2024/2025 1Q

Práctica

APA

Departament de Ciències de la Computació

Grau en Enginyeria Informàtica - UPC

Facultat d'Informàtica de Barcelona

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Índice general

CAPÍTULO 1

Información general

Esta es la **guía** para el correcto desarrollo de las prácticas de la asignatura APA ("Aprenentatge Automàtic"). El objetivo es aplicar los diferentes conceptos y técnicas impartidos durante el curso para resolver un problema real. Se deberá redactar un informe completo describiendo el trabajo realizado, los problemas encontrados y las soluciones previstas, así como los resultados finales y conclusiones del estudio.

El objetivo de la práctica es desarrollar un **modelo de clasificación o regresión** para resolver uno de los problemas que se pueden obtener de los repositorios de datos seleccionados (detallado en 2.2). También se puede optar por explorar cualquier problema que os motive e incluso traer vuestra propia propuesta.[1]

Se espera que se escriba un **informe** completo que describa el trabajo realizado, su motivación, los problemas encontrados, las soluciones aplicadas, y los resultados finales y conclusiones del estudio. El texto principal está *estrictamente* limitado a un máximo de 15 páginas (esto incluye gráficos, tablas y referencias; tened en cuenta que el código debe entregarse *por separado*).

No os limitéis a narrar los experimentos y copiar los resultados. El informe debe ser una explicación de lo que habéis hecho, una valoración de lo que cada modelo es capaz de hacer y una comparativa con sentido de los resultados. Pensad en el informe que presentarías a quien os ha encargado analizar los datos.

El principal lenguaje de programación utilizado para la parte de modelado es `python`. Recordad que hay *muchos* paquetes para `python` que probablemente contienen rutinas útiles que puede usar; solo aseguraos de mencionarlos en el documento final.

Se puede usar otro software siempre que sirva para un propósito específico o secundario.

Cualquier información adicional sobre los métodos o sobre los problemas debe ser **reconocida y/o debidamente citada**.

[1] Con aceptación explícita del profesor de laboratorio.

CAPÍTULO 2

Desarrollo

2.1. Procedimiento de los experimentos

Se recomienda encarecidamente que sigáis estos pasos en el desarrollo del trabajo práctico:

1. Elegid el problema y aseguraos de que podéis obtener los datos
2. Leed la documentación disponible sobre el problema y los datos; obtened y leed algunos trabajos previos relevantes sobre el mismo problema (o muy similar) y datos, si hay alguno
3. Preprocesad los datos y elegid las variables que vais a utilizar
4. Realizad una descripción estadística básica
5. Visualizad los datos
6. Elegid el método de remuestreo para ajustar, seleccionar y probar los modelos
7. Realizad un proceso de modelado completo, utilizando técnicas lineales/cuadráticas
8. Realizad un proceso de modelado completo, utilizando técnicas no lineales

2.2. Obtención de los problemas

Necesitáis decidir un problema que cumpla con todas las siguientes condiciones:

- El conjunto de datos tiene variables numéricas, categóricas o preferiblemente una combinación de ambas
- El conjunto de datos no se genera sintéticamente
- El conjunto de datos contiene más de 10 variables.

- Necesitáis tener suficiente información sobre el problema para poder entender y analizar tus resultados.
- No se aceptarán conjuntos de datos ya preprocesados. Es necesario un problema cuyos datos tengan algún trabajo de procesamiento que hacer
- El conjunto de datos contiene más de 500 muestras.
- El problema a resolver no es uno de los problemas simples usados habitualmente como ejemplos como iris, MNIST o wine o ya se encuentra en internet resuelto multitud de veces
- El problema es lo suficientemente complejo como para que no obtenga un acierto o un R^2 casi perfectos usando una regresión lineal/logística o naive bayes.

Podéis elegir cualquier conjunto de datos que cumpla las condiciones. Es recomendable que elijáis un problema de un campo que os guste y motive.

Si no sabéis dónde buscar problemas, podéis usar los siguientes repositorios:

1. El repositorio de aprendizaje automático de UCI: http://archive.ics.uci.edu/ml/
2. Kaggle https://www.kaggle.com/ (Solo podéis usar problemas que no tengan ya un notebook con su análisis y solución)
3. OpenML https://www.openml.org/
4. Data World https://data.world/
5. Cualquiera de los repositorios de datos cientificos del Dataverse (https://dataverse.org/), como por ejemplo el Repositori de dades de Recerca de Catalunya (https://dataverse.csuc.cat/) o el de la universidad de Harvard (https://dataverse.harvard.edu/)

También podéis usar el motor de búsqueda de conjuntos de datos de Google https://datasetsearch.research.google.com/ si estáis interesados en un tema específico.

Veréis que los problemas son muy diversos, no solo por el área de trabajo (biología, geofísica, medicina...) sino porque muestran características de datos diferentes. Por ejemplo, existen grandes diferencias en el número de variables y ejemplos, número de clases, dificultad intrínseca, valores faltantes, errores varios, variables mixtas nominales y/o continuas, etc. La mayoría de ellos son tareas del mundo real y muchos son bastante desafiantes, posiblemente superando el nivel de APA, especialmente en los recursos computacionales necesarios para resolverlos. En particular, evitad problemas que usen imágenes, a menos que el tamaño de la imagen sea muy pequeño.

Algunos problemas son más fáciles en algunos aspectos y más difíciles en otros. Por lo tanto, la selección del problema en particular no tiene mucha importancia para la nota. En concreto, no es nada recomendable que os pongáis a probar problemas para ver cómo se "comportan". Es recomendable que baséis la decisión en el interés que os despierte, aunque debéis evitad conjuntos de datos muy pequeños que sean demasiado fáciles (cualquier método predice con una precisión casi perfecta).

Evaluación preliminar del problema

Para evitar tener un problema demasiado fácil o demasiado difícil para el proyecto, una vez elegido el conjunto de datos debéis enviar un documento con un análisis básico del conjunto de datos elegido. Este documento debe tener:

- Una pequeña descripción del conjunto de datos y su fuente. Debe incluir cuál será su variable objetivo y si resolverá un problema de regresión o clasificación. También se deben comentar problemas específicos de los datos, como atributos relacionados con el tiempo y desbalance entre clases.
- Una simple visualización de las variables comparándolas con el objetivo.
- La matriz de correlación de las variables.

También deberéis agregar resultados de referencia con un modelo simple (regresión lineal/logística/naive bayes) aplicando un preproceso sencillo de los datos. Esto puede ayudaros a encontrar si vuestro conjunto de datos es demasiado fácil.

Debéis decidir qué **problema** queréis abordar lo antes posible y comunicar vuestra elección (por correo electrónico) enviando la evaluación preliminar del problema a vuestro profesor de laboratorio (`bejar@cs.upc.es`, `david.garcia.soriano@upc.edu`) no más tarde del **28 de Octubre**. Indicad también **todos** los nombres de los miembros del grupo.

2.3. Sobre el preprocesamiento de datos

Cada problema requiere un enfoque diferente con respecto a la limpieza y preparación de los datos, y la selección de la información particular que se va a utilizar puede variar; este proceso previo es muy importante porque puede tener un impacto profundo en el desempeño futuro; fácilmente puede llevar una parte significativa del tiempo de análisis. Por lo tanto se recomienda que se analicen bien los datos antes de hacer nada, a fin de evaluar la mejor manera de preprocesarlos[1]. En particular, se deberá prestar atención a los siguientes aspectos (no necesariamente en este orden):

1. Tratamiento de valores perdidos (missing values)
2. Tratamiento de valores anómalos (outliers)
3. Tratamiento de valores incoherentes o incorrectos
4. Codificación de variables no continuas o no ordenadas (nominales o binarias)
5. Posible eliminación de variables irrelevantes o redundantes (selección de características)
6. Creación de nuevas variables que puedan ser útiles (extracción de características)
7. Normalización de la variables (*e.g.* estandarización)
8. Transformación de las variables (*e.g.* corrección de asimetrías graves y/o curtosis en los valores de los datos)

[1] Ver el Laboratorio 1.

2.4. Sobre la selección de modelos y estimación del rendimiento

De acuerdo con el problema y los datos disponibles, se debe diseñar un conjunto de experimentos basados en protocolos válidos para seleccionar modelos y estimar honestamente el error de generalización (o cualquier otra medida de desempeño futuro) del modelo o solución propuesta.

Algunos problemas vienen con sus propios datos de prueba (datos usados para la estimación del error de generalización), otros no; en el último caso, debe obtener datos de prueba dividiendo todos los datos disponibles (una o varias veces, según el tamaño de los datos). Para la selección del modelo, probablemente será necesaria una validación cruzada de k particiones (la selección del mejor valor para k es vuestra decisión). Está metodológicamente prohibido utilizar como datos de test cualquier conjunto de datos que ya se haya utilizado para la creación, ajuste o selección del modelo.

2.5. Sobre la interpretación de modelos/ejemplos

Para evaluar la utilidad de un modelo también es importante obtener información sobre qué atributos se consideran más importantes en sus decisiones.

Debéis obtener/extraer esa información según el tipo de modelo usando los métodos de interpretabilidad que se han explicado en clase.

2.6. Entrega del proyecto

El informe final ha de ser descriptivo, que habéis hecho, que dificultades habéis encontrado y que soluciones habéis dato a los problemas, que resultados ha obtenido cada modelo y como se comparan entre ellos, cuales son las limitaciones de cada modelo en el conjunto de datos.

El **informe final** que debe entregar ha incluir las siguientes **secciones**:

1. Una descripción breve del trabajo y sus objetivos, y de los datos disponibles, y cualquier información adicional que se haya recopilado y utilizado
2. Una breve descripción de los estudios previos sobre el conjunto de datos y los resultados relacionados (mirad la bibliografía adjunta a los datos)
3. El proceso de exploración de los datos (preprocesamiento, extracción/selección de características, agrupamiento y visualización)
4. El protocolo de remuestreo (entrenamiento/prueba, validación cruzada, etc.) que se ha utilizado, como se han escogido los parámetros del remuestreo en función de las características de los datos
5. Los resultados obtenidos usando **tres métodos lineales/cuadráticos** (indicando el mejor conjunto de parámetros para cada uno):
   - *a*) Si la tarea es **clasificación**, cualquiera de los siguientes métodos es aplicable: regresión logística, regresión multinomial (perceptrón de una sola capa), LDA, QDA, Naive Bayes, k-vecinos más cercanos, SVM lineal, SVM cuadrático
   - *b*) Si la tarea es **regresión**, cualquiera de los siguientes métodos es aplicable: regresión lineal, ridge regresión, LASSO, k-vecinos más cercanos, SVM lineal, SVM cuadrática

6. Los resultados obtenidos usando **tres métodos no lineales** (indicando el mejor conjunto de parámetros para cada uno); para tareas de **clasificación** y **regresión**, cualquiera de los siguientes métodos es aplicable: el MLP, SVM con kernel RBF, Random Forest, gradient boosting, o cualquier combinación de clasificadores

7. Una comparativa de los resultados de los modelos con alguna explicación/análisis sobre su rendimiento

8. Una descripción y justificación del modelo final elegido, y una estimación honesta de su rendimiento, que le hace mejor para el problema que se ha resuelto

9. Los resultados del análisis de interpretabilidad de los modelos:

   - Qué atributos son más importantes para los modelos y si hay alguna diferencia entre modelos lineales y no lineales
   - La explicación de un modelo lineal y un modelo no lineal de vuestra elección de una selección de ejemplos que parezcan relevantes (por ejemplo: valores atípicos, ejemplos con valores extremos, ejemplos en los valores máximos/mínimos para el valor de salida, ejemplos clasificados incorrectamente, ejemplos donde no se usan los atributos más relevantes, ejemplos que parecen prototípicos...)

10. Una breve parte final que contenga:

    a) Una autoevaluación de éxitos, fracasos y dudas

    b) Conclusiones científicas y personales

    c) Posibles extensiones y limitaciones conocidas

11. Referencias a todas sus fuentes utilizadas: libros, páginas web, código, artículos científicos...

Entrega:

Deberéis entregar un informe escrito (un archivo pdf) y el **código completo** (scripts python/notebooks) y un breve archivo de texto con las instrucciones necesarias sobre cómo ejecutar el código si es el caso.

No entreguéis simplemente un notebook, debéis hacer un informe explicativo de lo que habéis hecho y del análisis de los resultados. Fijaos que la competencia de la asignatura se evalúa con el informe.

El informe *no* debe incluir explicaciones de los métodos vistos en clase, a menos que sea relevante para sus necesidades. La entrega se hará exclusivamente por medio del "Racó" en un **archivo único comprimido** que incluya el informe y el código/notebooks. La fecha límite es **12 de Enero, 2025**.

CAPÍTULO 3

Evaluación

3.1. Evaluación de tareas

La calificación se basará en parte en la **claridad** del informe, así que aseguraos de que el informe final esté bien organizado y claramente escrito. Debe haber una parte introductoria que explique los conceptos básicos del trabajo y una sección de conclusiones, básicamente indicando lo que se ha aprendido en comparación con lo que sabía antes de que comenzara el trabajo; las limitaciones y posibles extensiones de vuestro trabajo deben ser detalladas y explicadas.

El trabajo también será evaluado a partir de **calidad técnica**. Esto significa que las técnicas que se utilicen deben ser razonables, los resultados indicados deben ser precisos y los resultados técnicos deben ser correctos y completos.

En resumen, estas son las condiciones para una puntuación alta (en este orden):

1. El (buen) uso de las técnicas y métodos presentados en clase
2. El cuidado y rigor en la obtención de los resultados (protocolo de validación, significación estadística)
3. La originalidad del análisis probando técnicas/métodos vistos en la asignatura para mitigar problemas que puedan tener los datos
4. La calidad de los resultados obtenidos (error de generalización, simplicidad)
5. La calidad del informe escrito (concisión, exhaustividad, claridad).

3.1.1. Criterios de evaluación

Los siguientes elementos se utilizarán como guía para evaluar la tarea. A cada elemento se le otorgará una calificación entre 0 y 10 según esta tabla. Tened en cuenta que no hay ningún elemento sobre la calidad del modelo, eso se debe a que tener un mayor acierto o R^2 no afectará a la nota. Eso no quiere decir que no se haya de hacer un mínimo para tener un resultado aceptable. Lo que afectará a

la calificación es justificar adecuadamente cada paso y utilizar metodologías adecuadas para que cada resultado sea real.

1.	**Ejecución de código**
8-10	El código tiene un enlace o script válido para descargar los datos, tiene un archivo `requirements.txt` con todas sus dependencias, tiene un README.md con instrucciones válidas y se puede ejecutar sin error siguiendo estas instrucciones.
5-7	El código tiene un enlace válido para descargar los datos, tiene un README.md con instrucciones válidas y se puede ejecutar sin error siguiendo estas instrucciones.
0-4	No hay documentación sobre cómo obtener los datos o ejecutar el código y el código devuelve un error cuando se ejecuta.
2.	**Estudio del conjunto de datos**
8-10	Se hace un estudio estadístico de los datos y sus relaciones y se aplican técnicas de visualización y reducción de dimensionalidad adecuadamente
5-7	Se calcula la estadística descriptiva básica de los datos y se visualizan las distribuciones y relaciones entre variables
0-4	El estudio se limita a describir la tipología de los datos
3.	**Preprocesamiento de datos**
8-10	El preprocesamiento de datos es adecuado al conjunto de datos, maneja correctamente sus particularidades y es coherente con los modelos utilizados posteriormente. Cada decisión al respecto está debidamente explicada en el documento. El preprocesamiento se aplica adecuadamente a las diferentes particiones de datos.
5-7	El preprocesamiento de datos es suficiente para solucionar la mayoría de los "problemas"del conjunto de datos, por ejemplo valores faltantes. No se explican las decisiones sobre el preprocesamiento.
0-4	No se ha realizado ningún preprocesamiento de los datos o el preprocesamiento realizado no tiene sentido. El procesamiento previo se aplica incorrectamente a las diferentes particiones de los datos.
4.	**Protocolo de remuestreo**
8-10	El protocolo de remuestreo es adecuado para el problema y se usa correctamente. El protocolo de remuestreo elegido se explica adecuadamente en el informe.
5-7	El protocolo de remuestreo tiene algún error menor que no afecta a la validez de los resultados o no se explica correctamente.
0-4	El protocolo de remuestreo no tiene ningún sentido e implica que todos los resultados no son válidos. El protocolo de remuestreo elegido no se explica correctamente en el informe y tiene errores.

5.	**Selección de hiperparámetros**
8 -10	La selección de hiperparámetros utiliza una metodología adecuada y suficiente, se explica adecuadamente y se justifica toda decisión al respecto.
5-7	La selección de hiperparámetros utiliza una metodología adecuada pero no muy exhaustiva y no se explica lo suficientemente bien, ni se justifica.
0-4	La selección de hiperparámetros no se explica o se ignora por completo.
6.	**Selección del modelo final**
8-10	La selección del modelo utiliza una metodología adecuada, se explica adecuadamente y se justifica toda decisión al respecto. Se utiliza una metodología adecuada para estimar su error de generalización.
5-7	La selección del modelo tiene algún error menor o no está suficientemente justificada, pero no afecta a la validez de los resultados.
0-4	La selección de modelo no se explica o se ignora por completo. El error de generalización no se estima o se estima mal.
7.	**Adaptación al contexto de su problema**
8-10	Todas las decisiones tomadas son adecuadas al contexto del problema y debidamente justificadas.
5-7	La mayoría de las decisiones tomadas son adecuadas al contexto del problema y debidamente justificadas.
0-4	No se han tomado decisiones adecuadas al contexto del problema y debidamente justificadas.
8.	**Justificación de las decisiones**
8- 10	Todas las decisiones de diseño tomadas durante el proyecto están justificadas y explicadas.
5-7	La mayoría de las decisiones de diseño tomadas durante el proyecto están justificadas y explicadas.
0-4	Ninguna decisión de diseño (o casi ninguna decisión de diseño) tomada durante el proyecto está justificada y explicada.
9.	**Error de generalización final**
5 -10	El error de generalización final se estima correctamente utilizando una metodología adecuada.
0-4	El error de generalización final se estima incorrectamente o se usa la partición de datos incorrecta.
10.	**Análisis de interpretabilidad**
7- 10	El análisis compara la relevancia de los atributos del problema en los diferentes tipos de modelos y analiza la explicación de una selección de ejemplos relevantes.
3-6	El análisis compara la relevancia de los atributos del problema en los diferentes tipos de modelos.
0-2	El análisis se limita a mostrar la relevancia de los atributos de los modelos.

3.2. Evaluación de competencias genéricas

Además, como probablemente sepáis, hay una **competencia genérica** (o *habilidad*) asociada a este curso: "*Comunicación eficaz*"[1], que supone un 10 % de la nota final de la asignatura. Esta será la **rúbrica** con la que se evaluará la competencia.

Cada indicador tiene cuatro posibles respuestas y cuatro notas: A, B, C, D. El profesor decide primero en cuál de los cuatro casos se encuentra el documento para cada indicador y luego pone una nota numérica de acuerdo a la tabla:

[0-5)	D
[5-6.5)	C
[6.5-8.5)	B
[8.5 -10]	A

Por ejemplo, si un indicador tiene una B, entonces se coloca una nota en el rango [6.5 - 8.5) en función de su grado. Luego se suman las calificaciones y se obtiene una calificación de la competencia. Esta nota cuenta un 10,0 % de la nota final de la asignatura. La conversión de la nota a A, B, C, D (aplicando la tabla al revés) da la nota final de la competencia.

3.2.1. Criterios de evaluación

1.	**Presentación general**
A	El documento tiene una portada con información relevante, utiliza márgenes apropiados, no corta tablas ni figuras y numera las páginas. Tiene un índice numerado.
B	El documento tiene una portada con la información relevante, numera las páginas y tiene un índice.
C	El documento numera las páginas, pero tiene portada mejorable, márgenes no muy adecuados, y no tiene índice (para documentos de tamaño mediano o grande.)
D	El documento no tiene portada, márgenes adecuados, número de páginas, ni índice para un documento de tamaño mediano o grande.
2.	**Introducción**
A	Introduce el tema principal y anticipa la estructura del trabajo. El propósito, exposición general del tema y los objetivos son muy claros, y permiten tener una completa y rápida idea del trabajo, incluidos los resultados.
B	El propósito, presentación general del tema y objetivos son razonablemente claros, pero no permiten hacerse una idea completa y rápida del trabajo.
C	El propósito, el tema y los objetivos se presentan, pero de una manera no objetiva y clara.
D	Sin Introducción, o lo que hay es completamente ineficaz.

[1] Expressió oral i escrita.

3.	**Organización y estructura general**
A	Las ideas se presentan en un orden lógico, siempre con coherencia y fluidez, y la lectura es interesante.
B	Las ideas se presentan en un orden lógico, generalmente con coherencia y fluidez.
C	Las ideas se presentan en un cierto orden, pero muchos detalles no están en un orden lógico o esperado, y distraen al lector.
D	Las deas no se presentan en un orden lógico, falta coherencia y el orden de los párrafos no refuerza el contenido sino que lo oscurece.
4.	**Corrección gramatical**
A	No hay errores ortográficos, sintácticos o tipográficos. El lenguaje utilizado es rico.
B	No hay errores ortográficos ni sintácticos, pero sí algunos errores tipográficos. El lenguaje utilizado es razonablemente variado.
C	Hay algunos errores ortográficos o sintácticos, o errores tipográficos. El lenguaje utilizado es justo.
D	Hay muchos errores ortográficos o sintácticos, o errores tipográficos. El lenguaje utilizado es pobre.
5.	**Estilo del lenguaje**
A	El documento está escrito con un lenguaje preciso y didáctico.
B	El documento está escrito con un lenguaje aceptablemente preciso y comprensible, aunque no sea completamente didáctico.
C	El documento utiliza un lenguaje poco preciso o gramaticalmente dudoso (ambigüedades, contradicciones...).
D	El documento utiliza un lenguaje lleno de inconsistencias y ambigüedades y, a menudo, es gramaticalmente incorrecto.
6.	**Elementos visuales (tablas, gráficos...)**
A	Se utilizan elementos visuales cuando conviene, están bien hechos (ni cargados, ni deficientes) y ayudan mucho a la comprensión de la obra.
B	Se utilizan elementos visuales cuando corresponde, se realizan de manera aceptable y, a menudo, ayudan a comprender el trabajo.
C	Se utilizan elementos visuales, pero no siempre de manera apropiada, no siempre están correctamente realizados realizados y no siempre ayudan en la comprensión del trabajo.
D	Authors no utilizan elementos visuales cuando es conveniente, o no están bien hechos o dificultan la comprensión del trabajo.

7.	**Conclusiones**
A	El documento finaliza con un resumen claro que describe lo que se puede deducir del trabajo realizado, qué trabajo se podría haber hecho de otra manera, cuál sería el trabajo a realizar si se quisiera seguir trabajando porque han aprendido haciendo el trabajo.
B	El documento finaliza con un resumen claro que describe lo que se puede deducir del trabajo realizado y lo que se ha aprendido al hacer el trabajo.
C	El documento finaliza con un resumen que describe en parte lo que se puede deducir del trabajo realizado.
D	El documento no incluye conclusiones o estas son manifiestamente incompletas o incorrectas.
8.	**Referencias (Libros, páginas web, periódicos, cursos, apuntes...)**
A	El documento identifica y cita correctamente la información utilizada y da referencias adicionales.
B	El documento identifica y cita la información utilizada.
C	El documento identifica y cita la información utilizada, pero no correctamente.
D	El documento no cita la información utilizada.