**Idea 1 — Perfil estadístico poblacional de rasgos faciales**

**Objetivo:** describir distribuciones y prevalencias de cada rasgo en la muestra.

**Paso a paso**

1. **Preparar fichero maestro** con variables categóricas codificadas (ver plantilla más abajo).
2. **Etiquetado inicial**: etiquetar todas las fotos con las variables de la ficha por al menos 1 observador (y preferible 2). Registrar CALIDAD\_FOTO.
3. **Limpieza**: revisar valores faltantes y validar codificaciones inconsistentes.
4. **Estadística descriptiva**: calcular frecuencias absolutas y relativas para cada rasgo; IC 95% para proporciones.
5. **Subgrupos**: comparar por sexo, grupo etario, región (Chi² / Fisher según frecuencias).
6. **Reportes**: tablas de frecuencias por rasgo y gráficos (barras, pirámides por sexo).

**Entregables**

* CSV limpio con codificaciones.
* Informe con tablas, gráficos y resumen ejecutivo.
* Script reproducible en R/Python para reproducir tablas y gráficos.

**Idea 2 — Agrupamiento de perfiles faciales (tipologías)**

**Objetivo:** encontrar tipologías (clusters) de rostros en la muestra.

**Paso a paso**

1. **Codificar variables**: transformar categorías a dummy / usar ACM para variables puramente categóricas.
2. **ACM** (Análisis de Correspondencias Múltiples): reducir dimensiones manteniendo relaciones de variables categóricas.
3. **Clustering**: aplicar k-means (sobre ejes principales del ACM) y/o clustering jerárquico. Evaluar k óptimo con silhouette/Calinski-Harabasz.
4. **Interpretación**: describir cada cluster por los rasgos más representativos (vía contribuciones del ACM o frecuencias dentro del cluster).
5. **Validación**: bootstrap o reclasificación por cross-validation; revisar estabilidad.

**Entregables**

* Dendrograma y gráfico de clusters (scatter en ejes ACM).
* Perfil de cada tipología (texto y tabla de probabilidades condicionales por rasgo).
* Script R/Python y reporte.

**Idea 3 — Modelos predictivos para rasgos no observables (imputación)**

**Objetivo:** predecir rasgos no visibles por baja calidad o cobertura parcial.

**Paso a paso**

1. **Seleccionar variables predictoras**: rasgos visibles, EDAD, SEXO, CALIDAD\_FOTO, metadatos (región).
2. **Dividir datos**: training/validation (p. ej., 70/30).
3. **Modelos**:
   * Para rasgos binarios → regresión logística (con interacciones si aplica).
   * Para rasgos categóricos multiclase → regresión multinomial u ordinal.
   * Regularización (Lasso/ElasticNet) para selección.
4. **Evaluación**: AUC para binarios, accuracy y matriz de confusión para multiclas, calibración.
5. **Implementar imputación**: para fotos con rasgo no observado, predecir probabilidades y/o imputar la categoría más probable con umbral de confianza.
6. **Registrar incertidumbre**: guardar probabilidades y marcar imputaciones como “imputado\_modelo” o “observado”.

**Entregables**

* Modelos entrenados + métricas.
* Hoja de reglas para aplicar imputaciones y umbrales.
* Script reproducible.

**Idea 4 — Fiabilidad entre observadores y ajuste de sesgos**

**Objetivo:** medir concordancia entre etiquetadores y corregir sesgos sistemáticos.

**Paso a paso**

1. **Seleccionar muestra de validación** (n ≈ 50–100 fotos) que cubra variedad de rasgos y calidades.
2. **Etiquetado multi-observador**: 3–5 observadores independientes codifican la muestra.
3. **Cálculo de concordancia**: Cohen’s kappa (pares) y Fleiss’ kappa (grupo). Para ordinales usar weighted-kappa.
4. **Modelos de efectos**: modelo de efectos mixtos para estimar varianza entre observadores y por rasgo.
5. **Calibración**: crear guía visual y sesiones de re-entrenamiento para observadores que muestren sesgos. Re-evaluar.
6. **Ajuste post-hoc**: si un observador tiene sesgo sistemático, aplicar correcciones (p. ej. recalibrar sus etiquetas mediante matriz de confusión estimada).

**Entregables**

* Matrices de concordancia y kappa por rasgo.
* Protocolo de etiquetado y guía visual.
* Recomendaciones de formación.

**Idea 5 — Análisis espacial y correlación con zonas de origen**

**Objetivo:** evaluar si rasgos o tipologías se distribuyen espacialmente.

**Paso a paso**

1. **Georreferenciar**: asignar códigos de departamento/municipio (asegúrate de confidencialidad).
2. **Calcular prevalencias por región** para rasgos y clusters.
3. **Pruebas básicas**: Chi² estratificado por región.
4. **Autocorrelación espacial**: Moran’s I o índices similares sobre índices compuestos (p. ej., índice de “nariz ancha” por municipio).
5. **Modelos espaciales**: GLMM con efecto aleatorio por municipio o modelos espaciales (CAR/ICAR) si corresponde.
6. **Mapas temáticos**: choropleth por prevalencia y mapas de hot-spots.

**Entregables**

* Mapas temáticos y reporte espacial con resultados estadísticos.
* Modelos y scripts reproducibles.

**Idea 6 — Fusión bayesiana de evidencias para cotejos con testimonios**

**Objetivo:** combinar la evidencia fotográfica (rasgos codificados) con descripciones testimoniales para obtener una probabilidad de consistencia.

**Paso a paso**

1. **Calcular probabilidades marginales** P(rasgo | población) (frecuencias desde Idea 1).
2. **Estimar verosimilitudes**: P(testimonio | misma persona) y P(testimonio | no misma persona). Esto se aproxima a partir de sensibilidad/ especificidad de observadores o mediante opinión experta.
3. **Likelihood Ratios (LR)** por rasgo: LR = P(obs|match) / P(obs|no match).
4. **Combinar LRs**: multiplicación (o sumar log-LR) para rasgos independientes; si no independientes, usar fusión bayesiana con pesos / calibración.
5. **Aplicar Bayes**: partir de un prior (p. ej., prevalencia de que una foto corresponde a la persona buscada) y actualizar con LR para obtener posterior.
6. **Incluir incertidumbre**: usar MCMC o aproximaciones bayesianas para intervalos.

**Entregables**

* Hoja de cálculo / pequeña app que calcula posterior dada lista de rasgos/testimonios.
* Guía de interpretación (umbrales LR y significancia).

**Cronograma propuesto (desde 16 sep 2025 → finales de octubre 2025)**

Organizo por **semanas**. Empieza la **Semana 1** que arranca el **16 de septiembre de 2025**.

**Semana 1 (arranca 16 sep)**

* Actividades transversales: definir plantilla CSV, protocolos éticos, y procedimiento de anonimización.
* Reunión inicial del equipo, asignación de roles (etiquetadores, analista estadístico, GIS, responsable QC).
* Crear guía visual preliminar para codificación de rasgos.

**Semana 2**

* Etiquetado inicial de un 20–30% de la muestra (fotos con mejor calidad).
* Empezar análisis descriptivo preliminar (Idea 1) sobre la submuestra.
* Seleccionar muestra para evaluación inter-observador.

**Semana 3**

* Evaluación entre observadores (Idea 4): cálculo de kappa y retroalimentación.
* Ajustes a la guía visual según resultados del kappa.
* Continuar etiquetado del resto de fotos.

**Semana 4**

* Completar etiquetado (o la mayor parte). Registrar CALIDAD\_FOTO.
* Ejecutar análisis descriptivo final (Idea 1): tablas y gráficos.
* Preparar datos para ACM (Idea 2) y clustering.

**Semana 5**

* Ejecutar ACM + clustering; seleccionar número de clusters y describir tipologías (Idea 2).
* Iniciar modelado predictivo (Idea 3): preparar training/validation y primeros modelos.

**Semana 6**

* Validación de modelos predictivos; documentar reglas de imputación (Idea 3).
* Ejecutar análisis espacial básico y mapas (Idea 5): prevalencias por municipio/ departamento.
* Preparar ejemplos de fusión bayesiana (Idea 6) usando proporciones de Idea 1.

**Semana 7 (últimas semanas de octubre)**

* Finalizar reportes (Idea 1–6): consolidar entregables (CSV, scripts, reportes PDF).
* Reunión de cierre: presentar resultados, discusión política y recomendaciones operativas (protocolo de uso para familiares/organizaciones).
* (Opcional) implementar una hoja Excel / app simple para calcular LR/posterior con ejemplos.

Nota: ajusta la intensidad por semana según tamaño de la muestra y disponibilidad del equipo. Para muestras grandes, ampliar semanas de etiquetado.

**Plantilla de ficha codificada (ejemplo de columnas)**

(usa valores numéricos para facilitar análisis; incluye CALIDAD\_FOTO y OBSERVADOR)

ID, NOMBRES\_APELLIDOS, NOMBRES, APELLIDOS, SEXO, ULTIMA\_FECHA, EDAD, DEPARTAMENTO, MUNICIPIO, CALIDAD\_FOTO, FRENTE, CEJAS, OJOS\_TAM, OJOS\_FORMA, OJOS\_SEPARACION, NARIZ\_FORMA, NARIZ\_ANC, LABIOS, MENTON, MANDIBULA, POMULOS, OREJAS, ROSTRO\_FORMA, CICATRIZ, LUNAR, BARBA\_MOSTRADA, OBSERVADOR, FECHA\_ETIQUETA, IMPUTADO\_FLAG

**Ejemplo de codificación (sugerida)**

* FRENTE: 1=ancha, 2=media, 3=estrecha
* CEJAS: 1=pobladas, 2=delgadas, 3=arqueadas, 4=rectas
* OJOS\_TAM: 1=grandes, 2=medianos, 3=pequeños
* OJOS\_FORMA: 1=almendrados, 2=redondos, 3=rasgados
* NARIZ\_FORMA: 1=recta, 2=aguileña, 3=respingada, 4=chata
* LABIOS: 1=gruesos, 2=medios, 3=delgados
* ROSTRO\_FORMA: 1=ovalado, 2=cuadrado, 3=redondo, 4=alargado, 5=triangular
* CALIDAD\_FOTO: 1=alta, 2=media, 3=baja
* IMPUTADO\_FLAG: 0=observado, 1=imputado

**Caso / Ejemplo con la foto y los datos suministrados**

**Metadatos (tal como diste):**

* ID = 7ea5c83ffc25eb8a2078b92dbbb5f0cd97ba400c8957e1afb10dff66d96dbeb8
* NOMBRES\_APELLIDOS = ALDEMAR ANTONIO MONTES ARCILA
* NOMBRES = ALDEMAR ANTONIO
* APELLIDOS = MONTES ARCILA
* SEXO = MASCULINO
* ULTIMA\_FECHA = 22/07/2012
* EDAD = 12
* DEPARTAMENTO = RISARALDA
* MUNICIPIO = SANTA ROSA DE CABAL

Observación importante: **no estoy “identificando”** la persona — sólo describo rasgos visibles en la foto para propósitos de la base estadística que deseas construir.

**1) Evaluación visual y codificación de rasgos (ficha completada — EJEMPLO)**

Tras examinar la foto (calidad: **media-baja** por escaneo y ruido), una codificación ejemplo sería:

* CALIDAD\_FOTO = 2 (media)
* FRENTE = 2 (media) — frente no especialmente ancha ni muy estrecha
* CEJAS = 4 (rectas, ligeramente pobladas)
* OJOS\_TAM = 3 (pequeños/medianos-pequeños)
* OJOS\_FORMA = 2 (redondos / ligeramente almendrados)
* OJOS\_SEPARACION = 2 (separados-medios)
* NARIZ\_FORMA = 1 (recta, puente moderado)
* NARIZ\_ANC = 2 (ancho medio)
* LABIOS = 3 (delgados)
* MENTON = 2 (ligeramente prominente, no muy puntiagudo)
* MANDIBULA = 1 (línea mandibular definida, algo cuadrada)
* POMULOS = 2 (moderados)
* OREJAS = 2 (medianas, pegadas moderadamente)
* ROSTRO\_FORMA = 2 (cuadrado/rectangular juvenil)
* CICATRIZ = 0 (no visible)
* LUNAR = 0 (no visible)
* BARBA\_MOSTRADA = 0 (no)

**Fila CSV ejemplo (valores simplificados):**  
7ea5...,ALDEMAR ANTONIO MONTES ARCILA,ALDEMAR ANTONIO,MONTES ARCILA,MASCULINO,22/07/2012,12,RISARALDA,SANTA ROSA DE CABAL,2,2,4,3,2,1,2,3,2,1,2,2,2,0,0,0,OBS1,2025-09-20,0

(OBS1 = observador 1; fecha ejemplo)

**2) Uso en cada idea — ejemplo operativo rápido**

**Idea 1 (f recuencias)**

* Tras codificar N fotos, calculas que LABIOS=delgados tiene frecuencia 0.38 (38%). Para nuestro registro, Aldemar estaría en ese 38%.

**Idea 2 (clustering)**

* Ejecutas ACM + k-means; Aldemar resulta asignado a **Cluster 3** con probabilidad 0.85. Cluster 3 descrito como: {rostro más bien cuadrado, labios delgados, nariz recta, ojos pequeños}.
* Entregable: ficha de cluster con centroides (proporciones de rasgos).

**Idea 3 (imputación)**

* Si la foto tuviera la boca parcialmente oscurecida, el modelo (logit) usa sexo+edad+nariz+ojos para predecir LABIOS. Si el modelo devuelve P(labios\_delgados)=0.72 (>0.7 umbral), imputas delgados y marcas IMPUTADO\_FLAG=1.

**Idea 4 (concordancia)**

* Si 3 observadores codifican la foto y kappa para LABIOS fue 0.82, la concordancia es alta; si kappa<0.6, revisaríamos ese rasgo y re-entrenaríamos.

**Idea 5 (espacial)**

* Agregas Aldemar a SANTA ROSA DE CABAL; calculas prevalencia de OJOS\_PEQUEÑOS por municipio y comparas con otros municipios (mapa choropleth).

**Idea 6 (fusión bayesiana)**

* Supón que un familiar testimonial dice “ojos pequeños” y “labios delgados”.
  + Frecuencia poblacional: P(ojos\_peq)=0.30, P(labios\_delg)=0.38 (estimates from Idea 1).
  + Sensibilidad testimonial por estudio previo: P(testimonio dice ojos\_peq | realmente ojos\_peq)=0.8; especificidad=0.85 (ejemplo).
  + LR\_ojos = 0.8 / (1 - 0.85) = 0.8 / 0.15 ≈ 5.33.
  + LR\_labios = 0.75 / (1 - 0.80) ≈ 3.75 (ejemplo).
  + LR\_total ≈ 5.33 \* 3.75 ≈ 20.0.
  + Si nuestro prior (antes de la evidencia) era 0.05 (baja), posterior odds = 0.05/(1-0.05) \* 20 ≈ 1 → posterior ≈ 0.5.
  + Interpretación: con esos testimonios y las LRs asumidas, la probabilidad de consistencia aumenta mucho (ejemplo ilustrativo). Siempre documentar supuestos.

Nota: los números de sensibilidad/especificidad usados arriba son **ejemplificativos**; debes estimarlos con datos reales o expertos.

**1) Perfil estadístico poblacional**

**Beneficio:** entrega un mapa claro de rasgos predominantes y permite comparar subgrupos.  
**Negativo:** solo describe, no ayuda a identificar individualmente; puede ser percibido como “poco útil” por actores que esperan identificación directa.  
**Alerta:** define bien que es un análisis poblacional, no individual, para que no haya malentendidos.

**2) Agrupamiento de perfiles (tipologías)**

**Beneficio:** sintetiza diversidad en perfiles fáciles de comunicar (ej. 3–5 tipologías).  
**Negativo:** los clusters dependen de la calidad de los datos y pueden cambiar si la muestra es pequeña o sesgada.  
**Alerta:** valida los clusters con diferentes métodos y documenta criterios de selección de k (número de grupos).

**3) Modelos predictivos para rasgos no observables**

**Beneficio:** recupera información en fotos de baja calidad, evita perder casos.  
**Negativo:** puede imputar rasgos erróneamente si el modelo se entrena con pocos datos o datos sesgados.  
**Alerta:** marcar siempre qué rasgos son **imputados** y no observados directos, para no confundir a usuarios futuros.

**4) Fiabilidad entre observadores**

**Beneficio:** garantiza rigor, asegura que diferentes personas describan igual.  
**Negativo:** requiere tiempo extra (doble o triple codificación).  
**Alerta:** entrena a los observadores antes de etiquetar en masa; si no, tendrás que rehacer trabajo.

**5) Análisis espacial**

**Beneficio:** permite ver si hay patrones regionales (útil para priorizar búsquedas en terreno).  
**Negativo:** con pocos casos por municipio, los resultados pueden ser inestables o dar asociaciones falsas.  
**Alerta:** trabajar con regiones agregadas si los casos son escasos (departamento > municipio).

**6) Fusión bayesiana con testimonios**

**Beneficio:** transforma testimonios en evidencia probabilística defendible.  
**Negativo:** depende de estimaciones de sensibilidad/especificidad que no siempre existen; se necesita justificar los supuestos.  
**Alerta:** documentar muy bien de dónde salen tus probabilidades (literatura, expertos, consenso). Si no, será criticado por falta de base.

**Algo clave para evitar percances más adelante**

**Estandarización y documentación.**  
El mayor problema en proyectos de este tipo no es la estadística en sí, sino que los datos terminan:

* mal codificados,
* sin protocolo de cómo se etiquetó,
* sin distinción clara entre **observado** y **imputado**,
* o con observadores que no coinciden en criterios.

**Mi consejo:**

1. Define un **manual de codificación** (con fotos de ejemplo de cada categoría).
2. Guarda siempre en la base de datos: **observador**, **fecha de codificación** y si fue **observado o imputado**.
3. Haz un **piloto pequeño primero** para probar la plantilla, medir tiempos y concordancia.

Con eso reduces el 80% de los riesgos que suelen dañar este tipo de proyectos.