Universidad Intercontinental de la Empresa

Proyecto

Aprendizaje Automático

Borja González Seoane

S03, 18 de septiembre de 2024

1. Descripción general de la actividad

Tal y como consta en el Contrato de Ensenañza-Aprendizaje, la AE03, con un peso del 25 % en la calificación final, consiste en la realización de un proyecto en el que se pretende que se apliquen la totalidad los conocimientos adquiridos en la asignatura. Esta actividad permitirá a los estudiantes trabajar en un proyecto de aprendizaje automático de principio a fin, desde la obtención de los datos hasta la presentación de los resultados, constituyendo de esta forma una experiencia de aprendizaje holística que integrará todas las unidades del currículo del curso.

Del 25 % de la calificación, el 20 % se corresponde con la entrega final y su exposición en la última sesión del curso, S15, mientras que el 5 % restante se podrá sumar en dos puntos de control (2,5 % cada uno) distribuidos a lo largo de curso, tal y como se detalla en la planificación de la Sección 3 del Contrato de Enseñanza-Aprendizaje. En estos puntos de control se comprobará si se está avanzando correctamente en el proyecto y se resolverán dudas que puedan surgir.

La tarea se llevará a cabo en equipos de trabajo de 3-4 estudiantes. Todos los equipos deberán integrar a estudiantes de ambos campus. Los equipos deberán registrarse en una wiki dispuesta para dicho fin en el Campus Virtual. Se tendrán en consideración los principios, política de la UIE, para una evaluación justa y equitativa en las actividades grupales.

Para la realización del proyecto se proporcionará un conjunto de datos en el Campus Virtual de la asignatura, con el que trabajarán todos los grupos. Este conjunto de datos permitirá desarrollar diferentes líneas de trabajo de aprendizaje automático; a saber, clasificación, regresión y clusterización. Cada equipo deberá trabajar en las tres líneas de trabajo, entrenar un modelo que resuelva cada uno de ellos y presentar los resultados obtenidos en una exposición final.

2. Detalle de los objetivos de la actividad

Los objetivos de la actividad, que deberán perseguir todos los equipos, son los siguientes:

1. Analizar el conjunto de datos proporcionado, realizando un EDA exhaustivo para comprender la estructura de los datos y las relaciones entre las variables.

2. Preprocesar los datos, realizando las transformaciones necesarias para que los modelos de aprendizaje automático puedan ser entrenados. Podrían ser necesarios distintos preprocesamientos para cada una de las líneas de trabajo a modelar.

3. Para cada una de las líneas de trabajo planteadas con respecto al conjunto de datos, entrenar un modelo y evaluar su rendimiento. Se valorará el probar diferentes arquitecturas de modelos para las tareas, así como diferentes estrategias de entrenamiento y validación, hiperparametrización, etc. Se trata de intentar emplear la mayor cantidad de técnicas aprendidas en la asignatura posible, siempre y cuando se justifique su uso racionalmente.

4. Presentar una exposición oral final, en la que se deberá explicar el proceso seguido, las decisiones tomadas y los resultados obtenidos. Se valorará la claridad y la capacidad de comunicación de los estudiantes, más allá de la calidad de los resultados o la solidez técnica del trabajo.

Para esta actividad se pretende que los equipos trabajen de forma autónoma, como si de un proyecto real se tratase. La profundidad de la experimentación, los distintos ensayos y errores acometidos durante el semestre, será especialmente valorada. Se insta pues a los estudiantes a que lleven un registro de todas las decisiones tomadas, los problemas encontrados y las soluciones propuestas, y que incorporen toda esta información en la presentación final.

Se espera que los estudiantes sean capaces de tomar decisiones y resolver problemas de forma autónoma, aunque siempre podrán contar con la ayuda del profesor para resolver dudas o problemas que puedan surgir.

3. Conjunto de datos y líneas de trabajo

El conjunto de datos proporcionado para la realización del proyecto está disponible en el Campus Virtual de la asignatura, en el archivo `proy_escuela_dev.csv` —*proyecto-escuela-partición de desarrollo (*development*)*—.

El contexto es el de un centro educativo que pretende mejorar la comprensión sobre su alumnado para poder ofrecer una atención más personalizada o accionar diferentes políticas educativas. Se trata de un centro que imparte educación secundaria y ciclos formativos de grado medio y superior. El responsable de datos de la institución ha trabajado en la recopilación del conjunto de datos proporcionado, integrando varias tablas de la base de datos centralizada y anonimizando algunas columnas para proteger la privacidad de los estudiantes.

El conjunto de datos contiene tanto características de los estudiantes como anotaciones acerca de su desempeño académico, además de información procedente de encuestas realizadas a los estudiantes. Es por ello que el responsable de datos ha optado por encriptar en la muestra algunas columnas que pudieran identificar a los estudiantes, a pesar de que probablemente algunas de ellas hubieran sido de interés para el análisis. Además, ha indicado que la clave de tabla dentro de la arquitectura de datos del centro es el DNI del estudiante y que el ID de la columna `id_estudiante_proy` se ha generado únicamente para el estudio encargado, no teniendo correspondencia con ningún identificador real.

Tras una primera toma de requisitos con la dirección del centro educativo, se han planteado tres líneas de trabajo para el proyecto:

1. **Clusterización:** el centro educativo quiere identificar perfiles de estudiantes para poder ofrecer una educación más personalizada. Así pues, el propósito sería hacer emerger algunos grupos de estudiantes con características similares. Estos clústeres deberán ser fáciles de explicar y de interpretar.

2. **Clasificación:** se pretende clasificar a los estudiantes en dos grupos, aquellos que van a obtener una beca y aquellos que no. La política del centro es ofrecer becas a aquellos estudiantes que tengan un rendimiento académico excepcional, en este caso aquellos que obtengan una nota final superior al percentil 90 de la distribución de notas.

3. **Regresión:** el objetivo sería predecir la nota final de los estudiantes basándose en la información de la que se disponga.

Como en un proyecto real de aprendizaje automático, es posible que surjan dudas acerca de los datos que sólo pueda resolver alguien con conocimiento experto en el dominio. Para simular este escenario, se habilitará un foro específico en el Campus Virtual: *Proyecto - Foro de aclaraciones sobre el conjunto de datos*. En este foro, los estudiantes podrán plantear cuestiones acerca del conjunto de datos que el profesor resolverá emulando ponerse en contacto con el responsable de datos del centro educativo.

Además, el susodicho foro también podría utilizarse para comunicar alguna novedad al respecto del uso del conjunto de datos por iniciativa del responsable de datos de la institución. Por ejemplo, se podría instar a los equipos de trabajo a dejar de utilizar algunas variables porque una auditoría interna habría detectado que no se deben emplear por motivos éticos o legales, especialmente de protección de datos de carácter personal.

Como comentarios generales, ha de tenerse en cuenta que como en cualquier base de datos del mundo real, pueden existir errores o inconsistencias, principalmente debidos a la recogida de datos. Por otra parte, aunque la institución haya disponibilizado una muestra de su base de datos, es responsabilidad del equipo de trabajo seleccionar de la misma las variables que considere más relevantes para cada una de las líneas de trabajo planteadas, que no necesariamente han de ser todas las disponibles o todas para las tres líneas de trabajo.

4. Entrega y exposición final

En la última sesión del semestre, S15, se realizará la exposición final de los proyectos. Cada equipo dispondrá de 20 minutos para presentar los resultados obtenidos, explicar el proceso seguido y las decisiones tomadas. Se valorará la claridad y la capacidad de comunicación de los estudiantes. Todos los estudiantes del equipo deberán participar en la exposición de forma equitativa.

Además de la propia exposición, se dispondrá en el Campus Virtual una tarea para depositar los siguientes entregables antes del inicio de la S15:

1. Un informe en formato PDF con una extensión máxima de 10 páginas, en el que se detallen los objetivos del proyecto, el proceso seguido, los resultados obtenidos y las decisiones tomadas.
2. Las transparencias de la presentación final en formato PDF.
3. Un archivo comprimido ZIP con el código fuente del proyecto, incluyendo los *scripts* o *notebooks* de preprocesamiento, entrenamiento y evaluación de los modelos, así como cualquier otro archivo necesario para la reproducción de los resultados.
4. Los artefactos con los pesos de cada uno de los modelos finales de cada una de las líneas de trabajo, en otro archivo comprimido ZIP. **El profesor empleará estos archivos de pesos, así como el código fuente, para evaluar los modelos sobre una partición de datos que no ha sido vista por los estudiantes.** En caso de que el código no funcione correctamente o de que no sea posible reproducir los resultados con los pesos suministrados, la calificación de la actividad se verá afectada.

5. Criterios de evaluación

La evaluación de la actividad se realizará teniendo en cuenta los siguientes aspectos, que se recomienda que los estudiantes tengan en cuenta a la hora de realizar el proyecto:

1. Avance satisfactorio del proyecto en el primer punto de control (2,5 %). Objetivos mínimos: se ha completado el análisis exploratorio de los datos y se ha implementado el código necesario para preprocesar los datos. Se tiene una versión preprocesada de los datos para comenzar a trabajar en el modelado de alguna de las líneas de trabajo.

2. Avance satisfactorio del proyecto en el segundo punto de control (2,5 %). Objetivos mínimos: se ha trabajado en el modelado de al menos una de las líneas de trabajo. Se dispone de una versión preliminar de un modelo entrenado que presenta un comportamiento razonable.

3. Calidad técnica general del proyecto (10 %), tras la entrega final. Se valorará la calidad técnica del proyecto, incluyendo la profundidad del EDA, la correcta implementación de las transformaciones de los datos, la elección de los modelos y la calidad de los resultados obtenidos. Transversalmente se valorará la limpieza y claridad del código fuente, así como la correcta organización del proyecto en *scripts* o *notebooks* fáciles de navegar y entender.

4. Calidad de la presentación final (5 %). Se valorará la claridad y la capacidad de comunicación de los estudiantes, más allá de la calidad de los resultados obtenidos o la solidez técnica del trabajo. Se valorará la capacidad de los estudiantes para justificar el proceso seguido y las decisiones tomadas.

5. **Fase competitiva (5 %). El profesor dispone de una partición de datos no vista por los estudiantes,** `proy_escuela_test.csv` **—*proyecto-escuela-partición de test*—, pero perteneciente al mismo conjunto de datos. Se empleará esta partición para evaluar los modelos finales de cada una de las líneas de trabajo de cada uno de los equipos. De esta forma, se obtendrán nuevas métricas con las que ponderar la capacidad de generalización de los modelos obtenidos. Se establecerá una ponderación de equipos en función de estas métricas y se escalará el 5 % en consonancia para premiar a los equipos con mejores resultados.**