Parte 1:

1) En cuanto a las variables que se mencionan en la base de hogares, que pueden ser predictoras de pobreza, pero que no se menciona en las bases individuales, algunas de las que llaman la atención son:
    - pisos interiores: particularmente si son de tierra o ladrillo suelto
    - La cubierta exterior del techo es de: particularmente si es chapa de cartón o saña, tabla, paja con barro, paja sola
    - ¿Tiene baño / letrina?
    - El baño tiene: si responde que es inodoro sin botón / cadena y con arrastre de agua (a balde) o letrina (sin arrastre de agua)
    - La vivienda está ubicada en villa de emergencia (por observación)
    - Combustible utilizado para cocinar: si es garrafa de gas de garrafa o leña y carbón
    - Baño (tenencia y uso): si el uso es compartido o no tiene
    - Menores de 10 años ayudan con algún dinero trabajando?
    - Menores de 10 años ayudan con algún menor pidiendo?
2) Primero definimos los file path para cada base de datos y pedimos a Jupyter que las lea. Luego filtramos las bases por región y nos quedamos con las que el valor es 1 (que es GBA). Usamos la función pd.merge de Pandas para combinar los dos data frames usando CODUSU y NRO_HOGAR como claves de unión.
3) En este punto corregimos las variables binarias para que tomen solo valores 1 o 2. Eliminamos columnas y filtramos las filas donde los ingresos, la edad, las horas trabajadas, la cantidad de habitaciones y la cantidad de personas que viven en el hogar tienen valores coherentes.
4) En este inciso creamos las variables habitantes por habitación (dividiendo IX_TOT por IV2), Ingreso total familiar por personas mayores a 10 años y por último creamos la variable Ingreso total por asistencia por hogar, que representa cuánta asistencia estatal o de organismos como iglesias recibe una familia.
5) Las cinco variables que seleccionamos y sus interpretaciones son: 'IV12_3', 'IX_TOT', 'V5', 'V19_B', 'II9'.
    - **IV12_3 (Ubicación en villa emergencia):**
        - Media: 1.991577
        - Desviación estándar: 0.091402
        - Mínimo: 1
        - Máximo: 2
        - Esta variable indica si el hogar está ubicado en una villa emergencia (1) o no (2). La desviación estándar baja sugiere que la mayoría de los hogares están categorizados como no ubicados en villa emergencia.
    - **IX_TOT (Cantidad de personas que viven en el hogar):**
        - Media: 3.376997
        - Desviación estándar: 1.647869
        - Min: 1
        - Max: 12
        - Esta variable muestra el tamaño del hogar en términos de la cantidad de personas que lo habitan. La alta desviación estándar indica variabilidad en los tamaños de los hogares dentro del conjunto de

datos. Pensamos que puede ser un buen predictor de pobreza, dado que a excepción de alguna cuestión religiosa, hay evidencia que sugiere que las madres con menos recursos tienen más hijos. Estas razones van desde la falta de educación/recursos en materia de prevención del embarazo hasta teorías que sugieren que al tener menores recursos la expectativa de vida es menor, por lo tanto los padres tienen más hijos para que sus probabilidades de sustento en el futuro aumenten. Asimismo, otra razón por la que mayor cantidad de personas en el hogar podría sugerir que no hay una posibilidad económica para los hijos de abandonar el hogar de sus padres al formar una familia.

- **V5 (Reciben ayuda social - estado/iglesia):**
  - Media: 1.864653
  - Desviación estándar: 0.342144
  - Mínimo: 1
  - Máximo: 2
  - Comentario: Indica si el hogar recibe ayuda social del estado o de la iglesia (1) o no (2). La mayoría de los hogares no recibe ayuda social según la media cercana a 2. Esta podría ser también un buen predictor de pobreza, dado que quien necesita ayuda estatal o caritativa (en general) es porque no puede sustentar sus necesidades básicas con sus ingresos conseguidos mediante el trabajo. Asimismo, estas formas de asistencia suelen requerir una demostración de la escasez de recursos.
- **V19_B (Miembros de 10 años trabajan):**
  - Media: 1.999419
  - Desviación estándar: 0.024098
  - Mínimo: 1
  - Máximo: 2
  - Comentario: Indica si los miembros del hogar de 10 años trabajan (1) o no (2). La desviación estándar baja sugiere que la mayoría de los hogares tienen miembros de 10 años que no trabajan. Tambien podria ser un buen predictor de pobreza, dado que los chicos menores de 10 años, según los derechos del niño, deberían estar en la escuela y no trabajando. El hecho de que estén trabajando, sugiere la necesidad de ese hogar.
- **II9 (Baño - tenencia y uso):**
  - Media: 1.028464
  - Desviación estándar: 0.245358
  - Mínimo: 0
  - Máximo: 4
  - Comentario: Esta variable describe las condiciones de tenencia y uso del baño en el hogar, que va desde 1(Uso exclusivo del hogar), hasta 4 (No tiene baño). Es una pregunta donde el resultado o las respuestas son progresivos. Es decir, a medida que aumenta el número, peores son las condiciones de tenencia y uso del baño. La media cercana a 1 sugiere que la mayoría de los hogares tienen baño de uso exclusivo del hogar.

6) Este scatter plot nos muestra la relación entre el Ingreso Total Familiar (ITF) y el hecho de que los hogares vivan de planes o ayudas sociales (V5). La variable V5 es una variable dummy que toma valor 1 si el encuestado responde que recibe ayuda social y 2 si el encuestado responde que no recibe ayuda social. Podemos ver que los mayores niveles de ITF están sobre el 2 en el eje Y. Esto quiere decir que los hogares que reportan mayor ITF son aquellos que no utilizan ayudas sociales para vivir.
7) Cargamos y limpiamos la tabla de adult_equiv del TP3. Luego definimos una función para calcular adultos equivalentes basado en la edad y el sexo, aplicamos esta función a cada individuo en el dataframe mergeado y agregamos una columna que suma los adultos equivalentes para cada hogar.
8) Separamos los datos en dos bases distintas dependiendo de si las familias respondieron o no la pregunta sobre su ITF. Luego, para aquellas que respondieron, calculamos el ingreso necesario para no ser consideradas pobres y determinamos si cada familia es pobre en base a su ingreso total familiar comparado con el ingreso necesario.

Parte 2:

2.1. Definimos la función evalua_metodo que recibe un modelo como argumento y los datos de entrenamiento como detallados en consigna. Pedimos la accuracy y el ROC a Jupyter.

2.2. Definimos la funcion cross_validation que realiza validacion cruzada con k iteraciones.

Parte 3:

1) .
2) .
3) .
4) En validación cruzada, elegir un k muy pequeño (como 2 o 3) puede resultar en estimaciones del error del modelo que varían mucho, pero es rápido de calcular. Usar un k grande (como 10) ofrece estimaciones más estables y precisas, aunque requiere más tiempo de cómputo. Si k=n (el número de muestras), estamos en el caso de Leave-One-Out Cross-Validation (LOOCV), donde el modelo se evalúa n veces, una por cada muestra, lo cual es muy preciso pero también muy costoso en términos de tiempo de cálculo. Generalmente, un k de 10 es un buen equilibrio entre precisión y eficiencia