Apuntes

Basado en las transcripciones de las sesiones de repaso y las instrucciones del profesor, el examen será **eminentemente práctico**, centrado en la programación en R y la interpretación de resultados para la toma de decisiones de negocio.

Aquí tienes el desglose de en qué consistirá y qué tipo de ejercicios caerán:

1. Formato y Logística del Examen

- **Duración:** 2 horas.
- **Herramientas:** Se permite el uso de ordenador, R/RStudio, Excel, código impreso y, preferiblemente, **código en formato Word** (especialmente recomendado para exámenes online para evitar problemas con la cámara).
- **Archivos:** Se descargarán dos bases de datos (en Excel o CSV) protegidas por contraseña (la contraseña se da en el examen).
- **Evaluación:** Se corrige por **planteamiento y razonamiento**. Si el código da error pero explicas correctamente qué pasos ibas a seguir y por qué, puedes obtener puntuación parcial.
- **Plataforma:** Las respuestas se escriben en cuadros de texto en la plataforma de examen, pero **es obligatorio adjuntar el archivo de script (.R)** al final como justificación del trabajo.

2. Estructura y Tipos de Ejercicios

El examen constará de entre **15 y 20 preguntas**. Muchas son cortas (filtros, medias) y otras requieren desarrollar modelos. La **Actividad 3** se menciona explícitamente como una "preparación para el examen" porque combina regresión, clasificación y clustering.

Los bloques de ejercicios serán:

A. Preparación y Exploración de Datos (Fundamental)

Es la base para aprobar. Se espera que sepas: * **Cargar datos:** Leer archivos `.csv` o `.xlsx`. * **Calidad del dato:** Detectar y contar valores nulos (`is.na`, `colSums`) y decidir si eliminarlos (`na.omit`) o imputarlos. * **Tipos de variables:** Identificar numéricas vs. texto y convertir variables categóricas a `factor` (crucial para modelos de clasificación). * **Estadística descriptiva:** Interpretar el comando `summary` (media vs. mediana para detectar outliers).

B. Manipulación de Datos (Filtros y Tablas)

Habrá preguntas rápidas para ganar puntos ágiles: * **Filtros:** "¿Cuántas mujeres tienen un gasto superior a 1000?". Usar `filter` o corchetes. * **Tablas de frecuencia:** Usar `table` (conteo absoluto) y `prop.table` (porcentajes) para responder preguntas como "¿Cuál es la probabilidad de comprar una bicicleta?". * **Creación de columnas:** Crear una variable binaria (0/1) a partir de una columna de texto (ej. "Si" -> 1, "No" -> 0) usando `ifelse`.

C. Modelos de Machine Learning (El núcleo del examen)

Es muy probable (80% de probabilidad según la profesora) que caigan ejercicios de estos tres tipos:

1. **Regresión Logística y Árboles de Decisión (Clasificación Supervisada):**
   - *Objetivo:* Predecir una variable binaria (Sí/No, 0/1).
   - *Pregunta típica:* "¿Qué modelo tiene mayor precisión?". Deberás calcular la **Matriz de Confusión** y comparar el **Accuracy** (Exactitud).
   - *Pregunta de negocio:* "¿Qué variables son más importantes?". Usar `variable.importance` en árboles o mirar la significancia en regresión.
   - *Predicción de caso:* "Calcula la probabilidad de que un pasajero de 30 años, mujer, sobreviva". Usar la función `predict` con un `data.frame` nuevo.
2. **Clustering (No Supervisado):**
   - *Objetivo:* "Definir tipologías de clientes" o agrupar sin conocer las etiquetas.
   - *Pasos clave:* Seleccionar solo variables numéricas, determinar el número óptimo de grupos (K) usando `NbClust` (regla de la mayoría) o el método del Codo, y ejecutar `kmeans`.
   - *Interpretación:* Describir los grupos resultantes (ej. "El grupo 1 son clientes que gastan mucho").
3. **Regresión Lineal:**
   - *Objetivo:* Predecir un valor numérico continuo (ej. precio).
   - *Pregunta clave:* "¿Es significativa la variable precio?". Mirar si el **p-valor < 0.05** (si tiene asteriscos es significativa, si no, se puede despreciar).
   - *Interpretación:* "¿Cuánto aumenta Y si X aumenta en 1 unidad?". Mirar el coeficiente (`Estimate`).

3. Temas de Baja Probabilidad

- **Series Temporales:** Aunque está en el temario, la profesora indicó que es "mucho menos importante de cara al examen" y "muy raro que os entren series temporales". Si entra, sería usar `auto.arima` y `forecast` para predecir a corto plazo.

4. Consejos Clave para el Examen

- **Responde lo que se pregunta:** Si no te piden evaluar si el modelo es bueno, no pierdas tiempo haciendo la división de entrenamiento/testeo (train/test split). Si te lo piden, entonces es obligatorio.
- **Si te bloqueas:** Escribe en el cuadro de texto los pasos que darías ("Haría la limpieza de nulos, luego convertiría a factor…") aunque el código no funcione. Se valora el razonamiento.

- **Código preparado:** Ten tu script o Word listo para copiar y pegar estructuras de código (carga, limpieza, modelos), cambiando solo los nombres de las variables.

Basado en los comentarios explícitos de la profesora a lo largo de las sesiones de repaso y las transcripciones, aquí tienes una recopilación de las preguntas y conceptos con **mayor probabilidad** de caer en el examen, así como aquellos que tienen menos peso.

La profesora menciona literalmente: *"Regresión logística, árbol [de decisión] y clasterización son, vamos, a un 80% de posibilidades"*.

1. 🔥 Conceptos de "Muy Alta Probabilidad" (El núcleo del examen)

Se espera que el examen se centre en ejercicios prácticos donde tengas que programar y, sobre todo, **interpretar** los resultados.

A. Clasificación (Regresión Logística y Árboles de Decisión)

Es el bloque más enfatizado. Es muy probable que tengas que comparar dos modelos. * **Pregunta clave:** "¿Qué modelo de clasificación tiene una mayor precisión?" o "¿Cuál es mejor?". * **Lo que busca:** Que calcules la **Matriz de Confusión** de ambos modelos y compares el **Accuracy** (Exactitud). El que tenga mayor *Accuracy* es el mejor. * **Pregunta de interpretación:** "Interpreta la sensibilidad o especificidad". * **Concepto:** Recordar el ejemplo del **Test COVID**. Sensibilidad es detectar bien los positivos (enfermos); Especificidad es detectar bien los negativos (sanos). * **Pregunta de predicción:** "Calcula la probabilidad de que [un caso nuevo específico, ej. Leonardo DiCaprio] sobreviva/compre". * **Tarea:** Usar la función `predict` con un `data.frame` nuevo que contenga los valores que te dé el enunciado. * **Pregunta de negocio:** "¿Cuáles son las variables más importantes?". * **Tarea:** Usar `variable.importance` en árboles de decisión o mirar la significancia en la regresión.

B. Clasterización (Aprendizaje No Supervisado)

- **Pregunta clave:** "Definir tipologías de clientes" o agrupar datos.
- **Tarea:** Determinar el número óptimo de grupos (K). La profesora recomienda usar **`NbClust`** (regla de la mayoría) o el **método del Codo**.
- **Pregunta de interpretación:** "¿Qué características tiene el grupo 1?".
  - **Lo que busca:** Que mires las medias de las variables (ej. "el grupo 1 son los que más gastan") o que hagas un árbol de decisión sobre el clúster para entender cómo se ha dividido (el "truco" para explicar la caja negra).

C. Regresión Lineal

- **Pregunta clave:** "¿Qué variables son significativas?".
  - **Lo que busca:** Que mires el **p-valor** (`Pr(>|t|)`). Si es **< 0.05** (o tiene asteriscos), es significativa. Si es mayor, no lo es.
- **Pregunta de interpretación:** "¿Cuánto aumenta la variable objetivo si aumentamos X en una unidad?".
  - **Lo que busca:** Que interpretes el **coeficiente** (*Estimate*). Es el valor que multiplica a la variable.

2. 🛠️ Preguntas “Rápidas” (Puntos fáciles)

La profesora mencionó que habrá entre 15 y 20 preguntas, y muchas serán cortas para “rascar puntos”.

- **Filtros y Conteos:** “¿Cuántas mujeres tienen un gasto superior a 1000?”. Usar `filter` y `nrow` o `dim`.
- **Tablas de Frecuencia:** “¿Cuál es la probabilidad/porcentaje de X?”. Usar `table` y `prop.table`.
- **Creación de columnas:** Crear una variable binaria (0/1) a partir de texto (ej. si es “Yes” pon un 1, si es “No” pon un 0). La profesora dijo: *“Me cansaré de repetirlo, esta parte es muy importante”*.
- **Nulos:** “¿Existen celdas vacías?”. Usar `colSums(is.na(datos))`.

3. 📉 Conceptos de “Baja Probabilidad”

- **Series Temporales:** Aunque están en el temario, la profesora comentó repetidamente que *“es muy raro que os entren series temporales”* o que *“tiene menos probabilidades de entrar”*. Si entra, sería algo básico con `auto.arima`.
- **Teoría Pura:** *“No os entra teoría”*. El examen es práctico; no te pedirán definiciones de memoria, sino aplicar conceptos.

4. 💡 Consejos de la Profesora para el Examen

1. **Código preparado:** Ten tu archivo Word o impreso con los códigos (“chunks”) listos para copiar y pegar. No se evalúa que te sepas el código de memoria, sino que sepas usarlo.
2. **No te bloquees con errores:** Si el código falla, **escribe el planteamiento** en el cuadro de texto (“Haría un filtro, luego un modelo…”). Se corrige por planteamiento y razonamiento, no solo por la ejecución perfecta.
3. **Responde lo que se pregunta:** Si no te piden evaluar el modelo, no pierdas tiempo dividiendo en *train/test*. Si te piden evaluar la precisión, entonces sí es obligatorio dividir.
4. **Usa `na.omit`:** Para limpiar nulos en el examen, la profesora recomienda `na.omit` por ser lo más sencillo y seguro para evitar errores.