UNIVERSIDAD CATÓLICA SEDES SAPIENTIAE

Inteligencia Artificial

Ing. Juancarlos Santana Huamán
jsantana@ucss.edu.pe

Breve historia de los datos y las dificultades de administración asociadas

- Alrededor de 2,5 trillones (2 500 000 000 000 000 000) de bytes de datos se crean a diario, o aproximadamente 1,7 megabytes por segundo y persona en el planeta.
- Los datos no se pierden por arte de magia. Son los usuarios quienes los pierden. Sin embargo, en los entornos de trabajo actuales, distribuidos y centrados en la nube (cloud-first), la mayoría de los departamentos de TI tienen poca visibilidad de las pérdidas de datos provocadas por las personas y ninguna capacidad para gestionarlas. Este libro electrónico analiza cinco **fugas de datos** reales para descubrir cómo ocurrieron, las consecuencias para la empresa, y cómo podrían haberse evitado. Descubrirá:
  - Cómo hacer frente a los riesgos de apps de terceros con acceso OAuth
  - Qué hacer con los usuarios comprometidos
  - Cómo detectar y corregir los errores de configuración de los servidores
  - Las ventajas de un enfoque centrado en las personas de la prevención de la pérdida de datos

¿Qué es la clasificación de datos?

- La clasificación de datos es un método para definir y categorizar los archivos y otra información empresarial clave. Se usa principalmente en grandes organizaciones para crear sistemas de seguridad que siguen estrictas normativas de cumplimiento, pero también se puede usar en entornos reducidos. El uso más importante de la clasificación de datos es comprender la sensibilidad de la información almacenada para crear las herramientas adecuadas de ciberseguridad, controles de acceso y monitorización.
- La clasificación es el proceso de categorizar los activos de datos basándose en la sensibilidad de la información. Mediante la clasificación de datos, las organizaciones pueden determinar dos elementos clave:
  - Quién debe tener autorización para acceder a esta.
  - Qué políticas de protección aplicar al almacenarla y transferirla.
- La clasificación también ayuda a determinar los estándares normativos aplicables para proteger los datos. En general, la clasificación ayuda a las organizaciones a gestionar mejor sus datos para propósitos de privacidad, cumplimiento y ciberseguridad.

Motivos para realizar clasificación de datos

- Todas las organizaciones deben clasificar los datos que crean, gestionan y almacenan. Pero es aún más importante para entornos de grandes empresas. Esto es porque las grandes empresas tienen activos de datos distribuidos entre múltiples ubicaciones, incluyendo en la nube.
- Los administradores deben hacer seguimiento y auditar esta información para garantizar que tenga los controles de acceso y autenticación adecuados. La clasificación de datos permite a los administradores identificar las ubicaciones en donde se almacenan datos delicados y determinar cómo se debe acceder y compartir estos datos.
- La clasificación es un primer paso fundamental para cumplir con casi cualquier normativa de conformidad de datos. HIPAA, RGPD, FERPA y otros organismos reguladores gubernamentales exigen que los datos se etiqueten para que los controles de seguridad y autenticación puedan limitar el acceso. El etiquetado de datos ayuda a organizar y proteger la información. El ejercicio también reduce la duplicación innecesaria de datos, reduce los costes de almacenamiento, incrementa el rendimiento y hace posible registrarlo cuando se comparte.
- La clasificación de datos es la base de unas políticas eficaces para protección de datos y reglas para la prevención de pérdida de datos (DLP). Para tener unas reglas de DLP eficaces, primero se deben clasificar sus datos para asegurarse de conocer bien los datos almacenados en cada archivo.

Tipos de clasificación de datos

- Cualquier dato almacenado se puede clasificar en categorías. Para clasificar sus datos, es necesario formular diversas preguntas a medida que se va descubriendo y revisando. Use las siguientes preguntas de ejemplo al revisar cada sección de sus datos:
  - ¿Qué información almacena para sus clientes, empleados y proveedores?
  - ¿Qué tipos de datos crea la organización al generar un nuevo registro?
  - ¿Qué tan sensibles son los datos en una escala numérica (¿p. ej. 1-10, donde 1 indica la mayor sensibilidad?)
  - ¿Quién debe acceder a estos datos para seguir operando con productividad?
- Usando estas preguntas, es posible definir categorías generales para sus datos, por ejemplo:
- **Alta sensibilidad:** Estos datos deben ser protegidos y monitorizados para protegerlos contra agentes de amenaza. Con frecuencia quedan sujetos a las normativas de cumplimiento como información que requiere de estrictos controles de acceso y que también minimice la cantidad de usuarios que pueden acceder a los datos.
- **Sensibilidad intermedia:** Los archivos y datos que no se pueden revelar al público, pero que si sufriesen de una filtración no implicarían un nivel de riesgo significativo, se podrían considerar de riesgo intermedio. Requieren de controles de acceso, al igual que los datos altamente sensibles, pero una gama más amplia de usuarios puede acceder a estos.
- **Baja sensibilidad:** Estos datos típicamente constan de información pública que no requiere de mucha seguridad para protegerla de una filtración de datos.

Técnicas de clasificación de datos

- La clasificación de datos colabora estrechamente con otras tecnologías para proteger y gobernar mejor los datos. Si la organización sufre una vulneración de datos, la clasificación de datos ayuda a los administradores a identificar datos perdidos y potencialmente ayuda a rastrear al cibercriminal.
- Aquí mencionamos algunas técnicas de clasificación de datos:
  - **Gestión de acceso a la identidad (IAM):** Las herramientas de IAM permiten a los administradores determinan quién y qué puede acceder a los datos. Los usuarios con permisos similares se pueden agrupar. Los grupos reciben niveles de autorización y se gestionan como una unidad individual. Cuando un usuario se va, al usuario se le puede eliminar del grupo, lo que elimina todos los permisos de dicho usuario. Este tipo de reagrupación y organización optimiza la gestión de permisos en toda la red.
  - **Cifrado de datos:** Ciertos activos de datos deben estar cifrados tanto en reposo como en movimiento. Los datos “en reposo” se almacenan en cualquier dispositivo de almacenamiento, típicamente en un disco duro. Los datos “en movimiento” son aquellos que se transfieren a través de una red. El **cifrado de datos** los vuelve ilegibles para cualquier atacante que logre interceptarlos.
  - **Automatización:** La automatización funciona mano a mano con herramientas de monitorización para hallar, clasificar y etiquetar datos para su revisión administrativa. Algunas herramientas integran la inteligencia artificial (IA) y el aprendizaje automático (ML) para detectar, etiquetar y clasificar datos automáticamente. Las tecnologías también pueden ayudar a identificar amenazas que podrían usarse para robarla. Con los datos etiquetados, los administradores pueden usar la IAM para aplicar permisos y evitar que ciertas amenazas específicas obtengan acceso a datos almacenados.
  - **Análisis forense de datos:** El análisis forense es el proceso de identificar qué fue lo que salió mal y quién vulneró a la red. Después de una filtración de datos, el análisis forense de datos se encarga de recopilar y preservar evidencia para continuar con las investigaciones. El análisis forense de datos generalmente es un proceso de dos partes. Primero, las herramientas de automatización recopilan los datos; después, un analista humano identifica las anomalías y las investiga.

Niveles de clasificación de datos

- Mediante estos niveles, podrá clasificar mejor sus datos. La clasificación de datos se suele dividir en cuatro categorías:
  - **Datos públicos.** Los datos están disponibles al público, ya sea localmente o por internet. Los datos públicos requieren un nivel bajo de seguridad, porque su revelación no vulneraría los requisitos de cumplimiento.
  - **Datos de uso interno únicamente.** Memorandos, propiedad intelectual y mensajes de correo electrónico son algunos ejemplos de datos que deben estar restringidos a empleados internos.
  - **Datos confidenciales.** La diferencia entre los datos de uso interno únicamente y los confidenciales es que los confidenciales requieren de autorización para su acceso. Usted puede asignarles autorización a empleados específicos o a proveedores externos autorizados.
  - **Datos restringidos.** Los datos restringidos suelen estar relacionados con información gubernamental a la que solo pueden acceder individuos autorizados. La revelación de datos restringidos puede causar daños irreparables a la reputación e ingresos corporativos.

Alineación con una lista de activos

- Antes de comenzar con una revisión de clasificación de datos, deben estar en la misma página. Al comienzo de la revisión, crean una lista de activos para definir sus categorías de negocio. Por ejemplo, a lo mejor tiene archivos en los que almacena datos de tecnología, financieros y de los clientes. La definición de categorías alinea sus requerimientos de seguridad con sus datos.
- Este paso también implica el aplicar los niveles de clasificación de datos definidos en la sección anterior. Para cada categoría, es probable que tenga diferentes niveles de datos para cada grupo de archivos. Este primer paso crea una base para todo el proceso de clasificación de datos.

Proceso de clasificación de datos

- Cuando decida que es momento de clasificar datos para cumplir con los estándares de conformidad, el primer paso es implementar procedimientos para asistir con la ubicación y clasificación de datos, así como para determinar las medidas de ciberseguridad adecuadas. La ejecución de cada procedimiento depende de los estándares de conformidad de su organización y de la infraestructura que mejor proteja a los datos. Los pasos generales para clasificar datos son:
  - **Ejecutar una evaluación de riesgos:** Una evaluación de riesgos determina la sensibilidad de los datos e identifica cómo un atacante podría vulnerar las defensas de la red.
  - **Desarrollar políticas y estándares de clasificación:** Si se generan datos adicionales en el futuro, una política de clasificación permite optimizar y volver repetible al proceso, facilitando así el trabajo al personal y a la vez minimizando los errores en el proceso.

Proceso de clasificación de datos

- **Categorización de datos:** Ya con políticas establecidas y con una evaluación de riesgo realizada, categorice sus datos basándose en su sensibilidad, quién debe tener acceso y si las penalizaciones de conformidad deben ser hechas públicas.
- **Hallar la ubicación de almacenamiento de sus datos:** Antes de implementar las defensas de ciberseguridad adecuadas, usted debe saber en dónde se almacenan los datos. Identificar las ubicaciones de almacenamiento de datos indica el tipo de ciberseguridad necesario para proteger los datos.
- **Identificar y clasificar sus datos:** Con los datos identificados, ya puede clasificarlos. El software externo le ayuda con este paso para facilitarle la clasificación y rastreo de los datos.
- **Implementar controles:** Los controles empleados deberían solicitar autenticación y autorización a cada usuario y recurso que necesite acceso a los datos. Ese acceso debe otorgarse “solo si es necesario”, lo que significa que los usuarios solamente reciben acceso si necesitan ver los datos para ejecutar una función laboral.
- **Monitorización de acceso y datos:** La monitorización de datos es un requisito de cumplimiento y para la privacidad de sus datos. Sin la monitorización, un atacante podría disponer de meses enteros para **exfiltrar datos** de la red. Los controles de monitorización adecuados detectan anomalías y reducen el tiempo necesario para detectar, mitigar y erradicar una amenaza de la red.

Optimizar el proceso de clasificación de datos

- Si bien es posible optimizar el proceso de clasificación de datos y hasta automatizarlo parcialmente, el proceso requiere de elementos de revisión humana y procedimientos manuales.
- Los sistemas automatizados sugieren etiquetas y clasificaciones, pero es mediante una revisión humana que se determina si estas etiquetas son correctas. Los objetivos y estándares deben ser esbozados y definidos, cosa que precisa de revisores humanos y personal de TI.
- Las herramientas automatizadas señalan activos para que los revisen seres humanos. La lista muestra los objetos (tales como los datos de un cierto cliente) y las reglas (como HIPAA o PCI-DSS) que se aplican a cada uno. Algunas herramientas de automatización pueden indexar objetos. (La indexación es el proceso de clasificar y organizar datos para permitir búsquedas rápidas y eficientes en la red).
- Otras políticas también se aplican durante el proceso de clasificación de datos. El **reglamento general de protección de datos** (RGPD) es una normativa de la UE que les da a los consumidores el derecho a que sus datos sean eliminados. Las organizaciones deben cumplir con esto al almacenar datos de los consumidores en la UE. Algunas herramientas de clasificación de datos indexan objetos de manera tal que puedan eliminarse rápidamente cuando los clientes así lo soliciten.

Ejemplos de clasificación de datos

- Uno de los pasos más complejos de la clasificación de datos es comprender los riesgos. Si bien los estándares de conformidad supervisan la mayor parte de los datos privados delicados, las organizaciones deben adherirse a las normativas de conformidad aplicables a diferentes datos almacenados en archivos y bases de datos. La clasificación de datos ayuda a proteger datos y a garantizar la conformidad. Es fundamental para cumplir con los requisitos del RGPD. (De hecho, las organizaciones deben indexar los datos de los consumidores en la UE, para que estos puedan borrarse si así se solicita).
- El RGPD también exige la protección de información personal secundaria de los clientes, como el origen étnico, las opiniones políticas, su raza y sus creencias religiosas. Para ello, las organizaciones deben clasificar estos datos y determinar los permisos adecuados entre los diversos activos digitales. La clasificación es la que determina quiénes pueden acceder a estos datos para que no se les dé un uso indebido. Solo entonces pueden evitar el revelar información privada de los consumidores, así como costosas filtraciones de datos.

Ejemplos de clasificación de datos

- Tres pasos para clasificar para el RGPD serían:
  - **Localizar y auditar los datos.** Antes de la clasificación, los administradores deben identificar dónde se almacenan los datos y qué reglas los afectan.
  - **Crear una política de clasificación.** Para garantizar el cumplimiento, es necesario crear estándares y procedimientos de clasificación de datos para definir cómo su organización almacena y transfiere datos delicados.
  - **Organizar y asignar prioridades a los datos.** Con la priorización, su organización puede determinar la clasificación de datos y los permisos necesarios para acceder a esta.
- Aquí indicamos algunos ejemplos de sensibilidad de datos que podrían categorizarse como alta, media y baja.
  - **Alta sensibilidad:** Imagínese que su empresa registra números de tarjetas de crédito como método de pago de los clientes que le compran. Estos datos deben tener controles de autorización estrictos, auditorías para detectar solicitudes de acceso y cifrado aplicado a los datos almacenados y transmitidos. Una filtración de datos probablemente causaría daños tanto al cliente como a la organización, así que deben clasificarse como altamente sensible con estrictos controles de ciberseguridad.
  - **Sensibilidad intermedia:** Para cada proveedor externo, usted tiene un contrato con firmas que ejecutan un acuerdo. Estos datos no causarían daño a los clientes, pero aun así son información delicada que describe detalles del negocio. Estos archivos se considerarían de sensibilidad intermedia.
  - **Sensibilidad baja:** Los datos para consumo del público pueden considerarse de baja sensibilidad. Por ejemplo, el material de marketing publicado en su web no necesitaría de controles estrictos, dado que está disponible al público y ha sido creado para una audiencia general.

Uso de la inteligencia artificial (IA) para clasificación de datos

UNIVERSIDAD CATÓLICA SEDES SAPIENTIAE

- La clasificación de datos precisa de interacción humana, pero buena parte del proceso se puede automatizar. Para añadir la automatización a las capacidades de toma de decisiones. La automatización por IA garantiza que las organizaciones puedan identificar, clasificar y proteger sus documentos continuamente, lo que implica que el motor continuamente escanea y revisa nuevos documentos a medida que se agregan al entorno.
- El módulo de Aprendizaje Activo ingiere alrededor de 20 documentos por categoría para comenzar el proceso y mejorar su precisión. El motor de clasificación de datos emplea modelos de aprendizaje automático para identificar patrones. Cada grupo de archivos debe ser lo suficientemente diverso como para que los algoritmos de aprendizaje automático puedan volverse más precisos.
- Los modelos de aprendizaje automático predicen las etiquetas para los documentos y determinan la exactitud de las predicciones. Al revisor se le muestra un “nivel de confianza” para reevaluar los datos del modelo para otra ronda de clasificación de información. Si el modelo indica que la exactitud es baja, los examinadores humanos pueden actualizar los modelos para tener conjuntos de archivos más diversos que permitan incrementar la exactitud. El motor se “entrena” a sí mismo aprovechando la nueva información para producir resultados mejores y óptimos, así que asigna permisos de acceso a usuarios solamente en los archivos necesarios para realizar sus funciones laborales.
- El software de clasificación de datos basada en IA reduce buena parte de los gastos generales de un proceso que podría llevar meses. Escanea todos sus archivos de forma automática, identifica su contenido, asigna las categorías y niveles de datos correctos y después le permite determinar la seguridad de salvaguarda adecuada.

Importancia de la clasificación de datos

- El “nivel de sensibilidad” de los datos determina cómo los procesa y los protege. Incluso si usted sabe que los datos son importantes, es necesario evaluar sus riesgos. El proceso de clasificación de datos le ayuda a descubrir potenciales amenazas e implementar las soluciones de ciberseguridad más beneficiosas para su empresa.
- Al asignar niveles de sensibilidad y categorizar los datos, usted comprende las reglas de acceso para los datos clave. Usted puede monitorizar mejor los datos para hallar potenciales filtraciones de datos y, más importante aún, mantener un alto nivel de conformidad. Las normativas de conformidad le ayudan a determinar cuáles son los controles de ciberseguridad adecuados, pero antes debe ejecutar una evaluación de riesgos y clasificar los datos. Las organizaciones suelen requerir de ayuda externa para la clasificación de datos, de modo que la implementación de la ciberseguridad se pueda ejecutar más eficientemente.
- La exactitud de la clasificación de datos es fundamental para futuras estrategias de DLP; por tanto, muchas organizaciones, grandes y pequeñas, han optado por usar automatización basada en IA. La inteligencia artificial aprovecha los modelos de aprendizaje automático para determinar el nivel de clasificación y categoría adecuados.

Buenas prácticas de clasificación de datos

- Observar buenas prácticas de clasificación de datos hace mucho más eficiente tanto a la creación de políticas como al proceso en general. Las buenas prácticas definen los pasos a seguir para indexar y etiquetar correctamente los activos digitales para que ninguno se pase por alto o se gestione mal.
- Las organizaciones deben seguir estas buenas prácticas:
  - Identificar cuidadosamente en dónde se ubican todos los datos delicados, incluyendo la propiedad intelectual, en todas las ubicaciones de almacenamiento.
  - Definir categorías de datos, para que los datos delicados se puedan etiquetar y configurar con los permisos adecuados. Las categorías deben ser granulares, para que los permisos también lo sean. Las categorías también deben permitir a los administradores categorizar los datos en grupos.
  - Identificar los datos más sensibles y claves. Después, las herramientas de automatización pueden etiquetarlos con la clasificación y mandatos normativos correctos.
  - Capacitar a los empleados para que comprendan cómo manejar información delicada. Darles las herramientas necesarias para proteger datos delicados y seguir buenas prácticas de ciberseguridad.
  - Examinar todos los estándares normativos de modo que se cumpla con las reglas y se eviten sanciones.
  - Crear políticas que permitan a los usuarios identificar datos mal clasificados o sin clasificar, para arreglar el problema.
  - Usar IA en donde se pueda mejorar la exactitud y acelerar el proceso de clasificación de datos.

Clasificación de sensibilidad de datos

- La clasificación de datos requiere que evalúe el nivel de sensibilidad de los datos en toda su organización. Estos niveles generalmente varían de alto a medio a bajo y se correlacionan con lo dañino que sería si esos datos se perdieran, robaran o comprometerán.
- Clasificar los datos de esta manera ayuda a las organizaciones a entender dónde enfocar sus esfuerzos de mitigación de riesgos. Cuanto más sensibles sean los datos, más necesita su organización centrarse en protegerlos.

Datos de baja sensibilidad

- Los datos de baja sensibilidad son datos que tendrían poco o ningún impacto si se vieran comprometidos, perdidos o destruidos (aunque una organización aún puede implementar controles de seguridad para protegerse contra daños). Los datos de baja sensibilidad son para uso público y no requieren ninguna protección de confidencialidad. Generalmente se etiquetan como datos no restringidos o públicos, dependiendo de su modelo de clasificación.
- Ejemplos de datos de baja sensibilidad incluyen:
  - Información pública y páginas web, como ofertas de trabajo, publicaciones de blog, etc.
  - Comunicados de prensa
  - Directorio de empleados

Datos de sensibilidad media

- Los datos de sensibilidad media son datos que no tendrían un impacto catastrófico si fueran comprometidos, perdidos o destruidos, pero resultarían en algún riesgo para una organización. Por lo tanto, estos datos solo deben ser accesibles para el personal interno al que se le haya concedido acceso y comúnmente se etiquetan como internos o privados.
- Ejemplos de datos de sensibilidad media incluyen:
  - Correos electrónicos o documentos internos que no contienen datos confidenciales
  - Contratos con proveedores
  - Información de gestión de servicios de TI o telecomunicaciones

Datos de alta sensibilidad

- Los datos de alta sensibilidad son datos que, si se comprometen, pierden o destruyen, tendrían un impacto catastrófico en una organización. Por lo tanto, las organizaciones deben implementar los controles de acceso más estrictos para los datos de alta sensibilidad. Dado que el acceso está limitado según la necesidad de saber, los datos de alta sensibilidad usualmente se etiquetan como datos confidenciales o restringidos.
- Ejemplos de datos de alta sensibilidad incluyen:
  - Registros financieros, como números de tarjetas de crédito
  - Datos médicos y biométricos, incluida la información de salud protegida (PHI)
  - Registros de empleados, incluida la información de identificación personal (PII) como números de Seguridad Social
  - Datos de autenticación, como credenciales de inicio de sesión

Modelos y esquemas de clasificación de datos

- Un modelo y esquema de clasificación de datos define cómo una organización identifica y categoriza sus activos de datos. Típicamente, éstos definen de tres a cinco niveles basados en la criticidad y sensibilidad de los datos para ayudar a determinar los controles de seguridad apropiados.
- Las organizaciones deben diseñar sus propios modelos y esquemas de clasificación de datos según su necesidad de proteger datos propietarios, empresariales y/o de usuarios con diferentes niveles de sensibilidad y para cumplir con los requisitos de cumplimiento y regulación. Sin embargo, pueden comenzar con o basar sus modelos en diferentes modelos y esquemas de clasificación desarrollados por gobiernos y organizaciones comerciales.
- Por ejemplo, el gobierno de los EE.UU. utiliza un esquema de clasificación de tres niveles para datos basado en el impacto potencial en la seguridad nacional si se divulga:
  - **Confidencial**: La divulgación no autorizada de esta información probablemente causaría daño a la seguridad nacional.
  - **Secreto**: La divulgación no autorizada de esta información probablemente causaría un daño grave a la seguridad nacional.
  - **Altamente secreto**: La divulgación no autorizada de esta información probablemente causaría un daño excepcionalmente grave a la seguridad nacional.

Modelos y esquemas de clasificación de datos

- NIST (Instituto Nacional de Estándares y Tecnología) desarrolló un esquema de categorización de tres niveles basado en el impacto potencial no solo en la confidencialidad, sino también en la integridad y disponibilidad de la información y los sistemas de información aplicables a la misión de una organización:
  - **Bajo**: La divulgación no autorizada de esta información tendría un efecto adverso limitado en las operaciones de la organización, los activos de la organización o los individuos.
  - **Moderar**: La divulgación no autorizada de esta información tendría un efecto adverso serio en las operaciones de la organización, los activos de la organización o los individuos.
  - **Alto**: La divulgación no autorizada de esta información tendría un efecto adverso severo o catastrófico en las operaciones de la organización, los activos de la organización o los individuos.
- Las organizaciones pueden usar etiquetas secundarias dentro de estos niveles para especificar diferentes activos de datos y procedimientos de manejo o requisitos de cumplimiento y regulación. Por ejemplo, una organización que solo recopila registros financieros puede clasificarlos como “datos confidenciales”, pero una organización que recopila registros médicos puede clasificarlos de manera más específica como “información de salud protegida” para indicar que los requisitos de HIPAA se aplican a esos datos.

Ejemplos de clasificación de datos

- Aunque el esquema de clasificación de datos del NIST es ampliamente reconocido como un esquema de clasificación adecuado en certificaciones sectoriales, nacionales e internacionales, las organizaciones deben desarrollar sus propios esquemas de clasificación según sus necesidades únicas de gestión organizacional y de riesgos.
- Para inspirarnos, veremos algunos ejemplos de organizaciones y el modelo de clasificación y esquema que han implementado.

UW-Madison

- UW-Madison clasifica los datos en cuatro categorías, que se utilizan para determinar cómo proporcionar acceso a los datos a las personas. Las categorías son:
  - **Público**: La divulgación, alteración o destrucción no autorizada de estos datos resultaría en poco o ningún riesgo para la Universidad y sus afiliados. Cualquier dato mostrado en sitios web o publicado sin restricciones de acceso debe clasificarse como público.
  - **Interno**: La divulgación, alteración o destrucción no autorizada de estos datos podría resultar en algún riesgo para la Universidad y sus afiliados. Por defecto, cualquier dato que no esté clasificado explícitamente en las otras tres categorías debe ser clasificado como interno.
  - **Sensible**: La divulgación, alteración, pérdida o destrucción no autorizada de estos datos podría causar un nivel moderado de riesgo para la Universidad, sus afiliados o proyectos de investigación.
  - **Restringido**: La divulgación, alteración, pérdida o destrucción no autorizada de esos datos podría causar un nivel significativo de riesgo para la Universidad, sus afiliados o proyectos de investigación. Si la protección de los datos es requerida por ley o regulación o si UW-Madison está obligada a informar al gobierno y/o notificar al individuo en caso de acceso inapropiado a los datos, entonces deben clasificarse como restringidos.

Ejemplos de clasificación de datos

Harvard

- Harvard clasifica los datos en cinco niveles:
  - **N1**: N1 se refiere a información pública. La Universidad proporciona intencionadamente esta información al público. Investigaciones publicadas, catálogos de cursos, hallazgos regulatorios y legales, informes anuales publicados, patentes liberadas y políticas universitarias son todos ejemplos.
  - **N2**: N2 se refiere a información confidencial de bajo riesgo. La Universidad elige mantener esta información privada dentro de la comunidad de Harvard, pero su divulgación fuera de la comunidad no causaría daño material. Políticas y procedimientos del departamento, materiales de formación de Harvard, borradores de trabajos de investigación y solicitudes de patentes y subvenciones son todos ejemplos.
  - **N3**: N3 se refiere a información confidencial de riesgo medio. La Universidad intenta compartir esta información solo con aquellos que tienen una "necesidad comercial de saber" y la divulgación más allá de los destinatarios previstos podría causar un daño material a las personas o a la Universidad. Información no directorio de estudiantes, información no publicada de profesores y personal, información de transacciones presupuestarias/financieras y la información especificada como confidencial por contratos de proveedores y acuerdos de confidencialidad son todos ejemplos.
  - **N4**: N4 se refiere a información confidencial de alto riesgo. La Universidad tiene estrictos controles para esta información y la divulgación más allá de los destinatarios especificados probablemente causaría un daño grave a las personas o a la Universidad. Contraseñas y PINs, credenciales del sistema y claves de cifrado privadas son todos ejemplos.
  - **N5**: N5 está reservado solo para datos de investigación, según lo determinado por el IRB o el Acuerdo de Uso de Datos. Los datos que, si se divulgaran, podrían poner al sujeto en un grave riesgo de daño o los datos con requisitos contractuales para medidas de seguridad excepcionales deben clasificarse como N5.

Ejemplos de clasificación de datos

AWS

- AWS recomienda comenzar con un enfoque de clasificación de datos de tres niveles. Tanto las organizaciones públicas como comerciales que han adoptado la nube de AWS han podido satisfacer adecuadamente sus necesidades y requisitos de clasificación de datos utilizando el enfoque a continuación.

Data classification tier	System security categorization	Cloud deployment model options
Unclassified	Low to High	Accredited public cloud
Official	Moderate to High	Accredited public cloud
Secret and above	Moderate to Hig	Accredited private/hybrid/community cloud/public cloud

Procesamiento Básico de Datos con Python

- Los requisitos de programación en la ciencia de datos exigen un lenguaje muy versátil pero flexible que sea simple para escribir el código pero que pueda manejar un procesamiento matemático altamente complejo. Python es más adecuado para estos requisitos, ya que ya se ha establecido como un lenguaje para la informática general y científica.
- Además, se está actualizando continuamente en forma de una nueva adición a su gran cantidad de bibliotecas destinadas a diferentes requisitos de programación.
- Los siguientes, son fragmentos de Python que pueden ser útiles para principiantes para algunas tareas de procesamiento de datos diferentes.
- Las tareas del procesamiento de datos van desde procesamiento de texto y listas básico a procesamiento de conjunto de datos con Pandas.

Concatenar varios archivos de texto

- Comencemos con la concatenación de varios archivos de texto. Si tienes varios archivos de texto en un solo directorio que necesitas concatenar en un solo archivo, este código Python lo hará.
- Primero obtenemos una lista de todos los archivos txt en la ruta; luego leemos en cada archivo y escribimos su contenido en el nuevo archivo de salida; finalmente, volvemos a leer el nuevo archivo e imprimimos su contenido en la pantalla para verificarlo.

```
import glob
# Cargar todos los archivos txt en la ruta
files = glob.glob ('/ path / to / files / *.
txt')
# Concatenar archivos a un nuevo archivo
with open('2020_output.txt', 'w') as out_file:
    for file_name in files:
        with open(file_name) as in_file:
            out_file.write(in_file.read())
# Leer archivo e imprimir
with open('2020_output.txt', 'r') as new_file:
    lines = [line.strip() for line in new_file]
for line in lines:
    print(line)
```

Concatenar varios archivos CSV en un marco de datos

- Siguiendo con el tema de la concatenación de archivos, esta vez abordemos la concatenación de varios archivos de valores separados por comas en un solo marco de datos de Pandas.
- Primero obtenemos una lista de los archivos CSV en nuestra ruta; luego, para cada archivo en la ruta, leemos el contenido en su propio marco de datos; luego, combinamos todos los marcos de datos en un solo marco; finalmente, imprimimos los resultados para inspeccionar.

```
import glob
# Cargar todos los archivos csv en la ruta
files = glob.glob ('/ path/to/files/ *. csv')
# Crear una lista de dataframe, una serie por
CSV
fruit_list = []
for file_name in files:
    df = pd.read_csv (file_name, index_col =
None, header = None)
    fruit_list.append (df)
# Crear un marco combinado a partir de la lista
de marcos individuales
fruit_frame = pd.concat (fruit_list, axis = 0,
ignore_index = True)
print (fruit_frame)
```

Aplanar listas

- Quizás tengas una situación en la que estés trabajando con una lista de listas, es decir, una lista en la que todos sus elementos también sean listas. Este fragmento tomará esta lista de listas incrustadas y la acoplará a una lista lineal.
- Primero crearemos una lista de listas para usar en nuestro ejemplo; luego usaremos listas por comprensión para aplanar la lista de una manera Pythonic; finalmente, imprimimos la lista resultante en la pantalla para su verificación.

```
# Creación de lista de listas (una lista donde todos sus elementos son
listas)
list_of_lists = [['manzana', 'pera', 'plátano', 'uvas'], ['cebra',
'burro', 'elefante', 'vaca'],['vainilla', 'chocolate'], ['princesa',
'Príncipe']]
# Aplanar la lista de listas en una sola lista
flat_list = [element for sub_list in list_of_lists for element in
sub_list]
# Imprime ambos para comparar
print(f'List of lists:\n{list_of_lists}')
print(f'Flattened list:\n{flat_list}')
```

Ordenar lista de tuplas

- Este fragmento considerará la idea de ordenar tuplas según el elemento especificado. Las tuplas son una estructura de datos de Python que a menudo se pasa por alto y son una excelente manera de almacenar datos relacionados sin usar un tipo de estructura más complejo.
- En este ejemplo, primero crearemos una lista de tuplas de tamaño 2 y las llenaremos con datos numéricos; a continuación clasificaremos los pares, por separado por el primer y segundo elemento, imprimiendo los resultados de ambos procesos de clasificación para inspeccionar los resultados; finalmente, ampliaremos esta clasificación a elementos de datos alfanuméricos mixtos.

```
# Algunos datos emparejados
pares = [(1, 10.5), (5, 7.), (2, 12.7), (3, 9.2), (7,
11.6)]
# Ordenar pares por primera entrada
sorted_pairs = sorted (pares, clave = lambda x: x [0])
print (f'Ordenado por elemento 0 (primer elemento): \
n {sorted_pairs} ')
# Ordenar pares por segunda entrada
sorted_pairs = sorted (pares, clave = lambda x: x [1])
print (f'Ordenado por elemento 1 (segundo elemento): \
n {sorted_pairs} ')
# Extiende esto a tuplas de tamaño n y entradas no
numéricas
pares = [('banana', 3), ('manzana', 11), ('pera', 1),
('sandía', 4), ('fresa', 2), ('kiwi', 12) ]
sorted_pairs = sorted (pares, clave = lambda x: x [0])
print (f'Apares alfanuméricos ordenados por elemento 0
(primer elemento): \ n {sorted_pairs} ')
```

Procesamiento de datos con Pandas

- Al cargar nuestros datos, podemos ver una serie de tipos de características únicas. Tenemos características categóricas como “Employee_Name” y “Position”. Tenemos funciones binarias como “MarriedID”. Tenemos características continuas como “PayRate” y “EmpSatisfaction”.
- Tenemos funciones discretas como “DaysLateLast30” y finalmente tenemos funciones de fecha como “LastPerformanceReview_Date”.

```
import numpy as np
import pandas as pd
import datetime
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.pipeline import make_pipeline
from feature_engine import
missing_data_imputers as mdi
from feature_engine import
categorical_encoders as ce
from sklearn.model_selection import
train_test_split
%matplotlib inline
with open('HRDataset.csv') as f:
    df = pd.read_csv(f)
f.close()
df.head()
df.info()
```

Variabilidad de características alta o baja

- El primer paso que suelo dar es revisar el recuento único de valores por característica para determinar si alguna característica se puede eliminar rápidamente debido a una variabilidad muy alta o muy baja.
- En otras palabras, ¿tenemos características que tengan tantos valores únicos como la longitud del conjunto de datos o características que tengan un solo valor único?

```
for col in df.columns:
  print(col, df[col].nunique(), len(df))
```

- Podemos eliminar con seguridad “Employee_Name”, “Emp_ID”, “DOB” ya que la mayoría, si no todos, los valores son únicos para cada función. Además, podemos eliminar “DaysLateLast30” ya que esta característica solo contiene un valor único.

```
df.drop(['Employee_Name'], axis=1,
inplace=True)
df.drop(['EmpID'], axis=1, inplace=True)
df.drop(['DOB'], axis=1, inplace=True)
df.drop(['DaysLateLast30'], axis=1,
inplace=True)
```

Funciones duplicadas

- A continuación, al examinar el libro de códigos, que contiene las definiciones de cada función, podemos ver que tenemos muchas funciones duplicadas.
- Por ejemplo, “MarriedStatusID” es una función numérica que produce el código que coincide con los estatutos de casados en la función “MaritalDesc”. Podemos eliminar estas características.

```
df.drop(['MaritalStatusID', 'EmpStatusID',
'DeptID'], axis=1, inplace=True)
df.drop(['GenderID'], axis=1, inplace=True)
df.drop(['PerformanceScore'], axis=1,
inplace=True)
df.drop(['MarriedID'], axis=1,
inplace=True)
```

UNIVERSIDAD
CATÓLICA
SEDES SAPIENTIAE