

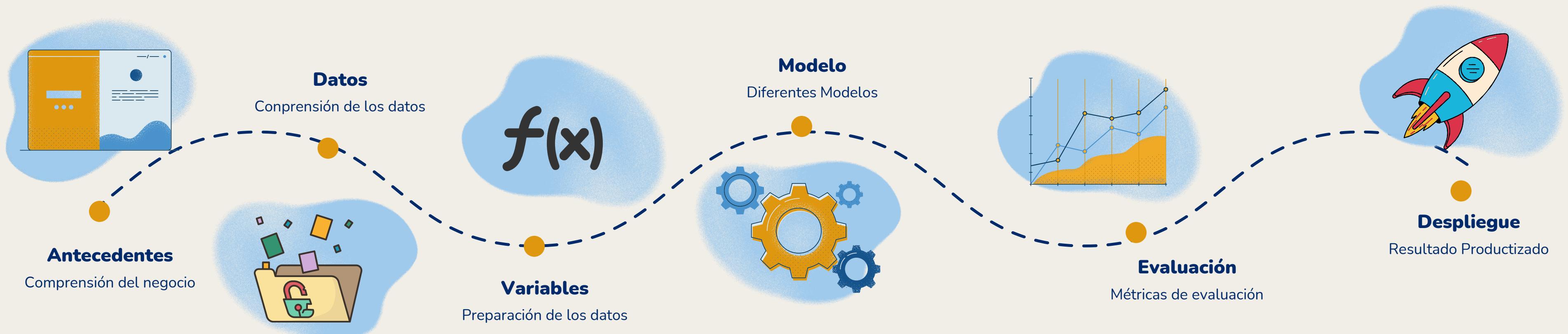
Detección de Anomalías

Presentado por:

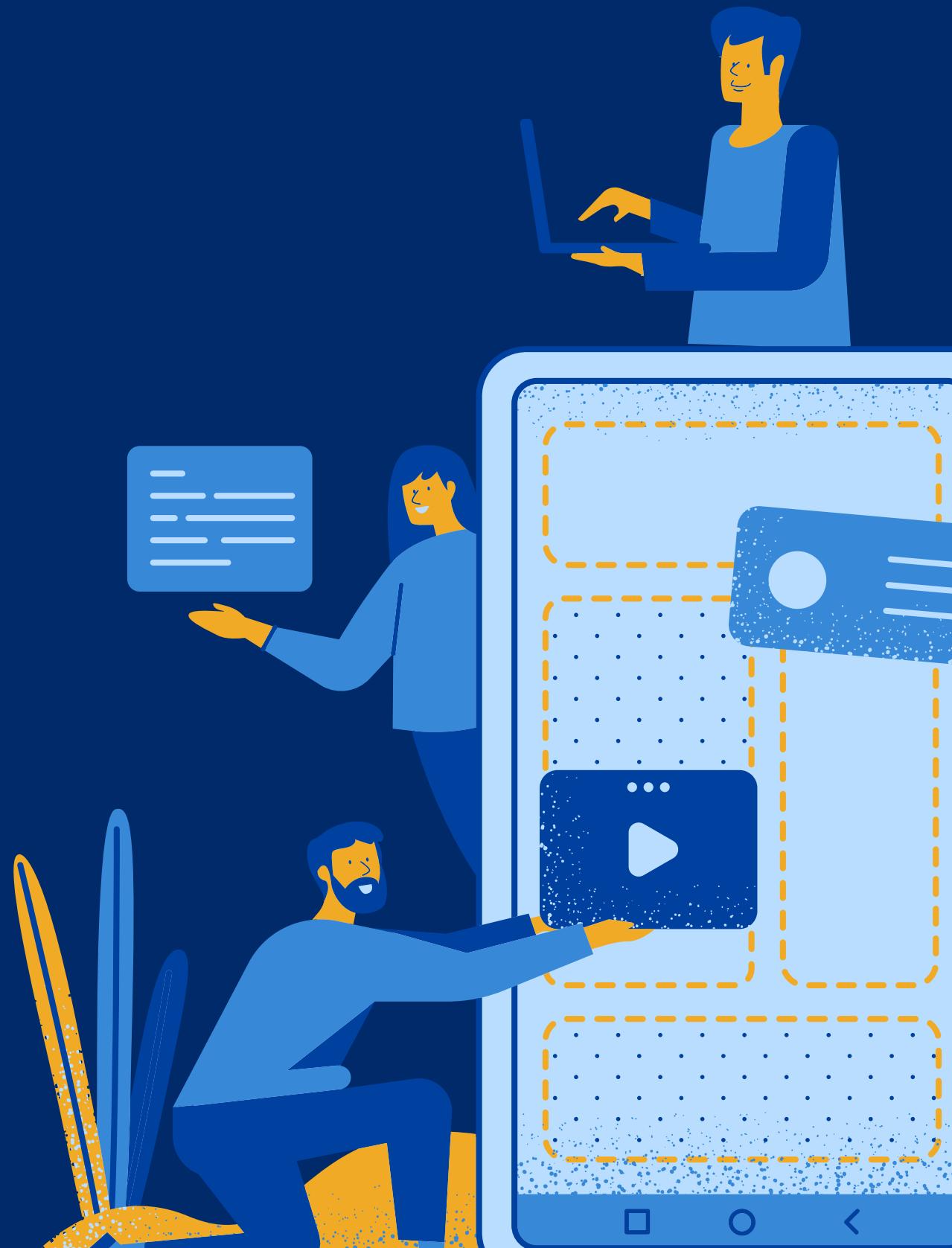
Marisol Correa Henao
Data Scientist



CRISP-DM vs KDD

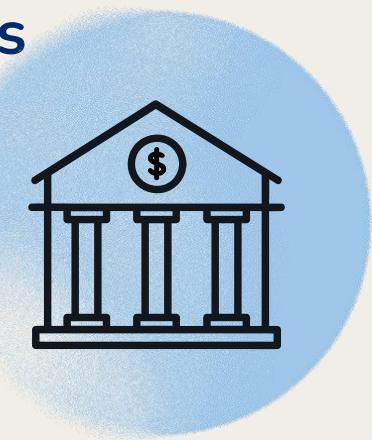


Antecedentes



Instituciones Financieras

Detección de patrones inusuales transaccionales



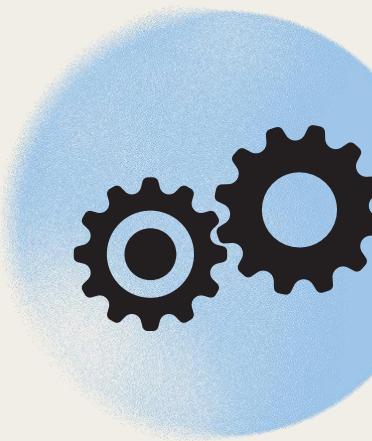
Redes Sociales

Detección de actividades sospechosas o no auténticas



Industria Manufacturera

Monitoreo del rendimiento de las maquinas



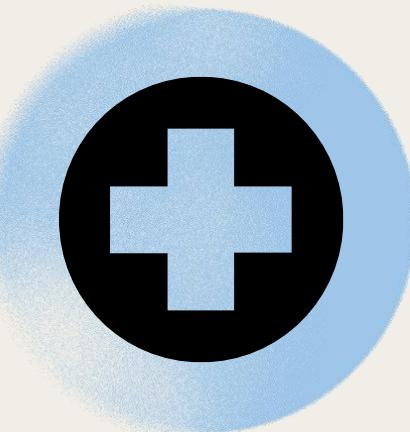
Sistemas de Seguridad

Identificar comportamientos sospechosos o intrusos en redes y sistemas informáticos



Salud

Diagnóstico temprano de enfermedades



Otros

Construcción, energía, entre otros...



Anomalías

Tipos Modelos

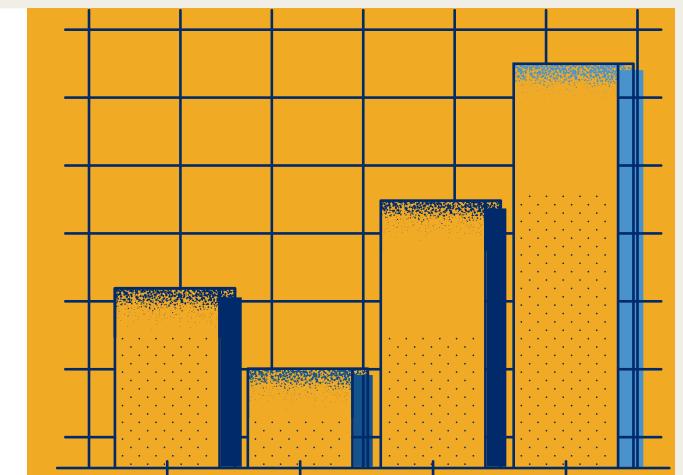
Supervisado

Aprendo



No Supervisado

Identifico



No Supervisado

Clustering

Agrupa basado en características similares.



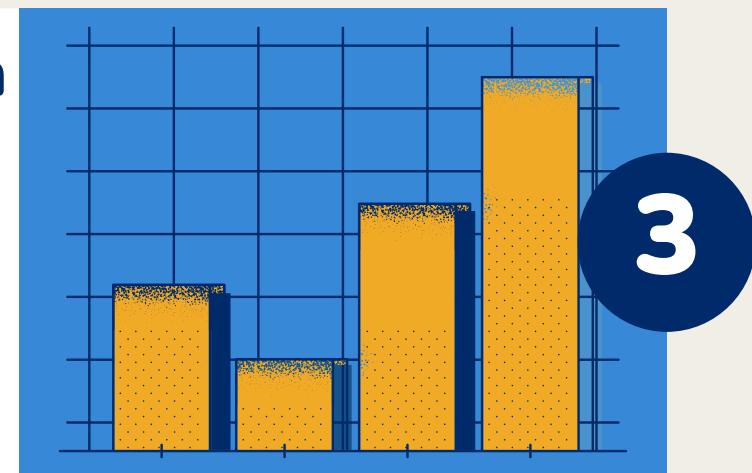
Association Rules

Descubre patrones y relaciones interesantes entre diferentes eventos o acciones.



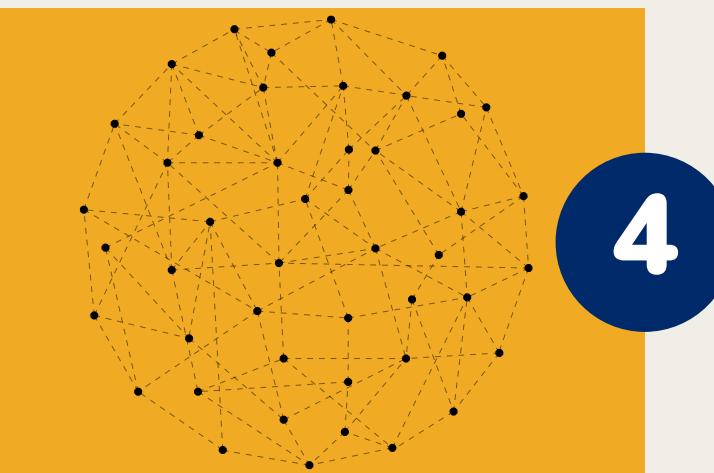
Anomaly Detection

Aprende patrones normales y luego identifica las desviaciones significativas.



Redes Neuronales Autoencoder

Reconstruye el comportamiento normal de los usuarios o entidades.



Algoritmos

Clustering

K-means Clustering



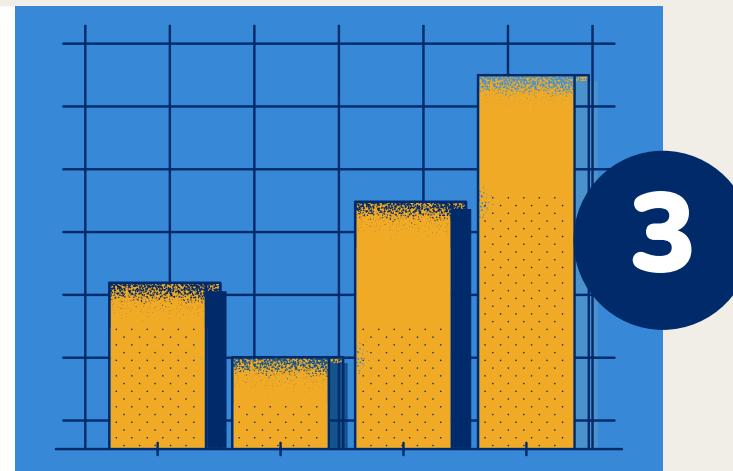
Association Rules

Apriori

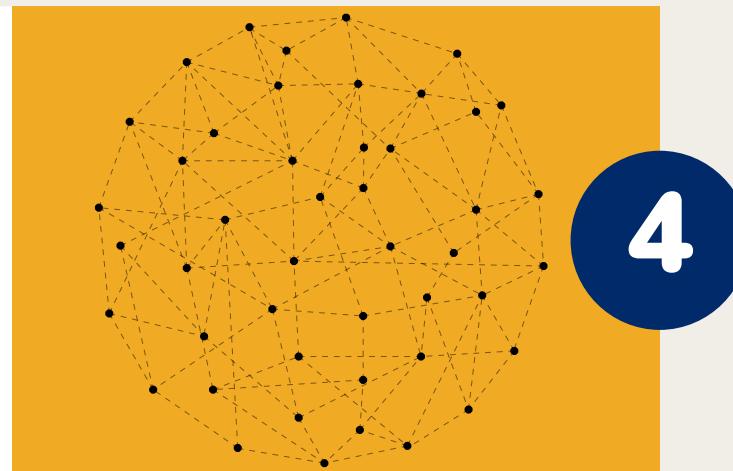


Anomaly Detection

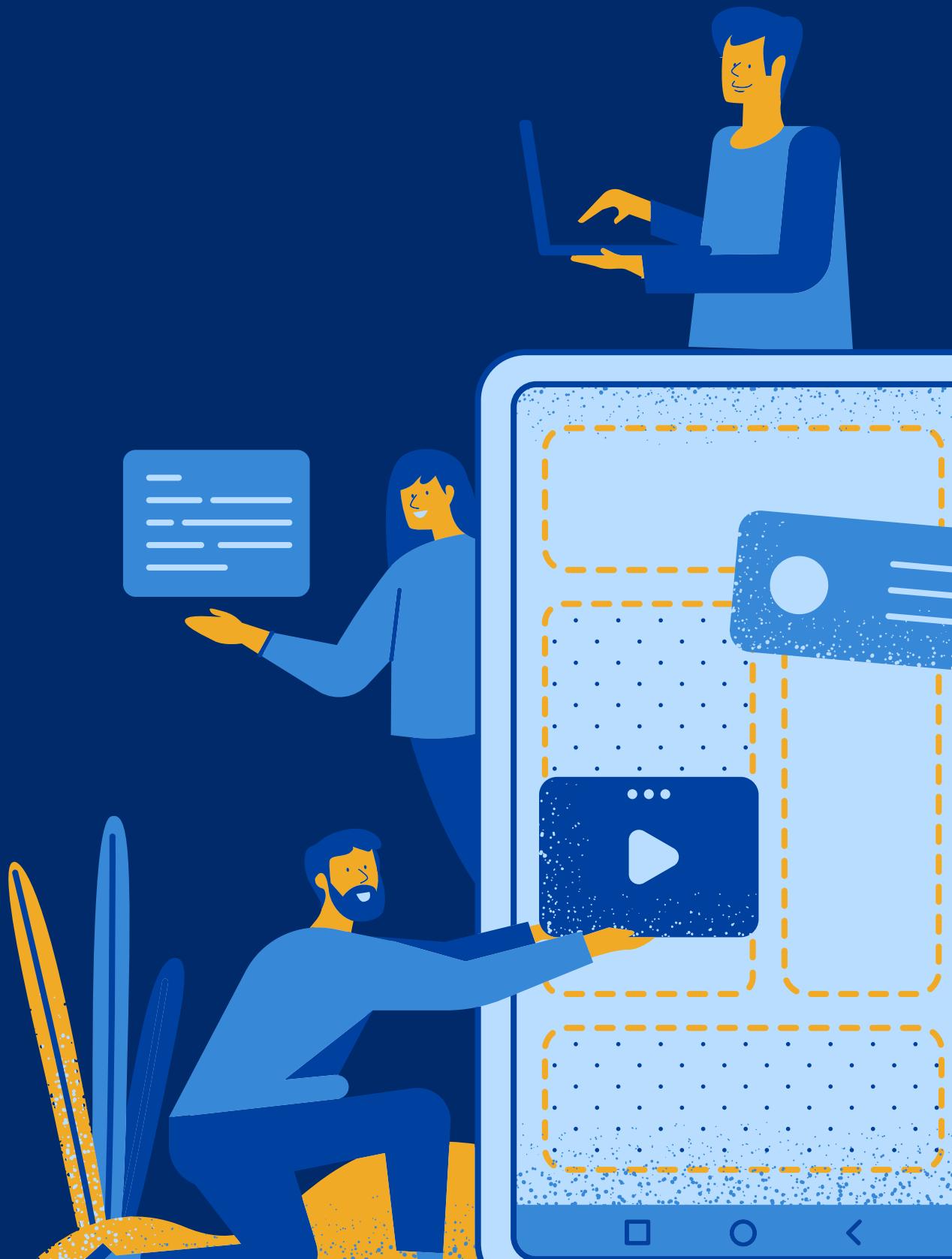
Isolation Forest



Redes Neuronales Autoencoder



Datos



Datos

ID

DISCRETAS

CONTINUAS

CATEGORICAS

ProcessId
ThreadId
UserId
EventId
Ip

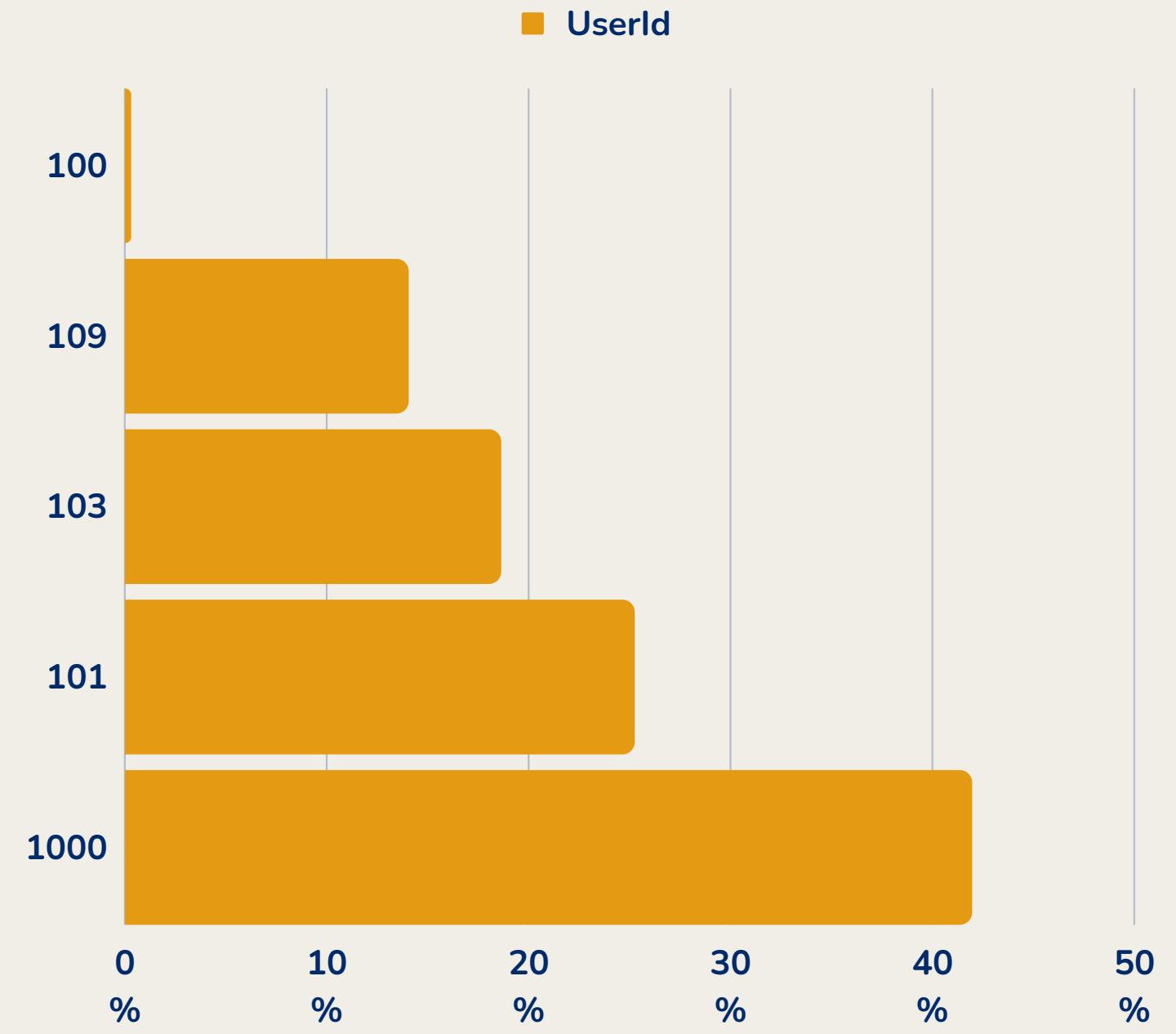
ArgsNum

Timestamp

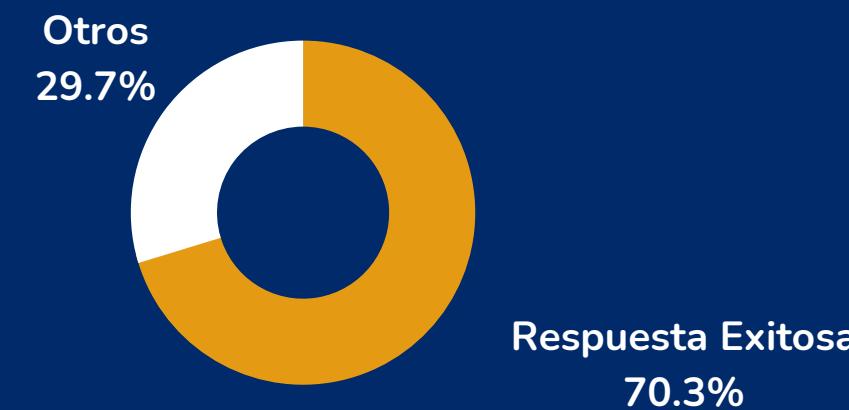
ProcessName
HostName
EventName
Args
User-agent
ReturnValue

Descriptivo Inicial

Los datos describen 378.425 eventos de 5 usuarios



Importante resaltar

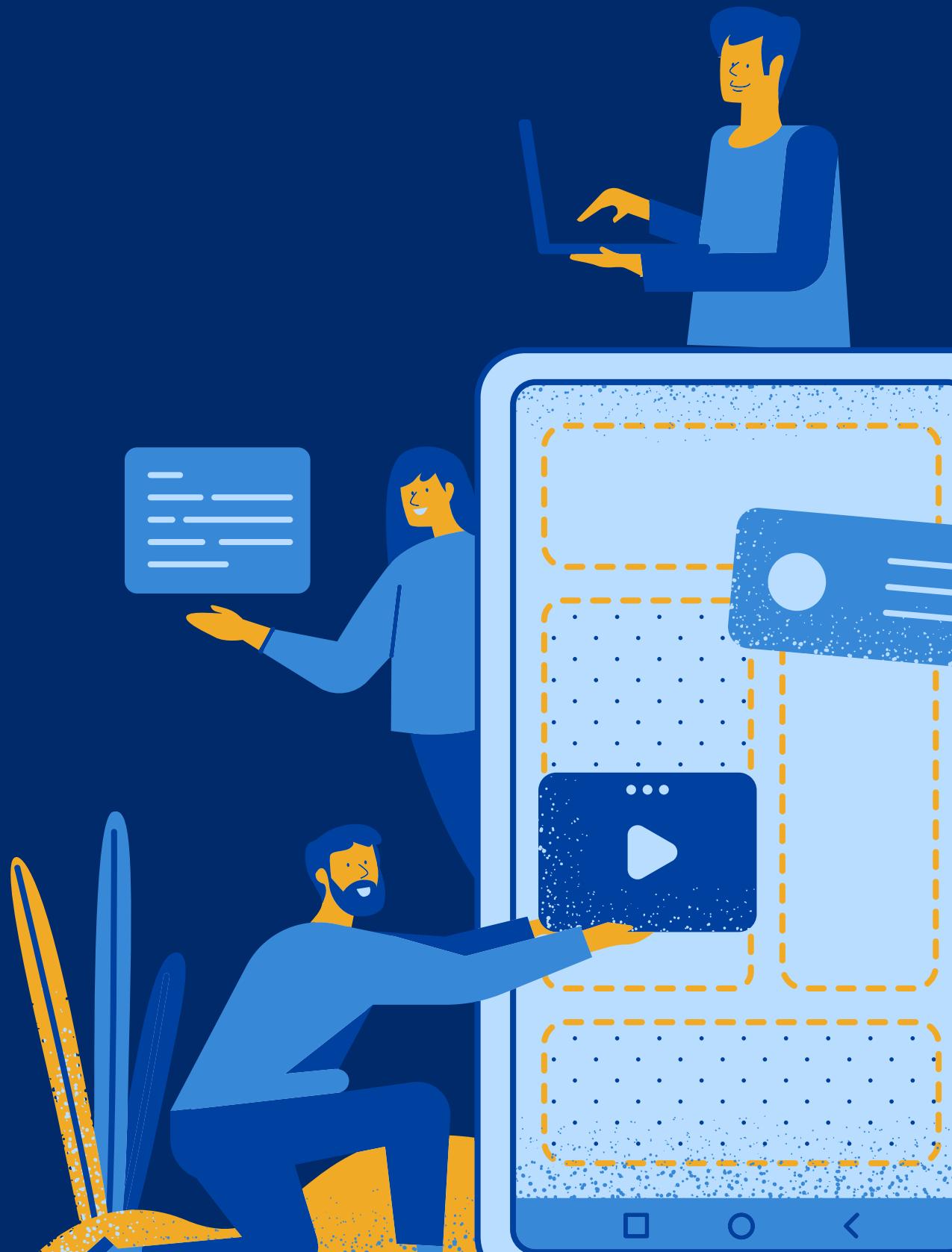


Datos relevantes



| | eventName |
|--------------------|-----------|
| openat | 28.61% |
| close | 28.58% |
| security_file_open | 27.76% |
| stat | 8.47% |
| fstat | 4.06% |
| otros | 2.52% |

Variables



Depuración

- Entendimiento de variables
- Corrección de variables
- Tokenización de strings
- Calculos timestamp
- Top Categorías
- Peticiones API datos externos



-Palabras frecuentes en argumentos por usuario
-Argumentos coherentes con proceso
-Argumentos con valores incoherentes
-Probabilidad de argumentos por usuario

-Tokenización, palabras clave en el process name
-Caracteres especiales en process name



-Tiempos promedio
-Frecuencias
-Usuario-proceso
-Usuario-evento
-Proceso-evento
-Usuario-ip

-Dispersión, Entropía, coeficiente de variación
-Percentiles de tiempos de peticiones
-Probabilidades de ip, eventos y procesos

-Probabilidad de eventos exitosos

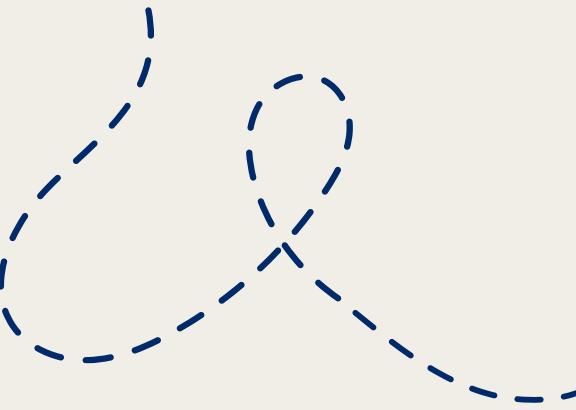
-Clasificación por tipo de eventos proceso, archivo, red

-Cantidad de direcciones ip
-Cantidad de eventos por ip
-Dispositivo, SO, navegador
-Geolocalización

-Ip sospechosas
-Association Rules



Args



args

```
[{"name": "domain", "type": "int", "value": "AF_UNIX"}, {"name": "type", "type": "int", "value": "SOCK_DGRAM|SOCK_CLOEXEC"}, {"name": "protocol", "type": "int", "value": 0}, {"name": "pathname", "type": "const char*", "value": "/proc/387/cgroup"}, {"name": "flags", "type": "int", "value": "O_RDONLY|O_LARGEFILE"}, {"name": "dev", "type": "dev_t", "value": 5}, {"name": "inode", "type": "unsigned long", "value": 39281}, {"name": "dirfd", "type": "int", "value": -100}, {"name": "pathname", "type": "const char*", "value": "/proc/387/cgroup"}, {"name": "flags", "type": "int", "value": "O_RDONLY|O_CLOEXEC"}, {"name": "mode", "type": "int", "value": 1640750884}, {"name": "fd", "type": "int", "value": 12}, {"name": "statbuf", "type": "struct stat*", "value": "0x7FFCEA3FFC20"}, {"name": "fd", "type": "int", "value": 12}]
```

| args | name | type | value | values | userid | processName | eventName |
|--|----------|-------------|-------------------------|------------------------------|--------|-----------------|--------------------|
| {"name": "domain", "type": "int", "value": "AF_UNIX"}, {"name": "type", "type": "int", "value": "SOCK_DGRAM SOCK_CLOEXEC"}, {"name": "protocol", "type": "int", "value": 0}, {"name": "pathname", "type": "const char*", "value": "/proc/387/cgroup"}, {"name": "flags", "type": "int", "value": "O_RDONLY O_LARGEFILE"}, {"name": "dev", "type": "dev_t", "value": 5}, {"name": "inode", "type": "unsigned long", "value": 39281}, {"name": "dirfd", "type": "int", "value": -100}, {"name": "pathname", "type": "const char*", "value": "/proc/387/cgroup"}, {"name": "flags", "type": "int", "value": "O_RDONLY O_CLOEXEC"}, {"name": "mode", "type": "int", "value": 1640750884}, {"name": "fd", "type": "int", "value": 12}, {"name": "statbuf", "type": "struct stat*", "value": "0x7FFCEA3FFC20"}, {"name": "fd", "type": "int", "value": 12} | domain | int | AF_UNIX | [af, unix] | 101 | systemd-resolve | socket |
| {"name": "domain", "type": "int", "value": "AF_UNIX"}, {"name": "type", "type": "int", "value": "SOCK_DGRAM SOCK_CLOEXEC"}, {"name": "protocol", "type": "int", "value": 0}, {"name": "pathname", "type": "const char*", "value": "/proc/387/cgroup"}, {"name": "flags", "type": "int", "value": "O_RDONLY O_LARGEFILE"}, {"name": "dev", "type": "dev_t", "value": 5}, {"name": "inode", "type": "unsigned long", "value": 39281}, {"name": "dirfd", "type": "int", "value": -100}, {"name": "pathname", "type": "const char*", "value": "/proc/387/cgroup"}, {"name": "flags", "type": "int", "value": "O_RDONLY O_CLOEXEC"}, {"name": "mode", "type": "int", "value": 1640750884}, {"name": "fd", "type": "int", "value": 12}, {"name": "statbuf", "type": "struct stat*", "value": "0x7FFCEA3FFC20"}, {"name": "fd", "type": "int", "value": 12} | type | int | SOCK_DGRAM SOCK_CLOEXEC | [sock, dgram, sock, cloexec] | 101 | systemd-resolve | socket |
| {"name": "domain", "type": "int", "value": "AF_UNIX"}, {"name": "type", "type": "int", "value": "SOCK_DGRAM SOCK_CLOEXEC"}, {"name": "protocol", "type": "int", "value": 0}, {"name": "pathname", "type": "const char*", "value": "/proc/387/cgroup"}, {"name": "flags", "type": "int", "value": "O_RDONLY O_LARGEFILE"}, {"name": "dev", "type": "dev_t", "value": 5}, {"name": "inode", "type": "unsigned long", "value": 39281}, {"name": "dirfd", "type": "int", "value": -100}, {"name": "pathname", "type": "const char*", "value": "/proc/387/cgroup"}, {"name": "flags", "type": "int", "value": "O_RDONLY O_CLOEXEC"}, {"name": "mode", "type": "int", "value": 1640750884}, {"name": "fd", "type": "int", "value": 12}, {"name": "statbuf", "type": "struct stat*", "value": "0x7FFCEA3FFC20"}, {"name": "fd", "type": "int", "value": 12} | protocol | int | 0 | [] | 101 | systemd-resolve | socket |
| {"name": "domain", "type": "int", "value": "AF_UNIX"}, {"name": "type", "type": "int", "value": "SOCK_DGRAM SOCK_CLOEXEC"}, {"name": "protocol", "type": "int", "value": 0}, {"name": "pathname", "type": "const char*", "value": "/proc/387/cgroup"}, {"name": "flags", "type": "int", "value": "O_RDONLY O_LARGEFILE"}, {"name": "dev", "type": "dev_t", "value": 5}, {"name": "inode", "type": "unsigned long", "value": 39281}, {"name": "dirfd", "type": "int", "value": -100}, {"name": "pathname", "type": "const char*", "value": "/proc/387/cgroup"}, {"name": "flags", "type": "int", "value": "O_RDONLY O_CLOEXEC"}, {"name": "mode", "type": "int", "value": 1640750884}, {"name": "fd", "type": "int", "value": 12}, {"name": "statbuf", "type": "struct stat*", "value": "0x7FFCEA3FFC20"}, {"name": "fd", "type": "int", "value": 12} | pathname | const char* | /proc/387/cgroup | [proc, cgroup] | 1000 | systemd | security_file_open |

Association Rules

| support | itemsets |
|----------|-----------------------|
| 0.014592 | (cloexec) |
| 0.054604 | (cmdline) |
| 0.097375 | (largefile) |
| 0.017120 | (lib) |
| 0.010257 | (locale) |
| 0.198068 | (o) |
| 0.199371 | (proc) |
| 0.196665 | (readonly) |
| 0.055044 | (stat) |
| 0.054649 | (status) |
| 0.014954 | (usr) |
| 0.013587 | (o, cloexec) |
| 0.012751 | (readonly, cloexec) |
| 0.054604 | (cmdline, proc) |
| 0.097375 | (o, largefile) |
| 0.097194 | (readonly, largefile) |
| 0.012257 | (lib, usr) |
| 0.196665 | (readonly, o) |

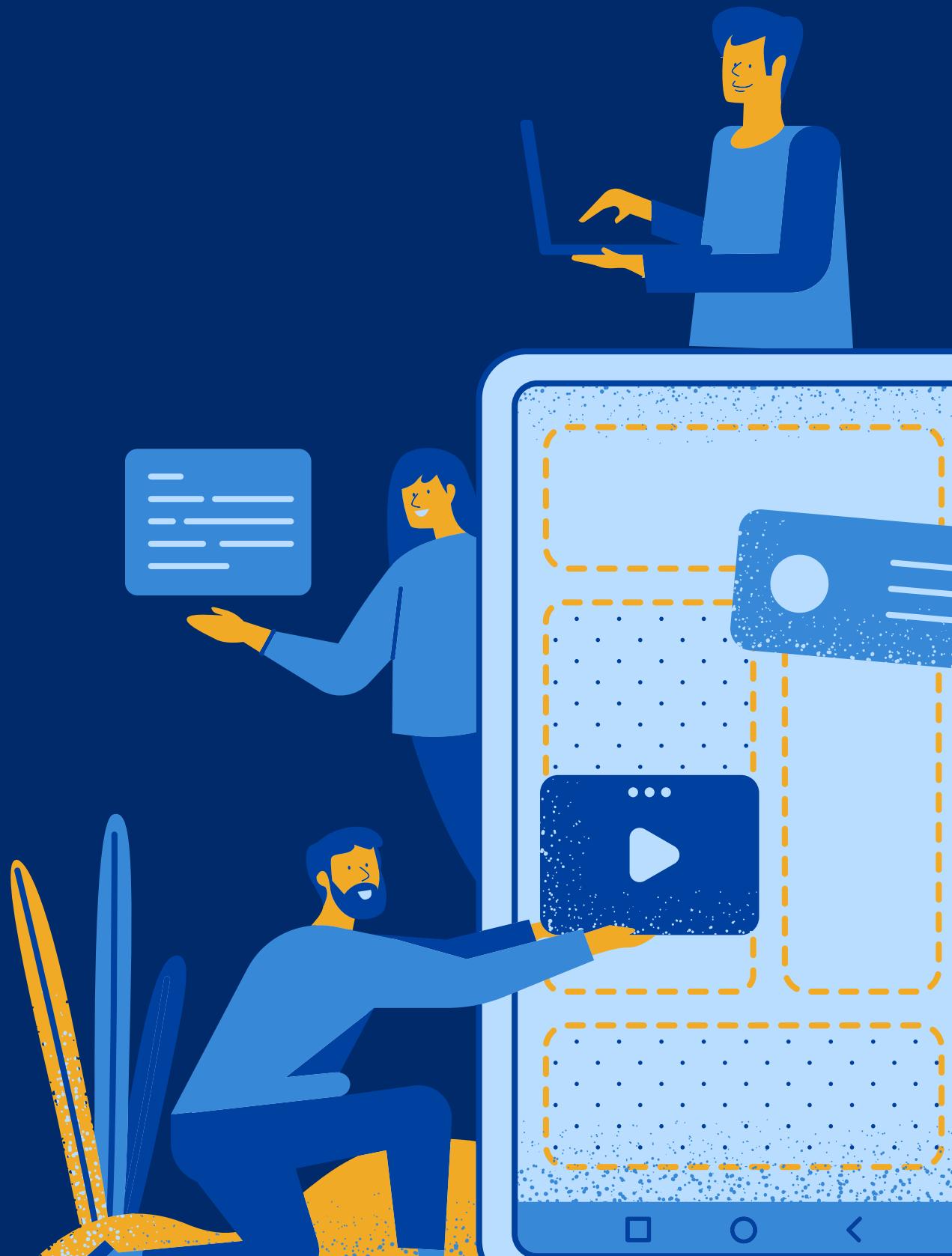


| antecedents | consequents | antecedent support | consequent support | support | confidence | lift | leverage | conviction | zhangs_metric |
|-------------|-------------|--------------------|--------------------|----------|------------|-----------|----------|------------|---------------|
| (cloexec) | (o) | 0.014592 | 0.198068 | 0.013587 | 0.931131 | 4.701068 | 0.010697 | 11.644290 | 0.798941 |
| (cloexec) | (readonly) | 0.014592 | 0.196665 | 0.012751 | 0.873825 | 4.443207 | 0.009881 | 6.366812 | 0.786413 |
| (cmdline) | (proc) | 0.054604 | 0.199371 | 0.054604 | 1.000000 | 5.015764 | 0.043718 | inf | 0.846871 |
| (largefile) | (o) | 0.097375 | 0.198068 | 0.097375 | 1.000000 | 5.048772 | 0.078088 | inf | 0.888444 |
| (largefile) | (readonly) | 0.097375 | 0.196665 | 0.097194 | 0.998143 | 5.075341 | 0.078044 | 432.704912 | 0.889593 |
| (usr) | (lib) | 0.014954 | 0.017120 | 0.012257 | 0.819653 | 47.875812 | 0.012001 | 5.449931 | 0.993976 |
| (readonly) | (o) | 0.196665 | 0.198068 | 0.196665 | 1.000000 | 5.048772 | 0.157712 | inf | 0.998254 |
| (o) | (readonly) | 0.198068 | 0.196665 | 0.196665 | 0.992918 | 5.048772 | 0.157712 | 113.436914 | 1.000000 |
| (stat) | (proc) | 0.055044 | 0.199371 | 0.055044 | 1.000000 | 5.015764 | 0.044070 | inf | 0.847266 |
| (status) | (proc) | 0.054649 | 0.199371 | 0.054649 | 1.000000 | 5.015764 | 0.043753 | inf | 0.846911 |

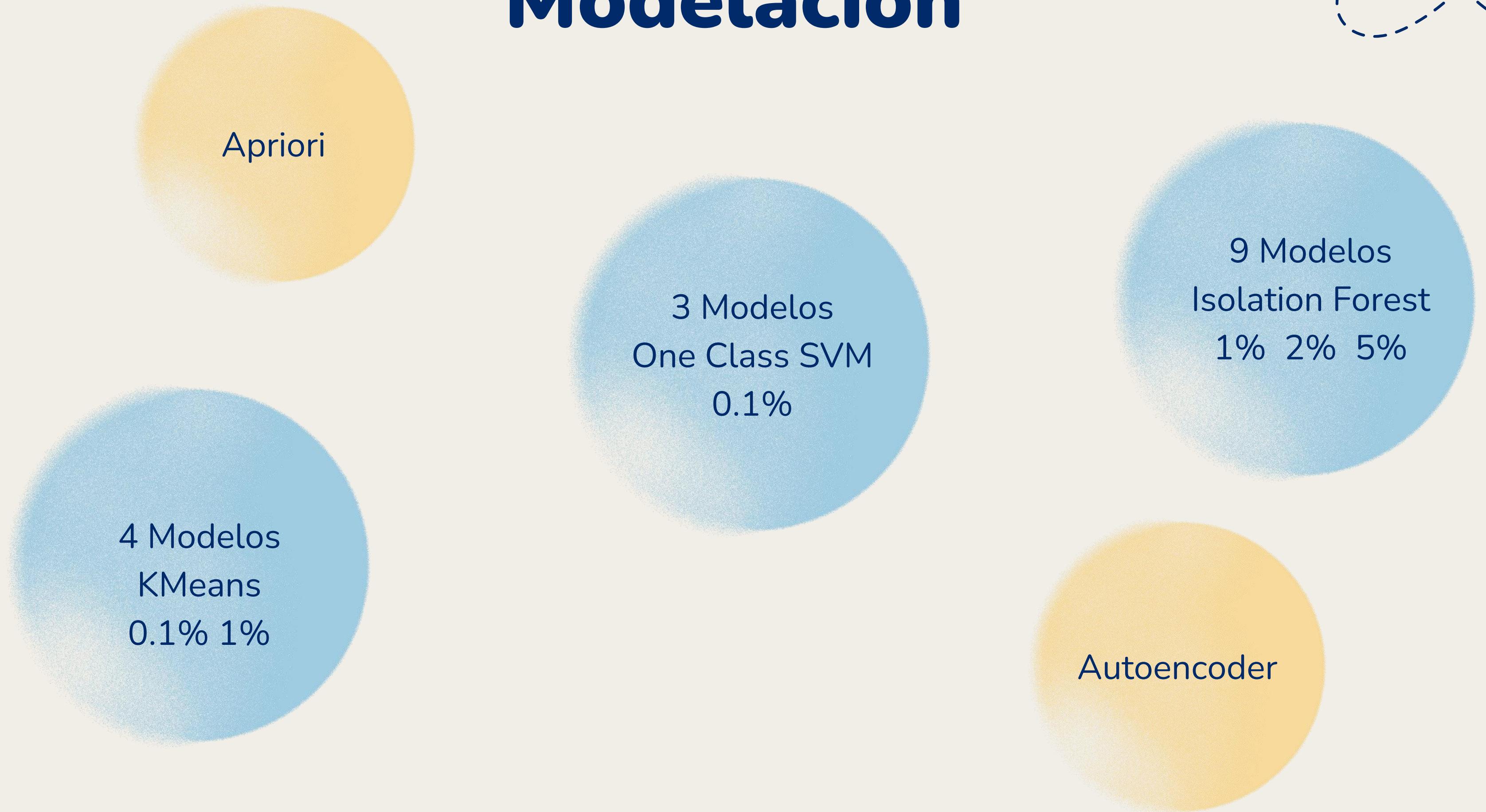
Event

| Tipo | Clasificación |
|----------|---|
| Proceso | setreuid, cap_capable, prctl, execve, clone, kill, sched, ss_exit, setuid, setregid, setgid, access, security_bprm_check |
| Red | Socket, Connect, Getsockname, Accept4, Bind, Accept |
| Archivos | openat,close, security_file_open, fstat, fchmod, stat, getdents64, unlink, dup3, dup2, dup, lstat, security_inode_unlink, unlinkat, umount, symlink |

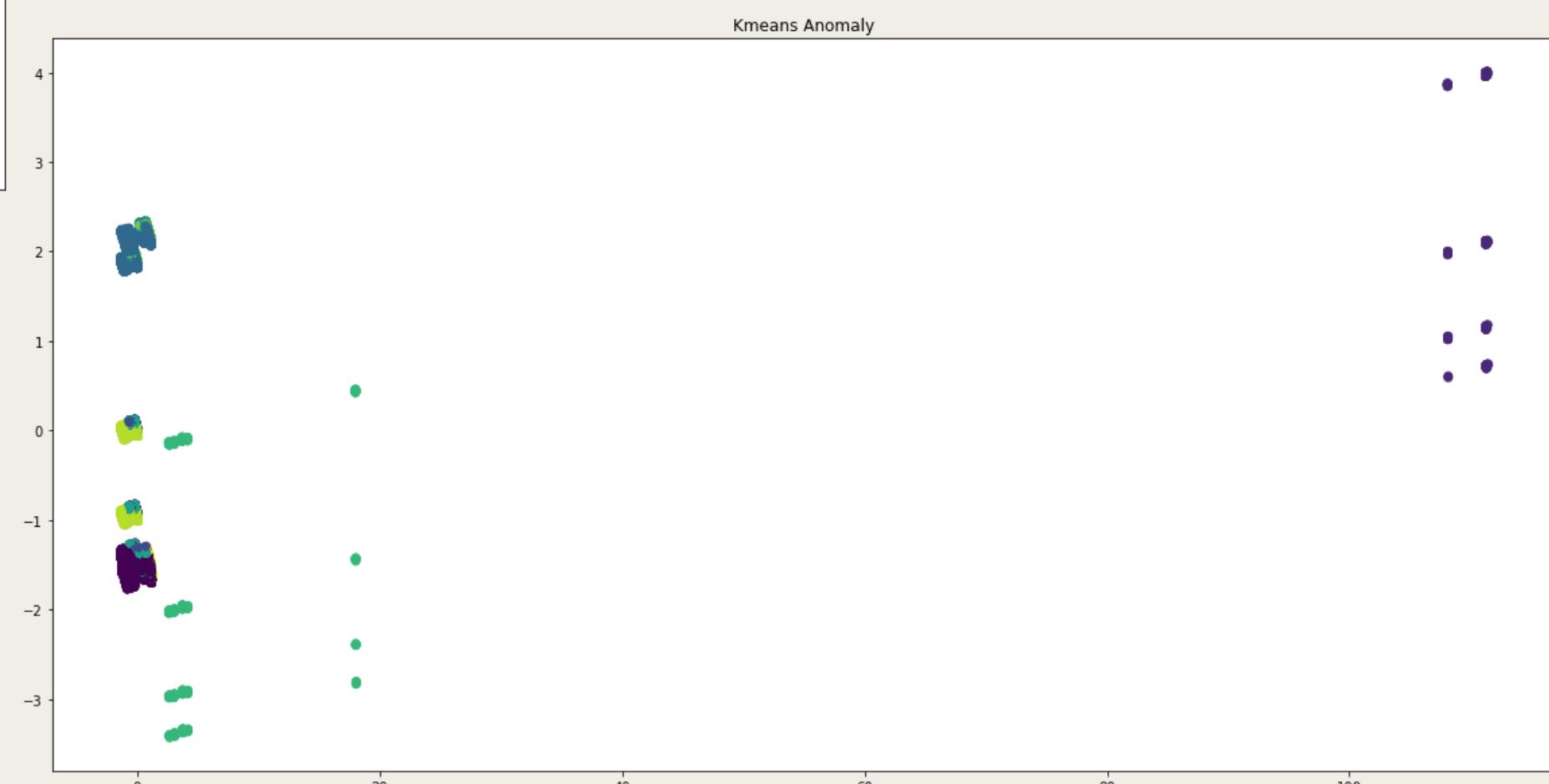
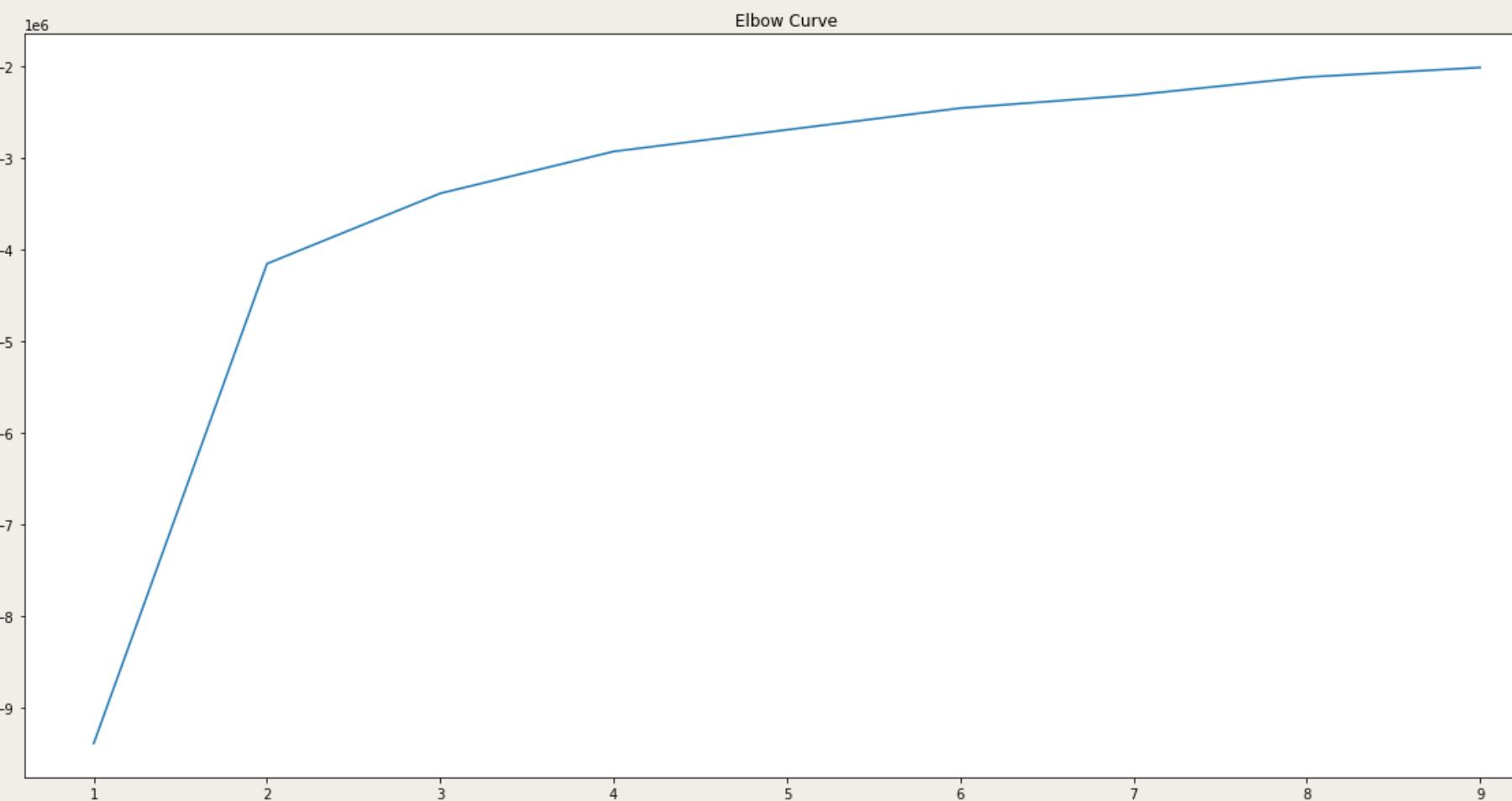
Modelado



Modelación

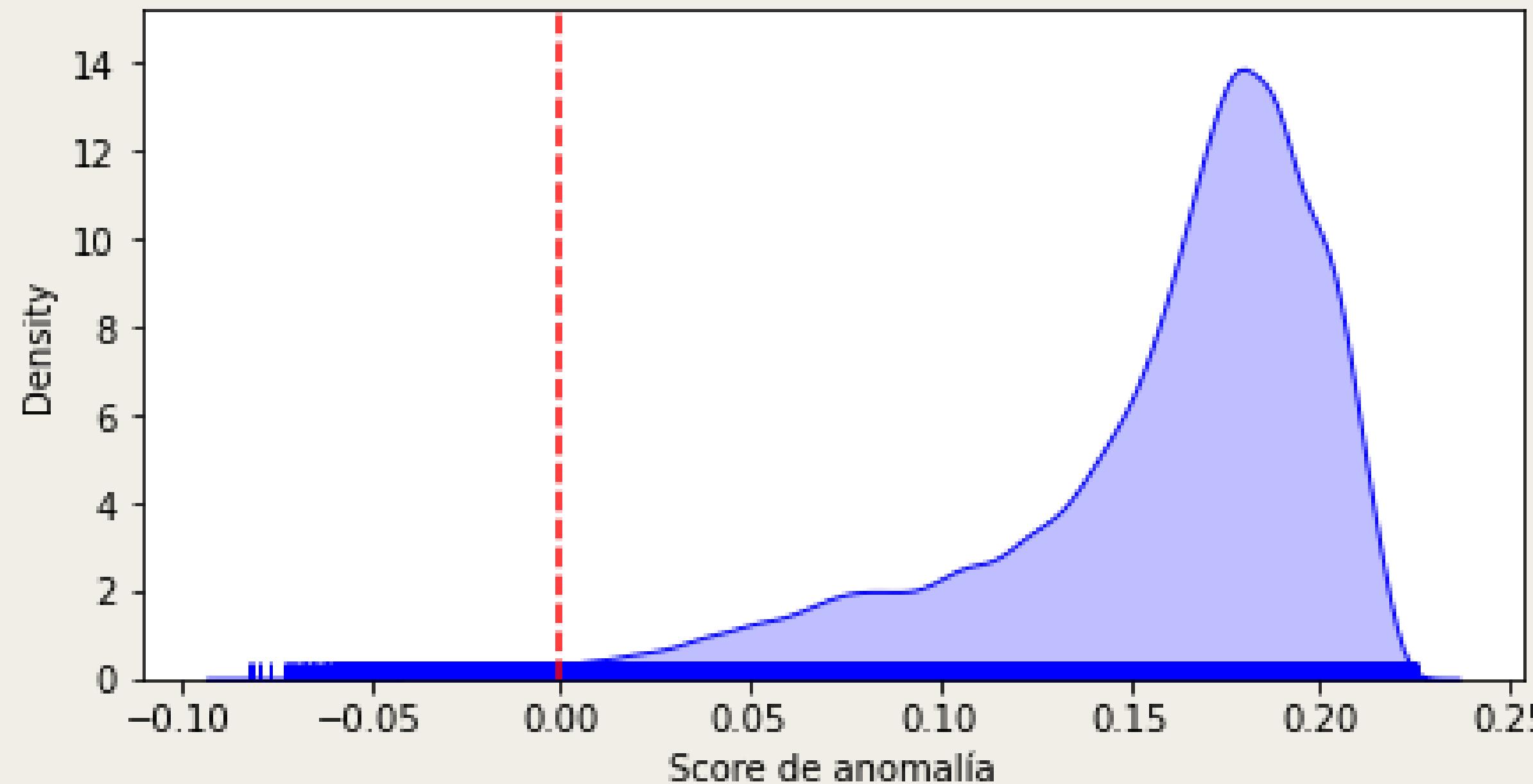


PCA Y KMEANS



Modelos

Distribución de los valores de anomalía



Evaluación

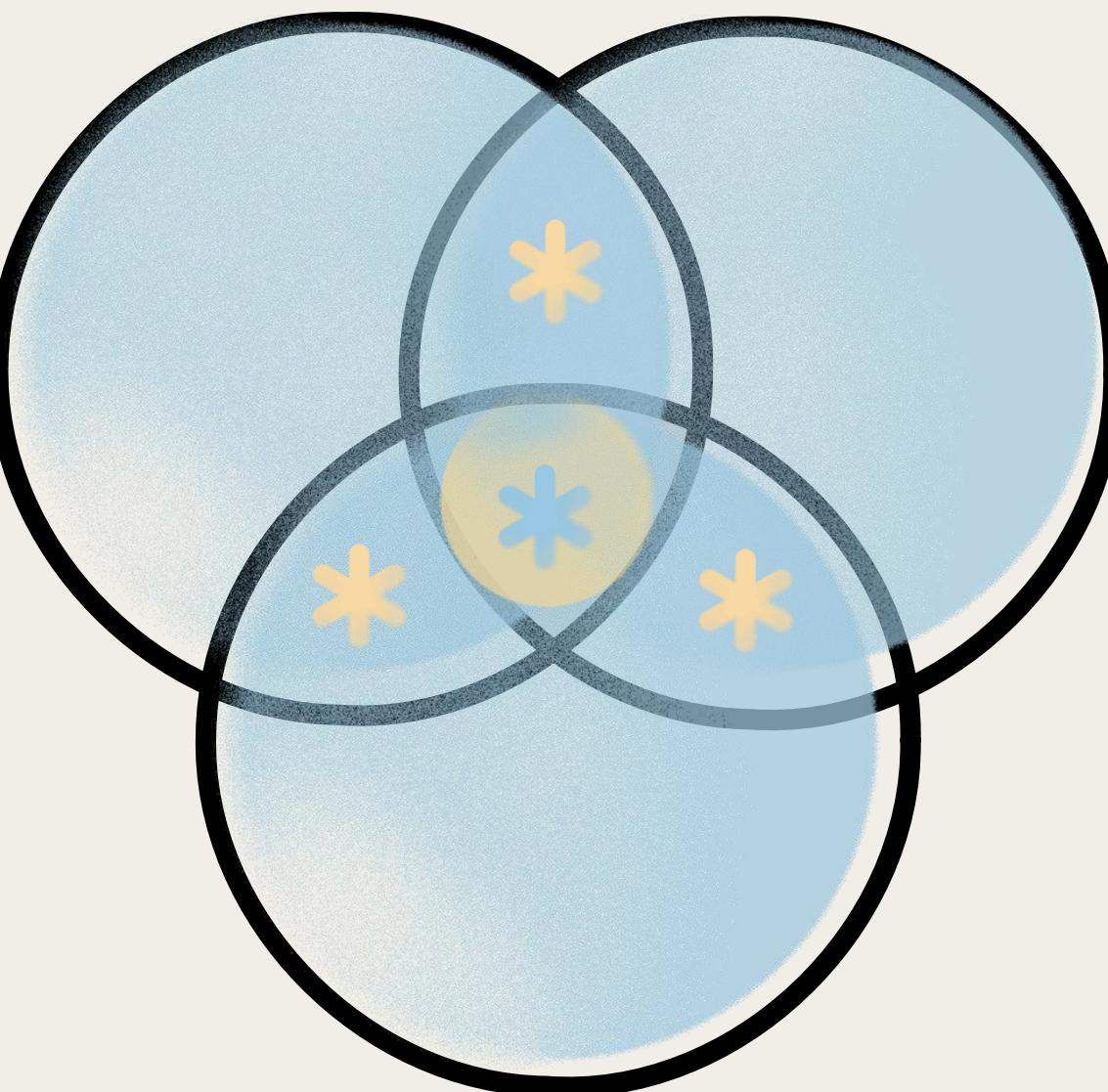
| Métrica | Descripción |
|-------------------|---|
| PCA | Reducción de Dimensionalidad |
| Silhouette | Homogeneidad intra cluster y heterogeneidad entre cluster |
| Decision function | Puntuación de anomalía de una instancia |
| Support | Proporción de transacciones que cumplen la regla |
| Zhang-metric | Importancia de los conjuntos de elementos frecuentes |

Evaluación

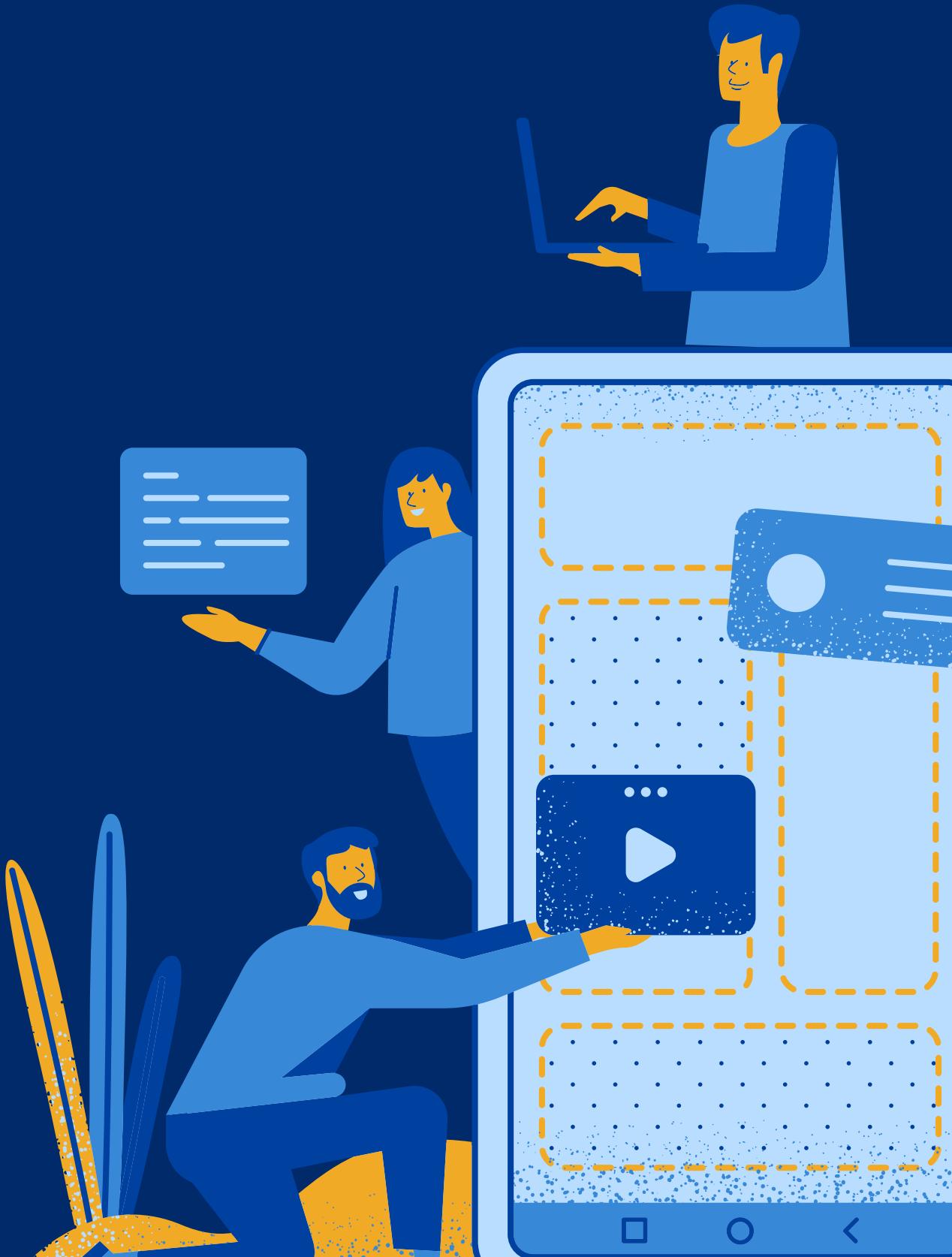
| Métrica | Descripción |
|------------|--|
| Confidence | Probabilidad condicional de que el consecuente ocurra dado que el antecedente ha ocurrido. |
| Lift | Fuerza de la asociación ante-conse |
| Leverage | Ganancia como frecuencia observada conjunta vs frecuencia esperada independiente |
| Conviction | Dependencia del consecuente con el antecedente |

Modelación

- * Anomalía Confirmada
- * Anomalía Recurrente
- Otras Anomalías



Resultados



Importante resaltar

Los Modelos
muestran
consistencia
respecto a lo que es
una anomalía en
nuestro conjunto de
datos



User 100

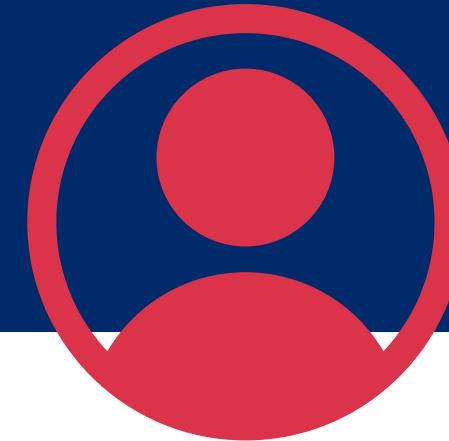
976 Obs



70,6%
Anomalías



- 67% ip sospechosas
- Eventos de archivos y de red
- Sin variación de argumentos ni procesos



- Eventos de tipo archivo
- Variación tiempos
- Argumentos anómalos

User 109

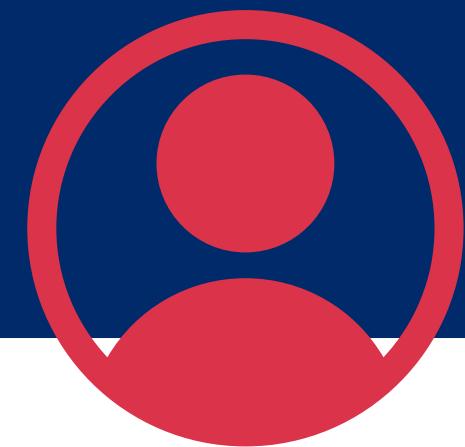


2%
Anomalías

53.015 Obs



- Ips no sospechosa
- Proceso ps



- 19% ip sospechosa de pakistan
- Procesos anómalos para el usuario sshd

User 103

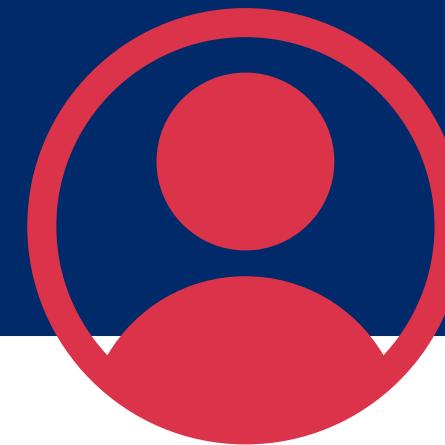
70.400 Obs



1,6%
Anomalías



- Ips no sospechosa
- Proceso ps



- 19% ip sospechosa de pakistan
- Procesos anómalos para el usuario systemd

User 101

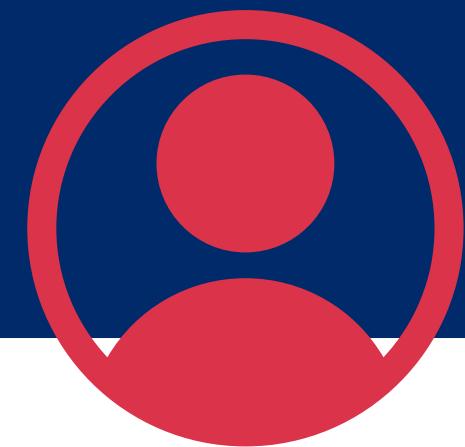
95.396 Obs



1,5%
Anomalías



- 67% ip sospechosa normal
- Dispositivos MAC, IPHONE
- Proceso ps



- Tiempos altos
- Procesos anómalos para el usuario

User 1000



3,7%
Anomalías

158.637 Obs



- 67% ip sospechosa normal
- Dispositivos MAC, IPHONE



- Variación alta en los argumentos
- Respuesta no exitosa

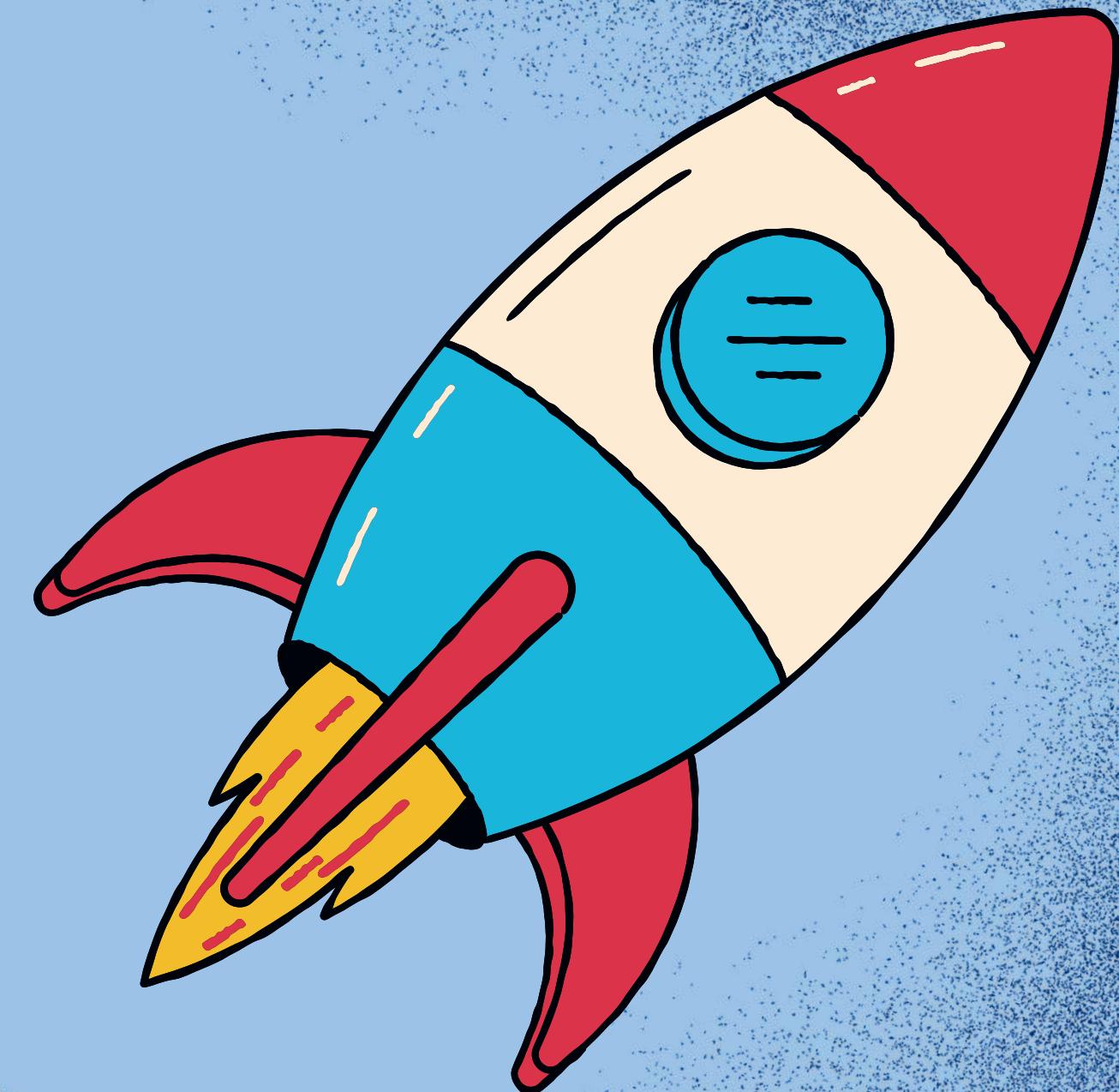
Oportunidades de Mejora

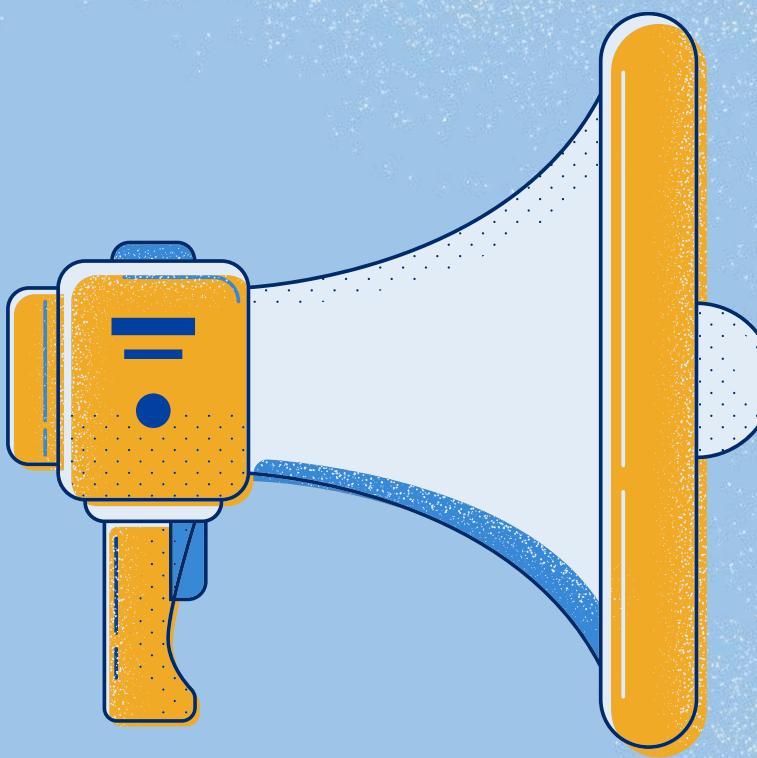
- Timestamp
- Profundizar en Argumentos
- Transformaciones u otros modelos
- Mejor Procesamiento
- Aprovechamiento Association rules



https://github.com/88marisol/Anomaly_Detection.git

Despliegue





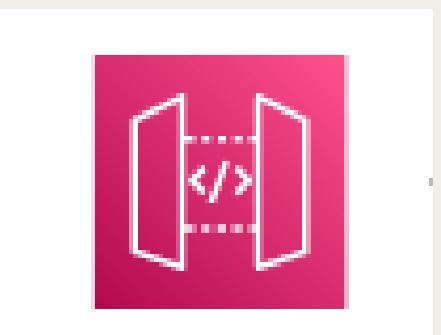
<https://w8063fiule.execute-api.us-east-1.amazonaws.com/test>



S3



Lambda



API Gateway

