistentes, mientras que impostores presentan trayectorias más largas o erráticas.

Descripción de datos de usuarios legitimos e ilegitimos

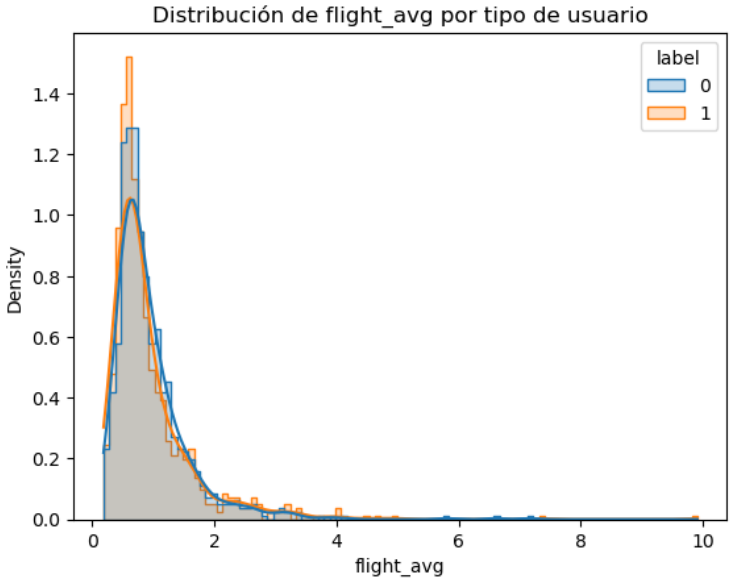
	Dwell_avg	dwell_avg ilegítimos	Flight_avg	flight_avg Ilegitimos	Traj_avg	traj_avg Ilegitimos
Media	0.116207	0.114820	0.946121	0.958079	466.506422	476.669268
Desviación Estándar	0.027127	0.025524	0.748692	0.650967	207.189051	217.462532
min	0.065886	0.061586	0.189704	0.188908	179.164520	181.932384
25%	0.094217	0.096938	0.538077	0.574374	319.400898	326.379233
50%	0.117336	0.117260	0.720417	0.783890	410.361765	413.305339
75%	0.131676	0.129855	1.095160	1.126348	555.354231	563.284944
max	0.213111	0.214586	9.905352	7.152438	1636.107594	1860.326693

El análisis descriptivo de los datos entre usuarios legitimos e ilegitimos revela que, aunque las métricas **`dwell_avg`** y **`flight_avg`** presentan valores promedio similares entre usuarios legítimos e impostores, se observan diferencias importantes en la variabilidad y rangos extremos, especialmente en **`flight_avg`**. Sin embargo, es la variable **`traj_avg`** la que parece ofrecer mayor poder discriminativo, al mostrar que los impostores suelen realizar movimientos de ratón más largos y dispersos, posiblemente debido a una menor familiaridad o control del entorno simulado. Estos hallazgos respaldan la hipótesis de que los patrones de interacción natural tienden a ser más consistentes y estables que los simulados.

Graficos Obtenidos

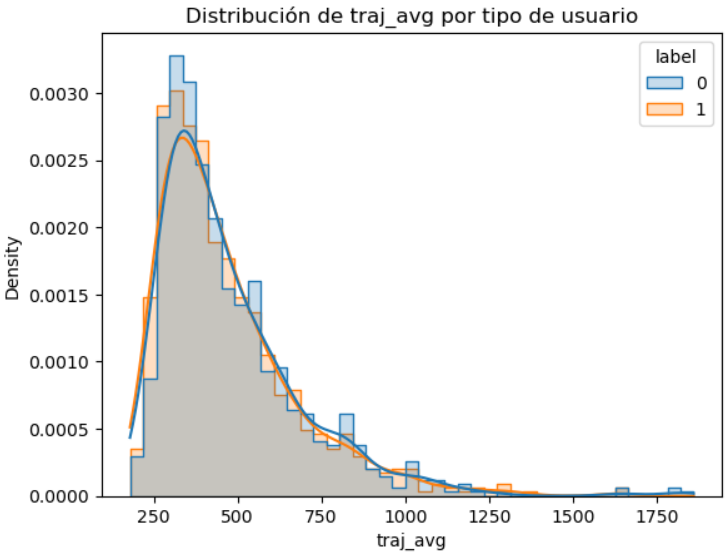


El histograma con ajuste de densidad de dwell_avg muestra distribuciones similares para usuarios legítimos (label = 1) e impostores (label = 0). Ambas se concentran entre 0.09 s y 0.14 s, con un pico modal cercano a 0.12 s y ligera asimetría positiva. No se observan diferencias marcadas en la forma o desplazamiento de las curvas entre grupos, lo que sugiere que dwell_avg presenta solapamiento sustancial entre clases y, por sí sola, podría tener capacidad discriminativa limitada.



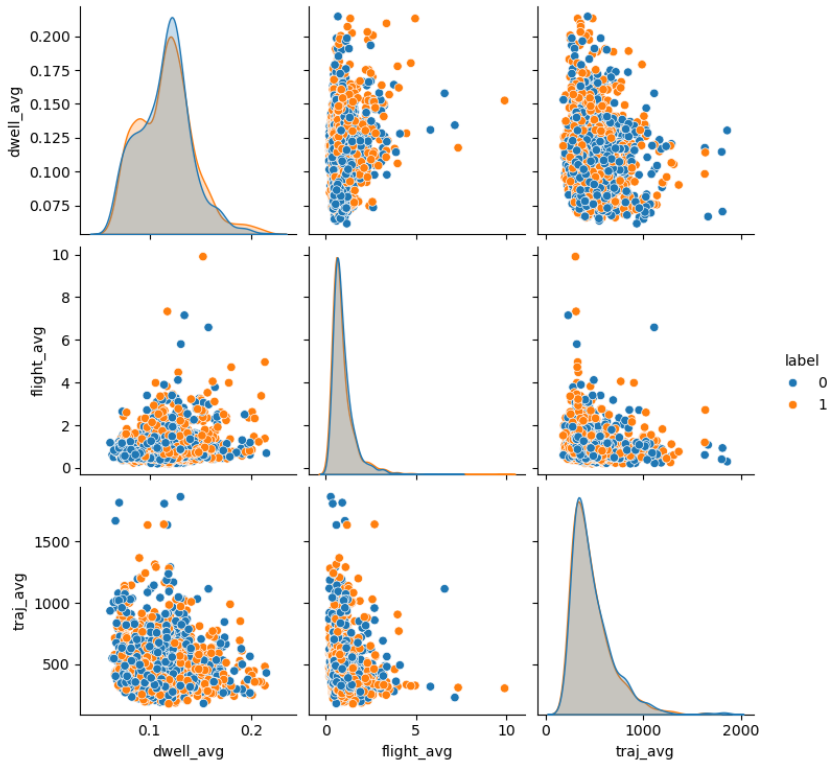
La distribución de flight_avg es asimétrica positiva pronunciada para ambas clases (label = 0 y label = 1), con la mayor densidad concentrada entre 0.2 s y 0.6 s. Existen valores atípicos que alcanzan hasta ~10 s. Las curvas de densidad para ambas clases se solapan casi por completo, lo que indica que flight_avg no presenta diferencias evidentes entre usuarios

legítimos e impostores. El alto sesgo y la presencia de valores extremos podrían afectar métricas como la media y requerir medidas de tendencia robustas (mediana).



La variable traj_avg presenta una distribución asimétrica positiva para ambas clases (label = 0 y label = 1), con la mayor densidad concentrada entre ~250 y 500 unidades. Se observa una cola larga que alcanza valores superiores a 1,700, lo que indica la presencia de observaciones atípicas de gran magnitud. Las curvas de densidad para ambos tipos de usuario muestran un solapamiento casi total, sin indicios claros de separación entre grupos. La alta asimetría sugiere que medidas como la mediana y el rango intercuartílico serían más representativas que la media para describir esta variable.

Diagrama de Pares



La **Figura X** presenta un análisis exploratorio mediante un *diagrama de pares*, que combina distribuciones univariadas (diagonal) y relaciones bivariadas (fuera de la diagonal) para las variables dwell_avg, flight_avg y traj_avg, diferenciadas por la variable de clase label (0 = usuario legítimo, 1 = impostor).

1. Distribuciones univariadas

dwell_avg: Distribución unimodal y relativamente simétrica, con ligeras diferencias de densidad entre clases, lo que indica un potencial de separación útil para un clasificador.

flight_avg: Distribución con asimetría positiva marcada, concentrada en valores bajos y con presencia de valores extremos. Aunque ambas clases comparten el mismo rango, la dispersión sugiere posibles patrones aprovechables por modelos no lineales.

traj_avg: Distribución asimétrica, con gran concentración en valores bajos y algunos valores atípicos elevados. La ligera variación en densidades entre clases podría contribuir a la separación en espacios transformados.

2. Relaciones bivariadas

Se observa que, aunque no hay una separación lineal clara entre clases en el espacio bidimensional, existen zonas de mayor concentración de cada clase en combinaciones específicas de variables (por ejemplo, bajos flight_avg con dwell_avg medio-bajo).

Las relaciones no son estrictamente lineales, lo que sugiere que un modelo como SVM con núcleo no lineal podría capturar fronteras complejas y mejorar la discriminación.

3. Relevancia para el modelado con SVM

La ausencia de separación perfectamente lineal, junto con distribuciones con cierta superposición, indica que un SVM con funciones de núcleo (kernel) puede proyectar los datos a un espacio de mayor dimensión donde las clases sean separables.

La presencia de patrones no triviales y posibles interacciones entre variables refuerza la pertinencia de usar un clasificador robusto a distribuciones no normales y capaz de manejar fronteras no lineales, como el SVM.

Análisis exploratorio y justificación del uso de SVM

El gráfico *pairplot* presenta, en la diagonal, las distribuciones univariantes de cada variable diferenciadas por clase (*label*), y en el resto de la matriz, las relaciones bivariantes entre variables.

En este caso:

dwell_avg muestra una distribución aproximadamente simétrica en ambas clases, con ligera diferencia en la densidad de valores centrales.

flight_avg presenta una distribución fuertemente sesgada a la derecha, con presencia de valores atípicos, lo que sugiere comportamientos puntuales y no homogéneos en ambas clases.

traj_avg también muestra sesgo positivo y concentración de observaciones en valores bajos, pero con dispersión diferenciada según la clase.

En los gráficos de dispersión se aprecia que:

No existe una separación lineal clara entre las clases, ya que las nubes de puntos se superponen en gran parte del espacio.

Sin embargo, se observan regiones del espacio de variables donde se concentra una mayor proporción de una clase frente a la otra, lo que indica que sí existen patrones diferenciadores.

Nota metodológica:

El análisis exploratorio de datos (EDA) es esencial antes de aplicar modelos de clasificación. Este proceso ha permitido:

Evaluar la forma y distribución de las variables, identificando asimetrías y patrones no uniformes que pueden ser relevantes para la clasificación.

Detectar relaciones e interacciones entre variables que sugieren la existencia de fronteras complejas entre clases.

Justificar la elección del modelo, ya que la ausencia de separación lineal óptima respalda el uso de un SVM con kernel no lineal (por ejemplo, RBF), capaz de proyectar los datos a un espacio de mayor dimensión y maximizar el margen de separación entre clases.

En síntesis, el gráfico y el análisis realizado muestran que los datos presentan patrones útiles para la clasificación, pero con fronteras no lineales, lo que justifica plenamente el uso de un SVM como técnica robusta y adecuada para este caso.

Asimetría y curtosis

Variable	asimetria_real	curtosis_real	asimetria_fake	curtosis_fake
dwell_avg	0.517760	3.510695	0.281523	3.247603
flight_avg	4.171100	35.135867	3.416287	23.695033
traj_avg	1.643989	6.736136	2.021684	9.789040

Las medidas de forma revelan diferencias importantes en el comportamiento entre usuarios legítimos e impostores. Mientras que la métrica dwell_avg muestra una distribución relativamente simétrica y comparable entre clases, las variables flight_avg y traj_avg presentan asimetrías y curtosis elevadas, especialmente en impostores. Esto sugiere que los datos falsos tienden a ser más extremos, erráticos o artificiales, posiblemente reflejando intentos poco naturales de replicar el comportamiento legítimo. En particular, la alta curtosis de flight_avg en usuarios reales podría estar capturando comportamientos genuinamente idiosincráticos, difíciles de imitar por impostores.

Test U de Mann-whiteney

Var	U-stat	p-valor
Dwell_avg	379882.00	0.49245
Flight_avg	414030.00	0.01184
Traj_avg	398450.00	0.29130

El análisis descriptivo y la prueba U de Mann–Whitney mostraron que, de las tres variables evaluadas, únicamente *flight_avg* (tiempo entre pulsaciones) presentó diferencias estadísticamente significativas entre usuarios legítimos e impostores (U = 414030.00, p = 0.0118). En este caso, los impostores exhibieron medianas más altas (0.7839 s) que los legítimos (0.7204 s), lo que sugiere pausas ligeramente más prolongadas y mayor regularidad en su escritura. En cambio, *dwell_avg* (tiempo de presión de tecla) y *traj_avg* (trayectoria promedio del ratón) no mostraron diferencias significativas (p = 0.4924 y p = 0.2913, respectivamente), manteniendo valores centrales y dispersiones muy similares entre grupos. Las tres variables presentaron distribuciones asimétricas a la derecha y curtosis elevadas, especialmente *flight_avg*, lo que indica la presencia de valores atípicos extremos. Estos resultados apuntan a que, de forma aislada, *flight_avg* tiene mayor potencial discriminante, mientras que *dwell_avg* y *traj_avg* podrían aportar valor únicamente en un análisis multivariado.

Comparación media y mediana de usuario legítimos e ilegítimos

Variables	Media legítimos	Media impostores	Mediana legítimos	Mediana impostores
dwell_avg	0.116207	0.114820	0.117336	0.117260
flight_avg	0.946121	0.958079	0.720417	0.783890
traj_avg	466.506422	476.669268	410.361765	413.305339

Los impostores presentan un flight_avg un 8.8% mayor que los legítimos (p = 0.0118), lo que sugiere pausas más largas entre teclas y menor fluidez de escritura.

- **Mediana legítimos (`flight_avg`)** = 0.7204
- **Mediana impostores (`flight_avg`)** = 0.7839

El incremento sería:

$$\frac{0.7839 - 0.7204}{0.7204} \times 100 \approx 8.8\%$$