

Estándar de competencias profesionales

Procesar los datos para su uso en sistemas de Inteligencia Artificial basados en aprendizaje automático

Familia Profesional	Inteligencia Artificial y Data
Nivel	3
Código	ECP2492_3
Estado	BOE
Publicación	RD 46/2022

Competencia profesional

Elementos de la competencia

EC1 Extraer datos de las fuentes, previa identificación y clasificación de las mismas, estableciendo el filtrado inicial y configurando las pasarelas que permitan su acceso para su posterior carga y análisis.

IC1.1 Las fuentes de datos se conectan, configurando las pasarelas que permiten su acceso y teniendo en cuenta:

- Su origen: sistemas de almacenamiento, sistemas IoT, plataformas de datos en 'streaming', integración con APIs u otro,
- Su naturaleza: estructurados, tales como datos cuantitativos u otros o no estructurados,
- Si se trata de datos formales, tales como los provenientes de fuentes jurídicas o bien datos no formales tales como audios, imágenes o textos provenientes de fuentes como las redes sociales.

IC1.2 Los filtros y consultas para la extracción de datos desde las fuentes se establecen, previa selección, configurándolos, considerando la posibilidad de distribución de datos y de procesamiento cercano a las fuentes para maximizar tiempos de proceso, conforme a las especificaciones de análisis y según la naturaleza de los datos.

IC1.3 Los datos se extraen desde sus orígenes, conforme a las especificaciones de análisis y según la naturaleza de los datos.

EC2 Determinar la estructura y relaciones correspondientes al dominio de aplicación en los datos disponibles, mediante exploración para el análisis según objetivos de negocio.

IC2.1 Los datos recogidos de fuentes diversas se almacenan para componer conjuntos de datos, identificando su naturaleza e integrándolos en el destino.

IC2.2 Los conjuntos de datos recogidos se exploran por separado, identificando su estructura y relaciones para el análisis de variables por parte de la persona responsable.

IC2.3 La descripción de cada conjunto de datos se documenta, incluyendo entidades, volumetrías, relaciones y descripción de atributos, entre otros, para trabajar en las especificaciones de análisis de los datos, según los objetivos de negocio.

EC3 Verificar los datos mediante técnicas estadísticas, escribiendo código para explorar sus atributos con el fin de evaluar la cobertura y calidad del conjunto, de modo que permita la selección de los atributos clave para el entrenamiento de modelos, siguiendo las especificaciones de análisis.

IC3.1 La calidad de los datos se valora, aplicando técnicas estadísticas, conforme a las especificaciones de análisis vinculadas al cumplimiento de las necesidades de negocio.

IC3.2 El conjunto de datos se evalúa, mediante estadísticas sobre cobertura y sesgo tales como frecuencia y distribución, valorando el cumplimiento de los objetivos de negocio conforme a las especificaciones de análisis.

IC3.3 La descripción del resultado de las verificaciones sobre el conjunto de datos se documenta, indicando cobertura, frecuencia y sesgo entre otros, para que se tomen decisiones de diseño sobre la selección y uso del mismo en el modelado.

EC4 Seleccionar los datos mediante la depuración, limpieza, construcción y posterior clasificación en conjuntos, para su uso posterior en el modelado analítico, según las especificaciones del diseño.

IC4.1 El conjunto de datos y sus atributos se seleccionan, atendiendo a los criterios y especificaciones del diseño, como resultado de las verificaciones realizadas.

IC4.2 El conjunto de datos de entrenamiento se limpia, depurándolo para su utilización en el modelado, normalizándolo, categorizándolo o imputando valores ausentes, entre otros.

IC4.3 El conjunto de datos depurado se construye, atendiendo a las especificaciones de diseño de los modelos a implementar y según las técnicas de Inteligencia Artificial en las que se basen dichos modelos, añadiendo variables sintéticas fruto de cálculos agregados tales como:

- Normalización de atributos a media cero, desviación típica uno,

- Reescalado de atributos dentro de un rango predefinido,
- Media móvil calculada con respecto a una ventana deslizante temporal.

IC4.4 El conjunto de datos depurado se divide en particiones, según indique el científico de datos, habitualmente entrenamiento, validación y test, para entrenar y evaluar el modelo.

Contexto profesional

Ámbito profesional

Sectores productivos

Ocupaciones y puestos de trabajo relevantes

Medios de producción

Equipos y sistemas cliente y servidor. Sistemas operativos, firmware y software de base. Soporte hardware para almacenamiento de información y/o almacenamiento de la misma en la nube. Software y herramientas de exploración y análisis estadístico de datos, lenguajes de programación específicos. Entornos de desarrollo y programación.

Información utilizada o generada

Normas externas de trabajo: (normativa aplicable de protección de datos y propiedad intelectual e industrial) Normas internas de trabajo: (modelo o plan de negocio, proyecto software, diseño y especificaciones de los datos utilizados, modelos de datos contenidos en los sistemas gestores de datos, normativa interna para elaboración de documentación) Documentación técnica: (manuales de las herramientas de análisis estadístico, desarrollo y consulta utilizadas, documentación de cursos de formación, soportes técnicos de asistencia).