Proyecto

1. Contexto

El desarrollo exitoso de soluciones de Machine Learning requiere una comprensión total del pipeline completo de esta disciplina, desde la exploración inicial de los datos hasta la implementación y evaluación de modelos predictivos. A diferencia de las prácticas que se centraban en aspectos específicos, este proyecto adopta una perspectiva holística que simula el flujo de trabajo real de un científico de datos. El énfasis está puesto en la resolución integral de problemas que presenta un dataset complejo, aplicando de manera cohesiva todas las técnicas estudiadas en la asignatura.

Este enfoque integral refleja la realidad profesional del Machine Learning, donde los proyectos raramente se limitan a la aplicación de una sola técnica. La capacidad de navegar fluidamente entre diferentes metodologías, identificar sinergias entre técnicas supervisadas y no supervisadas, y mantener una visión global del problema son habilidades fundamentales para el éxito profesional en el campo de la ciencia de datos. El proyecto no solo evalúa el dominio técnico de los algoritmos, sino también la capacidad de pensamiento crítico, resolución de problemas y síntesis metodológica que distingue a los profesionales competentes en Machine Learning, preparando para abordar desafíos reales desde la identificación de problemas de calidad de datos hasta la construcción de modelos predictivos robustos.

2. Formación de los grupos

- **Los grupos estarán formados por 3 personas** con la excepción de 2 grupos de 4. Estos grupos serán elegidos libremente por los estudiantes. Para decidir quiénes podrán formar los grupos de 4 personas, se realizará un sorteo el 8 de octubre en horario de clase.
- Un único miembro de cada equipo deberá mandar un email a alex.tascon@deusto.es comunicando los integrantes del grupo **antes del 9 de octubre a las 17:00**. Aquellos estudiantes que tras esta hora límite aún no hayan formado grupo serán unidos de forma aleatoria y se les enviará una notificación informando de sus equipos.

3. Formato de la entrega

Presentación intermedia opcional

- En primer lugar, habrá una **presentación intermedia opcional el 6 de noviembre**. Esta presentación sirve para que, a mitad del desarrollo del proyecto, los grupos puedan recibir feedback de su progreso y expectativas.
- Aquellos que quieran realizarla deberán registrarse en la Tarea habilitada para ello en ALUD **antes del 2 de noviembre a las 23:59**.
- El formato de esta presentación es libre: si bien deben salir a la pizarra todos los integrantes del grupo, pueden elegir que el soporte sea un PPT, el propio código o ninguno. Simplemente se debe aprovechar el slot de 10 minutos para tratar aquello que resulte más importante para cada grupo.

Entrega final

- La entrega constará de 2 partes:
  - Un **único fichero .ZIP** que recoja **todos los ficheros** utilizados en el proyecto: cuadernos .ipynb, scripts de Python, ficheros de datos…
  - Un **fichero PDF para la memoria del proyecto**.
- El fichero .ZIP deberá, como mínimo, contener todo el código y ficheros necesarios para la ejecución del proyecto. El código deberá permitir que los resultados sean reproducibles para verificar aquellos presentados en las celdas de *output* de los cuadernos, la memoria y la defensa oral.
- El fichero PDF de memoria del proyecto no tiene un mínimo ni máximo de páginas. Cada grupo escribirá lo que considere adecuado según el desarrollo de su proyecto. Se valorará un documento claro, exhaustivo y bien justificado. No necesariamente más páginas implicarán mejor nota.
- Ambas partes de esta entrega deberán ser subidas por uno de los integrantes del equipo a las Tareas correspondientes habilitadas en ALUD **antes del 4 de diciembre a las 15:00**.

Presentación

- Cada equipo deberá preparar y llevar a cabo una presentación de 15 minutos sobre su proyecto en la que cada integrante deberá hablar el mismo tiempo que el resto.
- Tras esto, habrá hasta 10 minutos adicionales para que, por un lado, el resto de la clase pueda realizar preguntas y comentarios al grupo que presenta; y por otro, el profesor pueda realizar, si así lo desea, entre 1 y 2 preguntas a cada integrante para evaluar su nivel de conocimiento del proyecto, con el impacto correspondiente en la nota individual.
- Es por esto último por lo que se recomienda que todos los integrantes conozcan de primera mano todas las partes del proyecto, aunque la manera de trabajar que hayan utilizado implique un reparto de tareas.
- Para la presentación se podrá utilizar como soporte un PPT que se habrá tenido que mandar por email a alex.tascon@deusto.es **antes del 9 de diciembre a las 23:59**.
- Estas presentaciones tendrán lugar los días **10 y 11 de diciembre. Antes del 4 de diciembre a las 23:59** los grupos habrán tenido que elegir su slot en la Tarea habilitada para ello en ALUD.

Aquellas entregas realizadas fuera de plazo recibirán un 0/10.

4. Instrucciones del proyecto

El objetivo principal del proyecto es resolver una o varias problemáticas presentadas por el o los conjuntos de datos seleccionados. Si bien se deben cumplir todos aquellos requisitos indicados a continuación, el foco debe estar en el desarrollo de un proyecto íntegro de Machine Learning, desde el aprovisionamiento de datos hasta el desarrollo de modelos que sirvan como solución.

Requisitos de tareas

- Se deben realizar ambas **tareas supervisadas**: **clasificación** y **regresión**.
- Se deben llevar a cabo al menos **2** análisis o **tareas no supervisadas**: **clustering**, **reducción de dimensionalidad, detección de anomalías**…

Recolección de datos

- Se debe escoger un dataset que contenga un mínimo de **5.000 registros y 10 características tanto numéricas como categóricas**.
- Puede provenir de **otras asignaturas** (si el proyecto es conjunto) o ser importado de alguna fuente de uso abierto en Internet. Se permite **combinar varios datasets** si es adecuado y correctamente justificado.
- Si este dataset principal no permite cumplir con todos los requisitos de tareas en el proceso de resolución de la problemática que presentan, el desarrollo puede dividirse. Se pueden usar **distintos datasets/distintos desafíos**. Ejemplo:
    - Dataset 1 sobre economía para regresión y detección de anomalías.
    - Dataset 2 sobre deporte para clasificación y reducción de dimensionalidad.

Metodología de trabajo

- El énfasis debe estar en **resolver el problema planteado** por el conjunto de datos.
- No se trata solamente de aplicar la técnica más moderna o eficaz. Se espera que el grupo realice un **preprocesamiento de datos adecuado**, **pruebe varias técnicas**, **compare su funcionamiento, analice los resultados, optimice modelos** y solo al final proponga una **solución final justificada**.

Entrega técnica

- La **estructura interna** del proyecto queda a elección del grupo. Se valorará positivamente un proyecto **organizado**, pero no hay estructura obligatoria.
- Se debe garantizar que el proyecto pueda ejecutarse de **forma determinista** (con semillas y dependencias fijadas), de forma que en cualquier momento se puedan reproducir los resultados.

*En caso de realizar el proyecto en conjunto con otra asignatura y que esto genere alguna problemática con respecto a estos requisitos e instrucciones, contactar con el profesor para estudiar una solución ad-hoc para cada caso. La fecha límite para esta discusión sería **el 6 de noviembre** durante las **presentaciones intermedias opcionales**, aunque se recomienda contactar lo antes posible una vez se conozca el conflicto.*

5. Evaluación

- El proyecto supone un **55% de la nota final** de la asignatura.
- La **nota máxima alcanzable** según la rúbrica es de **10,5 puntos**, siendo el 0,5 adicional un reconocimiento a aportaciones extra.
- No obstante, **la nota máxima del proyecto es de 10 puntos**. En caso de haber obtenido una nota superior según la rúbrica, la calificación final será de 10 sobre 10.

La evaluación se basará en la correcta implementación técnica, la profundidad de los razonamientos, la claridad de las conclusiones, el alineamiento de las soluciones al problema y la organización lógica del proyecto en su conjunto.

Criterio	Descripción	Peso
Alineamiento con la problemática	El proyecto está correctamente enfocado en resolver una o varias problemáticas planteadas por el/los dataset(s). Se evalúa la capacidad del grupo para contextualizar los datos, justificar las tareas realizadas en función de la problemática, y proponer una solución final coherente y bien motivada.	**12%**
Preprocesamiento de datos	Se evalúa la calidad de los análisis realizados, la limpieza de los datos, y las transformaciones aplicadas. Se valorará que estas decisiones estén bien justificadas y que contribuyan al buen desempeño posterior de los modelos.	**10%**
Aprendizaje supervisado	Se implementan y comparan distintas técnicas de regresión y clasificación. Se valorará la variedad de modelos probados, la justificación de su elección, el ajuste de hiperparámetros y la discusión sobre su rendimiento. La selección de la solución final debe estar fundamentada en los resultados obtenidos y en su adecuación al problema.	**23%**
Aprendizaje no supervisado	Se aplican al menos dos análisis distintos. Se valora la correcta elección de técnicas, la interpretación de resultados, la conexión con la problemática y la comparación entre distintos enfoques.	**18%**
Evaluación y análisis	Se emplean métricas adecuadas a cada tarea. Se realizan procesos de validación, análisis de overfitting/underfitting y curvas de aprendizaje. Se valora la capacidad para extraer conclusiones sólidas a partir de las evaluaciones y justificar la elección de los modelos finales.	**20%**
Calidad técnica y memoria escrita	El código es reproducible, está documentado y presenta bloques explicativos claros en los cuadernos. El proyecto se puede ejecutar de forma determinista gracias a la fijación de semillas y dependencias. La memoria en PDF expone con claridad el proceso seguido, las técnicas aplicadas, los resultados y las conclusiones, de forma exhaustiva pero sin redundancias innecesarias.	**12%**

Defensa oral	Presentación de 15 minutos en la que todos los integrantes participan de forma equilibrada. El grupo expone de manera clara los objetivos, el proceso y las conclusiones.	**5%**
Extra	*Se valorará el uso adecuado de técnicas, herramientas o reflexiones que vayan más allá de lo visto en clase.*	***+5%***

**Aquellas entregas realizadas fuera de plazo recibirán un 0/10.*

***Tras la presentación, el profesor podrá realizar 1-2 preguntas a cada integrante del equipo. Las respuestas a estas preguntas podrán modificar su nota individual en el proyecto de **-2 a +1 puntos**.*