

Estándar de competencias profesionales

Entrenar modelos en sistemas de Inteligencia Artificial basados en aprendizaje automático

Familia Profesional	Inteligencia Artificial y Data
Nivel	3
Código	ECP2493_3
Estado	BOE
Publicación	RD 46/2022

Competencia profesional

Elementos de la competencia

- EC1** Buscar correlaciones entre las variables utilizando herramientas que incorporan técnicas de la estadística y el aprendizaje automático, con anterioridad al diseño y entrenamiento de modelos y siguiendo las indicaciones proporcionadas por el científico de datos responsable.
- IC1.1** Las correlaciones entre variables, tanto lineales como no lineales se buscan, aplicando técnicas de la estadística y del aprendizaje automático ('machine learning'- ML-).
- IC1.2** Los procedimientos de categorización y codificación de las variables tales como mapas de bits ('one-hot vectors') u otros se aplican mediante software o herramientas.
- IC1.3** Las transformaciones matemáticas tales como logaritmos, exponenciales, funciones trigonométricas u otras se aplican antes del entrenamiento de los modelos, teniendo en cuenta la distribución de sus valores dentro del conjunto de datos.
- IC1.4** Las variables de entrada más explicativas se seleccionan para tenerlas en cuenta durante el diseño de los modelos, considerando aquellas que tienen mayor correlación con las variables de salida o variables objetivo.
- IC1.5** El documento que describe las variables se redacta, incluyendo todos los detalles analizados en relación a las variables tanto de entrada como de salida y a las correlaciones encontradas,

con el fin de servir para tomar decisiones durante el diseño de los modelos y para ser integrado en un informe final.

EC2 Reducir la dimensión de las muestras de los conjuntos de datos, mediante programación o herramientas software, para obtener una representación de los mismos mediante variables latentes siguiendo instrucciones de la persona responsable.

IC2.1 Las técnicas de reducción de la dimensión, tales como análisis de componentes principales - PCA-, t-SNE, autocodificadores ('autoencoders') basados en redes neuronales y otras, disponibles como funcionalidades en herramientas software existentes se utilizan, eligiendo aquellas que realizan una mayor reducción de la dimensión mientras conservan la mayor variabilidad (capacidad explicativa) de las muestras.

IC2.2 Los métodos y procedimientos se programan, usando el lenguaje definido para el proyecto, realizando la reducción de la dimensión de las muestras con objeto de crear una copia del conjunto de datos con las transformaciones ya aplicadas, o bien aplicando las transformaciones durante el entrenamiento, antes de que las muestras se utilicen como entrada a los modelos para su entrenamiento y evaluación.

IC2.3 El trabajo realizado se documenta, indicando:

- Las técnicas de reducción de la dimensión aplicadas al conjunto de datos con el que se esté trabajando,
- Los valores concretos de los parámetros de configuración de cada técnica,
- Los resultados obtenidos en cuanto a reducción de la dimensión y pérdida de explicación de la variabilidad,

IC2.3 enfatizando el balance entre la proporción de reducción y la pérdida de variabilidad en función de uno o más parámetros.

EC3 Visualizar los datos, representándolos gráficamente, con la finalidad de corroborar las correlaciones encontradas, verificando la reducción de la dimensión aplicada sobre el conjunto de datos con los que se esté trabajando.

IC3.1 Las gráficas se generan, escribiendo programas al efecto, utilizando varias representaciones, con el fin de verificar tanto las correlaciones encontradas como las transformaciones aplicadas para reducir la dimensión de las muestras, bajo la supervisión de la persona responsable.

IC3.2 La información recogida mediante gráficas se elabora, usando medios didácticos tales como diapositivas o infografías, organizándola para presentarla al equipo de trabajo, con el fin de sacar conclusiones de manera conjunta y así decidir colectivamente qué subconjunto de

todas las transformaciones probadas se utilizarán para probar su efecto con los distintos modelos a entrenar y evaluar.

IC3.3 Las gráficas se recogen en un informe, presentándolas en formatos tales como histogramas, mapas de dispersión y otras técnicas de representación, explicando cómo se han obtenido y aportando una valoración de cada gráfica, destacando aspectos cualitativos como la correlación entre las variables latentes fruto de las transformaciones y/o reducciones y las variables salida.

EC4 Diseñar modelos basados en aprendizaje automático ('machine learning' -ML-) para aplicarlos sobre el conjunto de datos con el objeto de abordar el problema planteado según su tipo -de regresión o de clasificación-, siguiendo las indicaciones proporcionadas por el científico de datos responsable.

IC4.1 Las técnicas de aprendizaje automático tales como redes neuronales, máquinas de soporte vectorial, modelos de mixturas de gaussianas, árboles de decisión u otras se eligen para diseñar los modelos en base al análisis exploratorio de los datos y el estudio visual realizados previamente.

IC4.2 Los modelos diseñados se ensayan, programando código para entrenarlos, realizando experimentos que permitan ver qué combinaciones de parámetros de configuración conducen a mejores resultados, tales como:

- El coeficiente de aprendizaje para las redes neuronales ('learning rate'), el número de capas, el número de neuronas en cada capa, los tipos de activación a utilizar en las capas ocultas,
- El número de grupos ('clusters') para el algoritmo 'k means', y
- El número de árboles a generar en el caso de 'random forests'.

IC4.3 Las métricas a utilizar para evaluar modelos tales como:

- La desviación porcentual entre valores predichos y reales en problemas de regresión como la predicción del volumen de ventas semanal por producto,
- La ratio de falsos positivos y la de falsos negativos para un problema de clasificación,
- Otras métricas relacionadas con las ratios anteriores como 'accuracy', 'precision', 'recall' o 'sensitivity', 'specificity', y área bajo la curva ROC ('Receiver Operating Characteristic'),

IC4.3 se seleccionan, teniendo en cuenta las técnicas a aplicar y el problema a resolver.

IC4.4 Los modelos diseñados en base a las métricas seleccionadas se evalúan, con el fin de decidir, bien la creación de nuevos modelos desde cero, bien el rediseño de modelos anteriores, en función de que los resultados obtenidos se ajusten a los esperados, siguiendo indicaciones de la persona responsable.

- IC4.5** Los comentarios en el código escrito se incluyen, detallando el papel que juega una variable, describiendo las funciones y algoritmos con el fin de ser reutilizado en el desarrollo de soluciones a otros problemas similares y siguiendo las normas establecidas por la organización.
- IC4.6** La documentación se genera, detallando los modelos diseñados con todas sus variantes, con el fin de que sirva como punto de partida para el informe que incluirá los resultados de evaluar el rendimiento de cada uno de los modelos.
- EC5** Evaluar los modelos diseñados, definiendo un subconjunto de test, cuando el conjunto de datos no lo tuviera previamente definido, y creando nuevas particiones de validación y entrenamiento una vez separado el subconjunto de test.
- IC5.1** El conjunto de datos se divide en dos subconjuntos, entrenamiento y test, si no estaban previamente definidos, usando las muestras del subconjunto de test solamente para evaluar el modelo y respetando que el subconjunto de test represente un porcentaje del conjunto de datos completo entre el 10% y el 30%.
- IC5.2** El subconjunto de entrenamiento se vuelve a dividir en dos nuevos subconjuntos, entrenamiento propiamente dicho y validación, escogiendo las muestras de manera aleatoria, pero conservando la misma distribución de clases.
- IC5.3** La validación cruzada de los modelos se lleva a cabo, generando distintas particiones de entrenamiento y validación, donde el número de distintas particiones lo indicará la persona responsable, repitiendo el proceso completo de entrenamiento y evaluación del modelo por cada una de las particiones, decidiendo en qué iteración del entrenamiento se guarda la versión del modelo para ser evaluado con el subconjunto de test.
- IC5.4** Cada modelo entrenado y validado con cada partición se evalúa con el subconjunto de test, mostrando la media aritmética, la varianza u otra medida de las métricas elegidas para evaluar los modelos a partir de las particiones creadas para la validación cruzada.
- IC5.5** El informe de la evaluación se crea, tomando como punto de partida el informe con todos los detalles de los modelos diseñados e incluyendo gráficas y tablas que muestren las métricas elegidas, para evaluar los modelos y su evolución durante el entrenamiento con respecto a los subconjuntos de validación y que muestren los resultados de los distintos modelos diseñados y sus variantes, siempre con respecto al mismo subconjunto de test con el fin de poder comparar rápidamente el rendimiento y la robustez de cada modelo y sus variantes.

Contexto profesional

Ámbito profesional

Sectores productivos

Ocupaciones y puestos de trabajo relevantes

Medios de producción

Lenguajes de programación, herramientas software y plataformas de desarrollo de Inteligencia Artificial para el tratamiento, análisis, visualización y transformación de los datos; Implementaciones de técnicas de aprendizaje automático.

Información utilizada o generada

Normas externas de trabajo (normativa aplicable de protección de datos, propiedad intelectual e industrial). Normas internas de trabajo (planificación del trabajo en base al proyecto, plan de seguridad) Documentación técnica (Guía de cómo se ha utilizado el sistema de Inteligencia Artificial elegido para realizar los experimentos, referencias a artículos de investigación y libros relacionados con el área del aprendizaje automático, manuales de las herramientas software utilizadas para entrenar los modelos, documentación de requisitos hardware tanto de la plataforma como de las herramientas software o 'toolkits'.