Hipótesis y Pregunta

**Predicción de la natalidad trimestral en Perú mediante algoritmos de regresión con datos sociodemográficos del Registro de Nacidos Vivos (2015–2025)**

**Propuesta 1**

**Hipótesis:**

Un modelo de regresión lineal múltiple, enriquecido con características temporales y sociodemográficas, puede predecir con precisión aceptable la tasa de natalidad trimestral a nivel nacional y departamental en Perú, superando enfoques basados únicamente en promedios históricos.

**Pregunta de investigación:**

¿Cómo pueden utilizarse técnicas de machine learning, específicamente modelos de regresión, para predecir la tasa de natalidad trimestral en Perú a diferentes niveles geográficos, y qué variables sociodemográficas demuestran mayor poder predictivo?

**Propuesta 2**

**Hipótesis:**Un modelo de regresión lineal que utilice la edad de la madre, su nivel educativo y la estacionalidad del año podrá predecir con precisión la cantidad de nacimientos trimestrales en Perú, siendo la edad de la madre la variable más importante.

**Pregunta:**

¿Qué combinación de variables del registro de nacidos vivos (edad, educación, ubicación, época del año) permite predecir mejor la tasa de natalidad trimestral en Perú, y cuál es el peso relativo de cada factor?

Metodología

# **METODOLOGÍA - RESUMEN EJECUTIVO**

## **📊 Fuente de Datos y Preparación**

### **Dataset Utilizado:**

* Fuente: Archivo Nacional de Datos de Colombia (ANDA) - Estadísticas Vitales EEVV
* Período: 2009 a 2019
* Registros: 642,657 nacimientos
* Variables: 23 columnas con información demográfica y de salud

### **Preprocesamiento de Datos:**

Se siguió la metodología CRISP-DM con los siguientes pasos:

1. Limpieza inicial:
   * Eliminación de columnas irrelevantes (AREANAC, SIT\_PARTO, AREA\_RES, SEG\_SOCIAL, IDPERTET)
   * Remoción de datos duplicados
   * Completado de valores faltantes con el promedio entre valores adyacentes
   * Eliminación de registros que permanecían con valores nulos
2. Transformación:
   * Normalización de datos con método Min-Max (escala 0-1)
   * Análisis exploratorio de distribuciones y patrones
   * Estudio de correlaciones entre variables

## **🔄 Método de Ensamble Propuesto**

### **Algoritmos Base:**

* Random Forest (principal)
* Árboles de Decisión
* Regresión Lineal
* k-NN (k-Nearest Neighbors)

### **Técnica de Ensamble: BAGGING**

(Bootstrap Aggregating)

#### **Implementación en 4 Pasos:**

Paso 1 - Muestreo Bootstrap:

text

Se crean L subconjuntos del dataset original:

{Z₁¹, Z₂¹, Z\_R¹}, {Z₁², Z₂², Z\_R²}, ..., {Z₁ᴸ, Z₂ᴸ, Z\_Rᴸ}

* Muestras con reemplazo
* Tamaño igual al conjunto original

Paso 2 - Modelado Individual:

text

Se entrena un modelo en cada subconjunto:

W₁, W₂, ..., W\_L → f₁(x), f₂(x), ..., f\_L(x)

Paso 3 - Aprendizaje Paralelo:

* Cada modelo aprende independientemente
* Entrenamiento concurrente en todos los subconjuntos

Paso 4 - Combinación de Predicciones:

text

Predicción final = Promedio de todas las predicciones:

f(x) = Σ f\_i(x) / L para i = 1 hasta L

## **🎯 Variables Clave Analizadas**

### **Variables Predictoras Seleccionadas:**

1. EDAD\_MADRE - Edad de la madre al parto
2. NIV\_EDUM - Nivel educativo de la madre
3. NIV\_EDUP - Nivel educativo del padre
4. EST\_CIVM - Estado civil de la madre
5. EDAD\_PADRE - Edad del padre
6. N\_HIJOSV - Número de hijos nacidos vivos

### **Variable Objetivo:**

* Número de hijos por madre (comportamiento reproductivo)

## **📈 Evaluación del Modelo**

### **Métricas de Performance:**

* R² (Coeficiente de Determinación):
  + Mide qué tan bien se ajustan los datos a la línea de regresión
  + Rango: 0 a 1 (mejor entre más cercano a 1)
* RMSE (Raíz del Error Cuadrático Medio):
  + Representa diferencia entre valores reales y predichos
  + Mejor entre más cercano a 0

### **Estrategia de Validación:**

* División 70%-30%: 70% entrenamiento, 30% testing
* Validación cruzada para estimar performance
* Comparación con modelos individuales

## **🔍 Análisis de Importancia de Variables**

### **Técnica Utilizada:**

* Importancia por Permutación
* Validación cruzada para estimar influencia
* Matriz de correlación para relaciones lineales

### **Variables Más Influyentes:**

1. Edad de la madre (42.3% de importancia)
2. Nivel educativo madre (34.9%)
3. Nivel educativo padre (10.4%)
4. Estado civil madre (4.1%)
5. Edad del padre (3.4%)

## **💡 Ventajas del Enfoque Propuesto**

### **Reducción de Varianza:**

* Bagging combina múltiples estimaciones
* Compensación de errores entre modelos
* Mayor estabilidad en predicciones

### **Mejora de Precisión:**

* 20% mejor RMSE vs árboles de decisión simples
* 30% mejor RMSE vs regresión lineal clásica
* 37-48% mejora en R² vs métodos individuales

Esta metodología demostró ser efectiva para identificar factores clave en el comportamiento reproductivo y puede servir como referencia para estudios similares en otros contextos demográficos.

# **METODOLOGÍA**

## **3.1. Fuente de Datos y Preparación**

### **3.1.1. Dataset Utilizado**

Para el desarrollo de esta investigación, se empleará el dataset "Registros de Nacidos Vivos en el Perú (2015-2025)" administrado por el Ministerio de Salud (MINSA). Este conjunto de datos contiene información detallada proveniente del Certificado de Nacido Vivo (CNV) con las siguientes características:

* Período: 2015 a 2025 (datos trimestrales)
* Variables disponibles: 22 atributos demográficos, de salud y socioculturales
* Cobertura: Nacional y desagregada por departamentos
* Actualización: Trimestral

### **3.1.2. Preprocesamiento de Datos**

Se seguirá la metodología CRISP-DM con los siguientes procesos:

Limpieza inicial:

* Eliminación de registros duplicados y inconsistentes
* Validación de rangos de valores (ej: edad materna entre 12-55 años)
* Manejo de valores faltantes mediante interpolación temporal
* Estandarización de formatos de fecha y categorías

Selección de variables:  
Se conservarán las variables más relevantes basadas en el estudio colombiano y la disponibilidad en el dataset peruano:

* FecNac\_Año, FecNac\_Mes (para series temporales)
* Edad\_Madre, Nivel\_Instrucción\_Madre
* ubiLbligeoInel (para análisis departamental)
* Estado\_Civil, Num\_embor\_madre, Hijos\_vivo\_madre

Transformación de datos:

* Agregación trimestral de nacimientos
* Codificación de variables categóricas (nivel educativo, estado civil)
* Normalización Min-Max para variables numéricas
* Creación de variables dummy para estacionalidad

## **3.2. Enfoque de Modelado**

### **3.2.1. Algoritmos Base**

Inspirado en el éxito del estudio colombiano, se implementarán los siguientes algoritmos:

Modelos Individuales:

* Regresión Lineal Múltiple: Como línea base para comparación
* Random Forest: Para capturar relaciones no lineales
* k-NN Regressor: Para patrones basados en similitud
* Decision Trees: Como componente base del ensamble

### **3.2.2. Método de Ensamble Bagging**

Se implementará la técnica de Bootstrap Aggregating siguiendo el mismo enfoque exitoso del estudio colombiano:

Paso 1: Muestreo Bootstrap

text

Creación de L subconjuntos de entrenamiento:

{D₁, D₂, ..., D\_L} donde cada D\_i se genera mediante muestreo con reemplazo del dataset original

Paso 2: Entrenamiento Paralelo

* Cada subconjunto D\_i entrena un modelo de Random Forest
* Configuración: 100 estimadores, profundidad máxima 10
* Parámetros optimizados mediante validación cruzada

Paso 3: Generación de Predicciones  
Cada modelo produce predicciones para:

* Tasa de natalidad trimestral nacional
* Tasa de natalidad trimestral por departamento

Paso 4: Combinación por Promedio

text

Predicción final = (1/L) \* Σ predicción\_modelo\_i

### **3.2.3. Adaptaciones Específicas para el Contexto Peruano**

Variables de Entrada Especializadas:

text

X = [Edad\_Madre, Nivel\_Educativo\_Madre, Trimestre, Año,

Codigo\_Departamento, Tendencia\_Temporal, Variables\_Estandarizadas]

Salidas del Modelo:

* Y\_nacional = Tasa de natalidad trimestral nacional
* Y\_departamental = Tasa de natalidad trimestral por departamento

## **3.3. Configuración Experimental**

### **3.3.1. División de Datos**

* Entrenamiento: 2015-2019 (70% del período)
* Validación: 2020-2022 (20% del período)
* Testing: 2023-2025 (10% del período)

### **3.3.2. Métricas de Evaluación**

Las mismas utilizadas en el estudio colombiano para permitir comparación:

Coeficiente de Determinación (R²):

text

R² = 1 - (Σ(y\_real - y\_pred)² / Σ(y\_real - y\_promedio)²)

* Interpretación: Proporción de varianza explicada
* Rango: 0 a 1 (mejor cerca de 1)

Raíz del Error Cuadrático Medio (RMSE):

text

RMSE = √(1/n \* Σ(y\_real - y\_pred)²)

* Interpretación: Error promedio en unidades originales
* Rango: ≥ 0 (mejor cerca de 0)

### **3.3.3. Validación**

* Validación Cruzada Temporal: TimeSeriesSplit con 5 folds
* Validación de Estabilidad: Múltiples ejecuciones con diferentes semillas
* Análisis de Residuales: Para detectar patrones no capturados

## **3.4. Análisis de Importancia de Variables**

### **3.4.1. Técnica de Importancia por Permutación**

* Permutación aleatoria de cada variable predictora
* Medición del decremento en performance
* Validación cruzada para robustez estadística

### **3.4.2. Análisis Esperado de Variables**

Basado en el estudio colombiano, se anticipa:

1. Edad de la madre (factor principal)
2. Nivel educativo materno
3. Ubicación geográfica (diferencias departamentales)
4. Factores estacionales (trimestres)

## **3.5. Ventajas del Enfoque Propuesto**

### **3.5.1. Robustez**

* Reducción de varianza mediante combinación de múltiples modelos
* Compensación de errores entre predictores individuales
* Estabilidad en predicciones a largo plazo

### **3.5.2. Adaptabilidad**

* Capacidad de capturar patrones no lineales
* Adaptación a cambios estructurales en tendencias
* Escalabilidad para análisis multinivel (nacional/departamental)

### **3.5.3. Interpretabilidad**

* Importancia de variables cuantificable
* Análisis de contribuciones marginales
* Transparencia en el proceso predictivo

## **3.6. Implementación Técnica**

### **3.6.1. Herramientas**

* Lenguaje: Python 3.9+
* Librerías: Scikit-learn, Pandas, NumPy, Matplotlib
* Procesamiento: Google Colab / Jupyter Notebooks

### **3.6.2. Flujo de Procesamiento**

text

Datos Crudos → Limpieza → Agregación Trimestral →

Entrenamiento Modelos → Ensamble Bagging →

Validación → Análisis Resultados

Esta metodología, basada en el enfoque probado exitosamente en Colombia pero adaptada al contexto y datos peruanos, permitirá desarrollar un modelo robusto y preciso para la predicción de tasas de natalidad, proporcionando así una herramienta valiosa para la planificación de políticas públicas en el Perú.

# **📋 EXPLICACIÓN DETALLADA: VALIDACIÓN DE RANGOS DE VALORES**

## **¿Qué significa "Validación de rangos de valores"?**

Es un proceso de control de calidad donde verificamos que los valores de cada variable estén dentro de límites lógicos y biológicamente posibles. Es como un "filtro de sentido común" para los datos.

## **🔍 EJEMPLO CONCRETO: EDAD MATERNA**

### **Límites propuestos: 12-55 años**

* Límite inferior (12 años): Edad biológica mínima para un embarazo viable
* Límite superior (55 años): Edad máxima común para embarazos naturales

### **¿Qué hacemos si encontramos valores fuera de rango?**

Caso 1: Edad = 8 años

* ❌ BIOLÓGICAMENTE IMPOSIBLE
* Acción: Eliminar registro (error de digitación)

Caso 2: Edad = 65 años

* ⚠️ POCO PROBABLE
* Acción: Verificar si es error o caso excepcional real

Caso 3: Edad = -1 o 999

* ❌ VALORES POR DEFECTO/ERROR
* Acción: Eliminar como dato corrupto

## **📊 OTROS EJEMPLOS EN TU DATASET**

### **Peso del recién nacido (PESO\_NACIDO)**

* Rango válido: 500 - 5,000 gramos
* Fuera de rango: 50g (inviable) o 10,000g (error)

### **Talla del recién nacido (TALLA\_NACIDO)**

* Rango válido: 30 - 60 cm
* Fuera de rango: 10 cm o 200 cm (imposible)

### **Duración del embarazo (DJ/L\_EMB\_PARTO)**

* Rango válido: 20 - 45 semanas
* Fuera de rango: 10 semanas (aborto) o 60 semanas (error)

## **🛠️ CÓMO LO IMPLEMENTARÍAS EN CÓDIGO**

python

def validar\_rangos(datos):

*# Edad materna*

datos = datos[(datos['Edad\_Madre'] >= 12) & (datos['Edad\_Madre'] <= 55)]

*# Peso del bebé*

datos = datos[(datos['PESO\_NACIDO'] >= 500) & (datos['PESO\_NACIDO'] <= 5000)]

*# Talla del bebé*

datos = datos[(datos['TALLA\_NACIDO'] >= 30) & (datos['TALLA\_NACIDO'] <= 60)]

*# Duración embarazo*

datos = datos[(datos['DJ/L\_EMB\_PARTO'] >= 20) & (datos['DJ/L\_EMB\_PARTO'] <= 45)]

return datos

*# Aplicar validación*

datos\_limpios = validar\_rangos(datos\_originales)

## **🎯 ¿POR QUÉ ES IMPORTANTE ESTE PASO?**

### **1. Calidad del Modelo**

* Datos incorrectos → Predicciones incorrectas
* Los valores extremos distorsionan los patrones reales

### **2. Validez Científica**

* Un modelo con datos biológicamente imposibles pierde credibilidad
* Garantiza que las conclusiones sean basadas en realidad

### **3. Ética de la Investigación**

* Evita sacar conclusiones de casos que no existen en la realidad
* Mantiene la integridad académica del trabajo

## **📈 EN LA PRÁCTICA PARA TU TRABAJO**

Ejemplo real que podrías encontrar:

* Registro con Edad\_Madre = 9 años → Eliminar (error)
* Registro con PESO\_NACIDO = 100g → Eliminar (inviable)
* Registro con Edad\_Madre = 48 años → Mantener (caso real posible)

Resultado esperado:

* Dataset más pequeño pero más confiable
* Modelo que aprende de patrones reales
* Conclusiones válidas y aplicables

# **🔧 EXPLICACIÓN DETALLADA: TRANSFORMACIÓN DE DATOS**

## **1. Agregación Trimestral de Nacimientos**

### **¿Qué es?**

Convertir los datos individuales de cada nacimiento en totales por trimestre.

### **¿Por qué se hace?**

* Tu objetivo es predecir tasa trimestral, no nacimientos individuales
* Reduce la cantidad de datos manejable
* Permite análisis de tendencias temporales

### **Ejemplo práctico:**

text

Datos originales:

- 15/01/2020: 1 nacimiento

- 20/01/2020: 1 nacimiento

- 10/02/2020: 1 nacimiento

- 05/03/2020: 1 nacimiento

→ Trimestre Q1-2020: 4 nacimientos

### **Implementación:**

python

*# Agrupar por año y trimestre*

datos\_trimestrales = datos.groupby(['Año', 'Trimestre']).agg({

'Nacimientos': 'count',

'Edad\_Madre': 'mean',

'Nivel\_Instruccion\_Madre': 'mode'

}).reset\_index()

## **2. Codificación de Variables Categóricas**

### **¿Qué es?**

Convertir texto en números que los algoritmos puedan entender.

### **Ejemplo: Nivel Educativo**

text

Texto → Número

"Primaria incompleta" → 1

"Primaria completa" → 2

"Secundaria incompleta" → 3

"Secundaria completa" → 4

"Superior universitaria" → 5

### **Ejemplo: Estado Civil**

text

"Soltera" → 0

"Conviviente" → 1

"Casada" → 2

"Divorciada" → 3

### **¿Por qué se hace?**

* Los modelos matemáticos solo entienden números
* Mantiene información ordinal (ej: más educación = número mayor)

## **3. Normalización Min-Max para Variables Numéricas**

### **¿Qué es?**

Escalar todas las variables numéricas al mismo rango (0 a 1).

### **Fórmula:**

text

valor\_normalizado = (valor\_original - mínimo) / (máximo - mínimo)

### **Ejemplo práctico: Edad de la madre**

text

Edades originales: [15, 20, 25, 30, 35]

Mínimo = 15, Máximo = 35

15 → (15-15)/(35-15) = 0.0

20 → (20-15)/(35-15) = 0.25

25 → (25-15)/(35-15) = 0.5

30 → (30-15)/(35-15) = 0.75

35 → (35-15)/(35-15) = 1.0

### **¿Por qué se hace?**

* Evita que variables con rangos grandes dominen el modelo
* Mejora la convergencia de los algoritmos
* Ejemplo: La edad (15-45) y el número de hijos (0-10) quedan en la misma escala

## **4. Creación de Variables Dummy para Estacionalidad**

### **¿Qué es?**

Crear variables "bandera" que indiquen el trimestre.

### **Ejemplo práctico:**

Para el primer trimestre (Ene-Feb-Mar):

text

Q1 = 1, Q2 = 0, Q3 = 0, Q4 = 0

Para el segundo trimestre (Abr-May-Jun):

text

Q1 = 0, Q2 = 1, Q3 = 0, Q4 = 0

### **Tabla resultante:**

| Año | Trimestre | Q1 | Q2 | Q3 | Q4 | Nacimientos |
| --- | --- | --- | --- | --- | --- | --- |
| 2020 | 1 | 1 | 0 | 0 | 0 | 15000 |
| 2020 | 2 | 0 | 1 | 0 | 0 | 14500 |
| 2020 | 3 | 0 | 0 | 1 | 0 | 14800 |
| 2020 | 4 | 0 | 0 | 0 | 1 | 15200 |

### **¿Por qué se hace?**

* Captura patrones estacionales (ej: más nacimientos en ciertos trimestres)
* Los modelos pueden aprender que el Q4 suele tener más/menos nacimientos
* Ejemplo: En muchos países hay "baby boom" en septiembre (concebidos en navidad)

## **🔄 FLUJO COMPLETO DE TRANSFORMACIÓN**

### **Datos Originales → Datos Listos para el Modelo**

Paso 1: Agrupar por trimestre  
Paso 2: Codificar educación y estado civil  
Paso 3: Normalizar edad, número de hijos, etc.  
Paso 4: Añadir columnas dummy para trimestres  
Paso 5: ¡Listo para entrenar el modelo!

### **Resultado Final:**

Un dataset donde cada fila es un trimestre con:

* Variables numéricas normalizadas (0-1)
* Variables categóricas codificadas (números)
* Indicadores de estacionalidad (dummy variables)
* Número total de nacimientos (variable objetivo)

## **💡 ANALOGÍA SENCILLA**

Imagina que estás cocinando:

* Agregación trimestral = Medir ingredientes por tazas, no por gramos individuales
* Codificación = Traducir receta de inglés a español
* Normalización = Usar la misma taza para medir todos los ingredientes
* Variables dummy = Etiquetar ingredientes por temporada (verano/invierno)

¿Te queda claro cada transformación? ¿Quieres que profundice en alguna en particular?

# **🎯 EXPLICACIÓN SENCILLA: VARIABLES DUMMY PARA ESTACIONALIDAD**

## **¿QUÉ SON LAS "VARIABLES DUMMY"?**

Imagina que son como interruptores de luz ⚡ para cada trimestre:

* ON (1) = Sí es este trimestre
* OFF (0) = No es este trimestre

## **📅 EJEMPLO MUY PRÁCTICO**

### **Sin variables dummy:**

| Año | Trimestre | Nacimientos |
| --- | --- | --- |
| 2020 | 1 | 15,000 |
| 2020 | 2 | 14,500 |
| 2020 | 3 | 14,800 |
| 2020 | 4 | 15,200 |

Problema: El modelo ve "Trimestre = 2" pero no entiende que es una categoría, piensa que es un número (¡que 2 es el doble que 1!)

### **CON variables dummy:**

| Año | Trimestre | Q1 | Q2 | Q3 | Q4 | Nacimientos |
| --- | --- | --- | --- | --- | --- | --- |
| 2020 | 1 | 1 | 0 | 0 | 0 | 15,000 |
| 2020 | 2 | 0 | 1 | 0 | 0 | 14,500 |
| 2020 | 3 | 0 | 0 | 1 | 0 | 14,800 |
| 2020 | 4 | 0 | 0 | 0 | 1 | 15,200 |

Ahora el modelo entiende:

* "Q1=1, otros=0" → Es el primer trimestre
* "Q2=1, otros=0" → Es el segundo trimestre
* etc.

## **🤔 ¿POR QUÉ HACER ESTO?**

### **Para capturar patrones como:**

* "En el primer trimestre (ene-mar) siempre nacen más bebés"
* "En el tercer trimestre (jul-sep) hay menos nacimientos"

### **Ejemplo real:**

Si el modelo descubre que:

* Q4 suele tener +5% de nacimientos
* Q2 suele tener -3% de nacimientos

Entonces puede predecir: "Para 2025-Q4, esperamos 5% más nacimientos que el promedio"

## **🔧 CÓMO SE CONSTRUYEN PASO A PASO**

### **Paso 1: Identificar el trimestre de cada fecha**

text

Fecha: 15/03/2020 → Trimestre 1 (Ene-Feb-Mar)

Fecha: 20/06/2020 → Trimestre 2 (Abr-May-Jun)

Fecha: 10/09/2020 → Trimestre 3 (Jul-Ago-Sep)

Fecha: 05/12/2020 → Trimestre 4 (Oct-Nov-Dic)

### **Paso 2: Crear 4 columnas nuevas (Q1, Q2, Q3, Q4)**

### **Paso 3: Poner 1 en la columna correcta**

text

Si Trimestre = 1 → Q1=1, Q2=0, Q3=0, Q4=0

Si Trimestre = 2 → Q1=0, Q2=1, Q3=0, Q4=0

Si Trimestre = 3 → Q1=0, Q2=0, Q3=1, Q4=0

Si Trimestre = 4 → Q1=0, Q2=0, Q3=0, Q4=1

## **🎪 ANALOGÍA: CARNÉ DE IDENTIDAD**

Piensa en las variables dummy como el carné de identidad de cada trimestre:

Carné del Trimestre 1:

* ¿Eres Q1? SÍ ✅ (1)
* ¿Eres Q2? NO ❌ (0)
* ¿Eres Q3? NO ❌ (0)
* ¿Eres Q4? NO ❌ (0)

Carné del Trimestre 2:

* ¿Eres Q1? NO ❌ (0)
* ¿Eres Q2? SÍ ✅ (1)
* ¿Eres Q3? NO ❌ (0)
* ¿Eres Q4? NO ❌ (0)

## **💡 EN TU PROYECTO PERUANO**

### **Patrones que podrías descubrir:**

* Q1 (ene-mar): ¿Aumento por "baby boom" de concepción en fiestas patrias?
* Q3 (jul-sep): ¿Disminución por temporada de frío?
* Q4 (oct-dic): ¿Aumento por concepción en meses festivos?

### **Resultado final:**

Tu modelo aprenderá automáticamente:

* "Cuando Q1=1, suma X nacimientos al pronóstico"
* "Cuando Q3=1, resta Y nacimientos al pronóstico"

## **📝 RESUMEN EN 1 FRASE**

"Variables dummy = Crear 4 banderas (Q1,Q2,Q3,Q4) donde solo una está activa por trimestre, para que el modelo aprenda los patrones de cada época del año."

# **🔍 RELACIÓN ENTRE LOS MODELOS: EXPLICACIÓN CLARA**

## **✅ SON LOS MISMOS MODELOS, SOLO CAMBIA LA TERMINOLOGÍA**

## **📊 TABLA DE EQUIVALENCIAS**

| Modelos que yo mencioné | Modelos que tú mencionas | ¿Son iguales? |
| --- | --- | --- |
| Regresión Lineal | Regresión lineal | SÍ, exactamente el mismo |
| k-NN Regressor | Regresión KNN | SÍ, mismo algoritmo |
| Decision Trees | Árboles de regresión | SÍ, mismo concepto |
| Random Forest | (No mencionado) | Es un ENSAMBLE de árboles |

## **🎯 EXPLICACIÓN DETALLADA POR MODELO**

### **1. REGRESIÓN LINEAL = Regresión lineal**

* Exactamente el mismo algoritmo
* Función: Encontrar línea recta que mejor se ajuste a los datos
* Tu proyecto: Predecir tasa natalidad como combinación lineal de variables

### **2. k-NN REGRESSOR = Regresión KNN**

* Exactamente el mismo algoritmo
* Función: Predecir basado en "vecinos más cercanos"
* Ejemplo: Si trimestres similares (mismas características) tuvieron X nacimientos, este trimestre probablemente también

### **3. DECISION TREES = Árboles de regresión**

* Exactamente el mismo concepto
* Función: Crear "reglas de decisión" en forma de árbol
* Ejemplo:
  + ¿Edad madre > 30? → Sí → ¿Nivel educativo > 3? → Sí → Probable 1 hijo

### **4. RANDOM FOREST**

* NO es igual a Árboles de regresión
* Es un AVANCE: Combina MUCHOS árboles de regresión
* Ventaja: Más preciso y estable que un solo árbol

## **🌳 ANALOGÍA: CONSTRUCCIÓN DE UNA CASA**

* Árbol de regresión = Un solo arquitecto haciendo planos
* Random Forest = Un equipo de 100 arquitectos, cada uno hace planos y votan el mejor diseño
* Regresión lineal = Usar solo reglas matemáticas simples
* KNN = Copiar el diseño de casas similares en el vecindario

## **🔧 EN TU PROYECTO PRÁCTICO**

### **Flujo recomendado:**

1. Empezar con Regresión Lineal (más simple)
2. Probar Árbol de Regresión (captura no linealidades)
3. Usar KNN (para patrones estacionales)
4. FINALMENTE: Random Forest (combinación poderosa de todos los árboles)

### **Ejemplo de resultados esperados:**

text

Modelo Precisión (R²)

Regresión Lineal 0.65

Árbol Regresión 0.72

KNN 0.68

Random Forest 0.85 ← ¡MEJOR!

## **💡 RESUMEN FINAL**

SÍ, estás usando los mismos modelos que mencioné, solo que:

* "Decision Trees" = "Árboles de regresión" (mismo)
* "k-NN Regressor" = "Regresión KNN" (mismo)
* "Regresión Lineal" = "Regresión lineal" (mismo)
* Random Forest es la EVOLUCIÓN de los árboles de regresión

¡Tu selección de modelos es perfecta para el proyecto! 🎉

# **METODOLOGÍA ACTUALIZADA**

## **3.1. Fuente de Datos y Preparación**

### **3.1.1. Dataset Utilizado**

Se empleará el dataset "Registros de Nacidos Vivos en el Perú (2015-2025)" del Ministerio de Salud (MINSA) con las siguientes características:

* Período: 2015 a 2025 (datos trimestrales)
* Cobertura: Nacional y desagregada por departamentos
* Variables clave: 22 atributos demográficos y socioculturales

### **3.1.2. Preprocesamiento de Datos**

Siguiendo la metodología CRISP-DM:

Limpieza y Validación:

* Eliminación de registros duplicados e inconsistentes
* Validación de rangos biológicamente posibles:
  + Edad materna: 12-55 años
  + Peso nacido: 500-5,000 gramos
  + Talla nacido: 30-60 cm
* Manejo de valores faltantes mediante interpolación temporal

Selección de Variables:

* FecNac\_Año, FecNac\_Mes (series temporales)
* Edad\_Madre, Nivel\_Instrucción\_Madre (predictores clave)
* Estado\_Civil, Hijos\_vivo\_madre (factores sociodemográficos)
* ubiLbligeoInel (análisis departamental)

## **3.2. Transformación de Datos**

### **3.2.1. Agregación Trimestral**

Conversión de datos individuales a totales por trimestre:

python

*# Agrupar por trimestre y departamento*

datos\_trimestrales = datos.groupby(['Año', 'Trimestre', 'Departamento']).agg({

'Nacimientos': 'count',

'Edad\_Madre': 'mean',

'Nivel\_Instruccion\_Madre': 'mode'

}).reset\_index()

### **3.2.2. Codificación de Variables Categóricas**

* Nivel educativo: Preescolar=1, Primaria=2, Secundaria=3, Superior=4
* Estado civil: Soltera=1, Conviviente=2, Casada=3, Otros=4

### **3.2.3. Normalización Min-Max**

Escalado de variables numéricas al rango [0,1]:

text

Edad\_normalizada = (Edad\_actual - 12) / (55 - 12)

### **3.2.4. Variables Dummy para Estacionalidad**

Creación de indicadores por trimestre:

| Trimestre | Q1 | Q2 | Q3 | Q4 |
| --- | --- | --- | --- | --- |
| 1 (Ene-Mar) | 1 | 0 | 0 | 0 |
| 2 (Abr-Jun) | 0 | 1 | 0 | 0 |
| 3 (Jul-Sep) | 0 | 0 | 1 | 0 |
| 4 (Oct-Dic) | 0 | 0 | 0 | 1 |

## **3.3. Modelos de Machine Learning Seleccionados**

### **3.3.1. Regresión Lineal Múltiple**

Objetivo: Establecer línea base y capturar relaciones lineales

Características:

* Modelo más simple e interpretable
* Ecuación: Y = β₀ + β₁X₁ + β₂X₂ + ... + βₙXₙ
* Donde Y = Tasa de natalidad, Xᵢ = variables predictoras

Ventajas para tu proyecto:

* Resultados fáciles de explicar
* Bajo costo computacional
* Base para comparar modelos complejos

### **3.3.2. k-NN Regressor (k-Vecinos Más Cercanos)**

Objetivo: Capturar patrones basados en similitud temporal

Características:

* k = 5 vecinos (optimizado por validación cruzada)
* Métrica de distancia: Euclidiana
* Busca trimestres históricos con características similares

Aplicación en tu proyecto:

* Si un trimestre tiene características similares a trimestres pasados que tuvieron alta natalidad, predice alta natalidad
* Captura patrones estacionales no lineales

### **3.3.3. Árboles de Decisión (Decision Trees)**

Objetivo: Modelar relaciones no lineales y complejas

Características:

* Profundidad máxima: 10 niveles
* Criterio de división: Reducción de varianza
* Parada temprana para evitar sobreajuste

Ejemplo de reglas que puede aprender:

text

SI Edad\_Madre > 30

Y Nivel\_Educativo > 3 (Superior)

Y Trimestre = 4

ENTONCES → Tasa\_Natalidad = X

## **3.4. Configuración Experimental**

### **3.4.1. División Temporal de Datos**

* Entrenamiento: 2015-2019 (5 años - 70%)
* Validación: 2020-2022 (3 años - 20%)
* Testing: 2023-2025 (3 años - 10%)

### **3.4.2. Métricas de Evaluación**

Coeficiente de Determinación (R²):

* Mide qué porcentaje de la variación es explicado por el modelo
* Rango: 0 a 1 (mejor cerca de 1)

Raíz del Error Cuadrático Medio (RMSE):

* Error promedio en unidades de tasa de natalidad
* Menor valor indica mejor precisión

### **3.4.3. Estrategia de Validación**

* Validación Cruzada Temporal: 5 folds
* Validación de Estabilidad: 3 ejecuciones con diferentes semillas
* Análisis de Residuales: Para detectar patrones no capturados

## **3.5. Proceso de Entrenamiento y Evaluación**

### **3.5.1. Entrenamiento Individual**

Cada modelo se entrena por separado con los mismos datos:

python

*# Ejemplo de flujo*

modelo\_lineal.fit(X\_entrenamiento, y\_entrenamiento)

modelo\_knn.fit(X\_entrenamiento, y\_entrenamiento)

modelo\_arbol.fit(X\_entrenamiento, y\_entrenamiento)

### **3.5.2. Evaluación Comparativa**

Tabla de resultados esperada:

| Modelo | R² | RMSE | Ventajas |
| --- | --- | --- | --- |
| Regresión Lineal | 0.65 | 0.45 | Interpretable |
| k-NN | 0.72 | 0.38 | Captura estacionalidad |
| Árbol Decisión | 0.75 | 0.35 | Captura no linealidades |

### **3.5.3. Análisis de Importancia de Variables**

Solo para Árboles de Decisión:

* Método: Importancia por reducción de impureza
* Resultado: Ranking de variables más influyentes
* Aplicación: Identificar factores clave en la natalidad peruana

## **3.6. Implementación Técnica**

### **3.6.1. Herramientas**

* Lenguaje: Python 3.9+
* Librerías: Scikit-learn, Pandas, NumPy, Matplotlib
* Entorno: Google Colab

### **3.6.2. Flujo de Procesamiento**

text

Datos Crudos

↓

Limpieza y Validación

↓

Agregación Trimestral

↓

Transformación (Codificación + Normalización)

↓

Entrenamiento 3 Modelos

↓

Evaluación Comparativa

↓

Análisis de Resultados

## **3.7. Ventajas de Esta Selección de Modelos**

### **3.7.1. Diversidad de Enfoques**

* Lineal: Relaciones simples y globales
* k-NN: Patrones locales y de similitud
* Árbol: Relaciones complejas y no lineales

### **3.7.2. Balance Complejidad-Interpretabilidad**

* Todos los modelos son interpretables
* Permiten entender qué factores afectan la natalidad
* Resultados comunicables a no técnicos

### **3.7.3. Adecuación al Problema**

* Capturan tanto tendencias temporales como factores sociodemográficos
* Apropiados para el volumen de datos disponible
* Computacionalmente eficientes

Vigente

# Datasets

1. Registros de Nacidos Vivos en el Perú (2015–2025)

<https://datosabiertos.gob.pe/dataset/registros-de-nacidos-vivos-en-el-per%C3%BA-2015%E2%80%932025>

1. Ubigeo de Perú (INEI)

<https://account.geodir.co/recursos/ubigeo-inei-peru.html>

1. Geo Perú

<https://visor.geoperu.gob.pe/>

# Publicaciones Científicas

1. Predicting Future Birth Rates with the Use of an Adaptive Machine Learning Algorithm: A Forecasting Experiment for Scotland

<https://pmc.ncbi.nlm.nih.gov/articles/PMC11276648/>

[Predicting Future Birth Rates with the Use of an Adaptive Machine Learning Algorithm: A Forecasting Experiment for Scotland](https://pmc.ncbi.nlm.nih.gov/articles/PMC11276648/pdf/ijerph-21-00841.pdf) -> pdf directo

[Prompt de consulta](#26h2ze6h5mbq) (Por qué deberíamos incluirlo según Gemini)

M. Tzitiridou-Chatzopoulou, G. Zournatzidou, y M. Kourakos, "Predicting Future Birth Rates with the Use of an Adaptive Machine Learning Algorithm: A Forecasting Experiment for Scotland", *International Journal of Environmental Research and Public Health*, vol. 21, no. 7, p. 841, junio de 2024.

1. Análisis de reducción de tasas de natalidad en Colombia: un enfoque con técnicas de aprendizaje automático

<https://hdl.handle.net/10495/44526>

L. V. Largo Quintero y S. Valencia Cadena, "Análisis de reducción de tasas de natalidad en Colombia: un enfoque con técnicas de aprendizaje automático", Tesis de especialización, Universidad de Antioquia, 2024.

1. Discovering Mexican Birth Rate Patterns Using Machine Learning Techniques

<https://www.researchgate.net/profile/Ivo-Pineda/publication/324987471_Discovering_Mexican_Birth_Rate_Patterns_Using_Machine_Learning_Techniques/links/5d0447e792851c90043c8848/Discovering-Mexican-Birth-Rate-Patterns-Using-Machine-Learning-Techniques.pdf>

Somodevilla, M., Limón, D., Pineda, I., de Celis, C. P., & Vilariño, D. Discovering Mexican Birth Rate Patterns Using Machine Learning Techniques.

1. Reducción de la tasa de natalidad en el Perú

<https://www.bcrp.gob.pe/docs/Publicaciones/Revista-Moneda/moneda-197/moneda-197-10.pdf>

Revista Moneda n° 197: <https://www.bcrp.gob.pe/publicaciones/revista-moneda/revista-moneda-197.html>

1. Predicción de factores clave en el aumento de la demografía en Colombia a través del ensamble de modelos de Machine Learning <https://doi.org/10.14483/23448350.19205>

Ejemplo:

1. Victor O.K. Li, "Hints on writing technical papers and making presentations", IEEE Transactions on Education, vol. 42, no. 2, pp. 134-137, mayo de 1999.

<https://repositorio.uns.edu.pe/handle/20.500.14278/4455>  
<https://repository.universidadean.edu.co/entities/publication/9a64e149-086b-443f-9f83-772b29bf30c4>

<https://www.scielosp.org/article/ssm/content/raw/?resource_ssm_path=/media/assets/rpsp/v14n1/16628.pdf>

<https://repository.ucc.edu.co/entities/publication/c91a6ba5-5dad-43a2-ad7b-70fa9ec1891a>

<https://go.gale.com/ps/i.do?p=IFME&sw=w&issn=&v=2.1&it=r&id=GALE%7CA637988656&sid=googleScholar&linkaccess=abs&userGroupName=anon~280884ca&aty=open-web-entry>

# Material extra

1. Modelos Predictivos (Freddy Hernández) 2024-12-06: <https://fhernanb.github.io/libro_mod_pred/>

Asesorías

Asesoría 1 - Villanueva

# Transcripción

(Transcrito por TurboScribe.ai. Actualizar a Ilimitado para eliminar este mensaje.)

Más proximidad, más homogeneidad, pero a veces no, ¿no? Incluso aquí en Lima, te vas de San Isidro a Los Conos y no encuentras la realidad. Y las tasas de natalidad que se pueden calcular con esa lata también varían, ¿no? Entonces, tal vez lo que, no sé, se le ocurre al Estado, le interesa saber cuál es la tasa, o la tasa o cuántos nacimientos per cápita va a haber el siguiente mes, por ejemplo. ¿Por qué? Porque me preveo sobre la asistencia a los sistemas de salud.

Entonces, acá encuentro que en este distrito de acá va a haber, no sé, 100 nacimientos por cada 10.000 pobladores. Una tasa, ¿no? No digo tasa porque, claro, cada distrito puede tener diferente tamaño. Entonces, más nacimientos, menos nacimientos.

Pero cuando yo lo normalizo con respecto al per cápita, la población, ya se hace uniforme, ¿no?, comparable, ¿no?, entre distritos. Por ejemplo, a mí me interesaría, o sea, pienso, ¿no?, que al Estado le interesaría saber cuántos nacimientos por cada 1.000 habitantes o 10.000 habitantes va a haber en el distrito. Para el siguiente mes.

Y qué entraría, ¿no? Es decir, tengo un modelo que lo que predice es la tasa de nacimientos, ¿esto? ¿Qué porcentaje, no? ¿Qué porcentaje? No sé, o sea, algo que sea una mediana organizada. O puede ser el mismo valor. Sí, justo acá también está.

Tiene lo que es el ubiqueo. Claro, por eso lo digo, porque ya lo vi. Sí, sí, y lo podemos ubicar acá, por ejemplo, acá encontré una página del ubiqueo del mes.

Sí, podrías hacer agregación, por eso podrías agregar la cantidad de nacimientos mes a mes de cada distrito. Para cada distrito tú puedes hacer una serie temporal, mes a mes, ¿no? Por ejemplo, tienes 5 años desde el 2000, ¿qué?, ¿2017? 2015. 2015, mes 1. Este es 2015, mes 2. Yo digo el mes para… Podría ser año, pero estás acá con muy poca edad.

Unos días vas a tener mucha edad y no vas a poder procesar tanto. El mes me parece algo adecuado. Además, es algo que a las agencias gobiernos les podría interesar, es decir, un mes adelante.

¿Cuánto van a ser en todos los distritos del Perú? Tal vez focalizar en una región. No son 1.800 distritos, pero podría focalizar en una región también. Sí, dime.

Sí, justamente ahorita mencionando en enfocarlo. ¿Usted nos recomendaría centralizarnos en un departamento y sus distritos o en todo a nivel peruano? Podría ser en todo, la data exige para todo el país. O podría ser algo representativo, Aragosta, Sierra y Cero.

Son 1.800 distritos. Lo que nos recomienda más o menos es básicamente ser un modelo proyectivo para decir lo que ya estás en natalidad como parte de preparación para que el gobierno se anteponga. Más o menos cada mes van a haber tantos nacimientos y así se prepara el Estado, más o menos así se prefiere.

Predecir el valor, por ejemplo, este distrito es de Pueblo Libre. Entonces yo tengo mes a mes el valor de la tasa de nacimiento. O la cantidad de nacimientos, si quieres, en absoluto.

Pero yo pensaría que debe ser un modelo general para todos los distritos. Entonces, ¿qué es lo que entra? Lo que entra serían las tasas de nacimiento de los meses previos. Una ventana temporal, no toda la data previa.

Puede ser el último año o los últimos dos años. Esto lo voy a... porque hay otro grupo que me ha consultado que quiere hacer pronóstico. Yo pronostico un valor futuro.

El valor siguiente, ¿cuál es? Esto se puede hacer con las técnicas que vamos a estudiar en el curso de Machine Learning. Uno de los atributos que puede entrar son los valores previos. El valor del mes anterior, del mes tras anterior y así.

12 valores previos puede ser. Tú puedes aumentar esa data con características del lugar. Porque ¿qué influye en la tasa de natalidad? Además de la tendencia, o sea, de los valores previos, se ha estado subiendo, me enteraría que va a seguir subiendo.

Pero además de esos valores, también influye cuestiones locales. Por ejemplo, crecimiento económico, variables de pobreza, desigualdad de educación. La educación puede influir.

La infraestructura que tenga, ¿qué se lo puede sacar de Geovisor? El índice de desarrollo humano. Si uno puede enriquecer los datos, además de los valores anteriores, históricos, puede enriquecer con variables locales que la puede sacar de otra fuente, como Geovisor. Y podrías hacer el ensayo, hacer un modelo solo con los valores previos.

Es decir, sea cual sea el distrito, meto sus valores previos y mi predice los va a seguir. Es decir, me sigue. Y adicionalmente puedo hacer otro experimento donde además de sus valores previos, meto información local.

Índice de desarrollo humano, índice de desigualdad. Hay un montón de mapas del distrito. Y evalúan qué tan bien se predice el valor del mes siguiente de la tasa de natalidad.

Lo pueden hacer con todos los distritos o tal vez con un subconjunto. El distrito de Perú hay 1800, probablemente no hay data en todo. De repente ahí filtras la data a grupas.

Cuántos nacimientos hay por cada distrito y filtras los siemas a datos. Capaz se podría manejar por provincia, por departamento, pero eso ya es muy amplio. Puede ser por departamento también.

Porque en un departamento como Lima hay diligencia gráfica. Te vas de fuera de Lima y es otra realidad. Pero ahí no justifican porque saber, y de hecho debe haber trabajos ahí, tienen que investigar.

Sobre pronosticar la tasa de nacimiento, tasa de natalidad futura, qué utilidades tienen. Cualquier paper que encuentre haciendo eso va a justificar. Si es importante para las agencias de gobierno.

Y los papers pueden ser de cualquier lado, internacionales. Porque el paper te va a servir, uno, para justificar el problema. Y otra para tener ideas de la metodología.

Sobre todo es cómo es que ellos han enfocado el pronóstico futuro. Qué tipo de modelo, qué variables han usado. Tanto locales como históricas.

De repente pueden usar temporada de lluvia. Bueno, no creo que influya. Pero no sé, variables que de repente no estamos pensando.

Y que desde ideas de repente consiguen esa data que ellos han usado. Es un ensayo. Al final esto es para aprendizaje.

Pero tiene cierto potencial de que esto también sea publicado. Depende de cómo lo enfoquen y los modelos que usen. Ahí el procesamiento de la data es clave.

El 80% del resultado final está en cómo trabaja la data. Yo en clase les voy a dar más o menos para hacer pronóstico cuando estemos viendo modelos de regresión. Cómo pueden hacer un modelo normal, no hecho para pronóstico.

Hacerlo para pronóstico. Básicamente trabaja en la data misma. Cómo acomoda la data.

Y cualquier modelo de los que vamos a estudiar o otros que vean, se puede aplicar. El trabajo más grande está en conseguir la data y adecuarla. Y el segundo ya es en entrenar los modelos y llevarlo a los bienes.

Todo parte de sus preguntas. Tu pregunta es, por ejemplo, una pregunta para esto que se ha hablado. ¿Cuánto influyen las variables locales en el pronóstico de la tasa de natalidad? Puede ser que no influya nada.

Puede ser que toda la información para pronosticar el valor siguiente esté en el histórico. De la misma, entonces no necesitas información local. O puede ser que sí, que sea decisivo que la data local que están usando haya aumentado fuertemente la precisión del pronóstico.

Si esa es tu pregunta, tu respuesta sería influir, influir. ¿Cuáles variables influyen? Por eso ahí en el formato se les pide, coloquen sus preguntas. No pueden ser 100, tienen que ser 3 o 4 más.

La otra sería qué tipo de modelos procesa mejor la data, hace mejor pronóstico. Y otras cosas, pueden hacer pronósticos. De repente se les ocurre, vamos a comparar el pronóstico distrital a departamental.

El distrital no se predice nada, pero el departamental depende de sí. Depende de las preguntas que ustedes se planteen. Pero con esa data puede sacar un montón de cosas.

Sí, me parece que está muy buena la data, bastante completa. Y es un buen ejercicio para ir haciendo ese modelo. Entonces básicamente sería con esto, ustedes nos recomiendan manejarlo mensual, más que anual.

Es lo que se me ocurre ahora. ¿Por qué anual no? Si tienes 6 años, ¿cuántos son? 10. Vas a tener 10 puntos temporales, uno por cada año.

Lo cual va a ser, no es que no se pueda, porque vas a tener 10 puntos temporales, pero va a depender de la agregación espacial que hagas. Si la agregación espacial es a nivel de distrito, en cada distrito vas a tener una serie temporal de 10 años. Y en teoría tú podrías, es como que tuvieses una serie temporal mucho más larga.

Le podrías sacar el jugo, la vuelta al problema. Pero de todas maneras tienes que reservar cierta data para validar. Porque si ustedes hacen un modelo que te dice, te va a pronosticar cuál es la tasa del siguiente año, que también tiene utilidad, incluso a veces hasta más útil para el Ministerio de Planificación a largo plazo.

Pero, ¿cómo yo sé que el modelo funciona bien? Entonces, lo que ustedes van a hacer en series temporales es, bueno, al menos se tendrían que reservar algunos años para probar. Tienen data de 2015 a 2025, digamos, que se reserva en el 2022, 23, 24 y 25. 23, 24 y 25.

Los tres últimos años. Entonces, van a entrenar el modelo con data de 2015 a 2022. Los van a entrenar para que pronostique el siguiente año.

Pero, ¿cómo van a evaluar ese modelo? Que también funciona. Lo van a evaluar en el 2023, 24 y 25. Ustedes saben qué pasó.

Y ahí calcula las métricas finales de pronóstico. Entonces, de esos 10 puntos ya descarta al menos 13, poco. Ya digamos que sea 3. Descarta 3 puntos.

Solo te queda 7. Puede que sea, tal vez, un intermedio, si mensuales mucho, que son 12 por 10, 120 puntos. Tal vez puedas hacer trimestrales. Un año, 4 puntos.

Un intermedio, ¿no? Pronosticar el trimestre. La tasa trimestral. Esa.

Me parece que sí, porque como acá también están los meses, es lo bueno. Los meses, el año en que nace, el objeto, hasta su peso. Ahora, claro, esto está diciendo pronosticar la cantidad.

No, sí, eso me parece bueno porque... Pero no sé si hay algo relativo a salud, enfermedad. Lo único que sale es de que en anteriores nacimientos, o sea, se ha tenido... ¿Hijos, nacimientos muertos o abortos? Pero es parte de la madre. Eso podría ser otra, ¿no? El pronóstico de nacimientos muertos, no vivos.

Sí, pero el tema es que como acá lo maneja, o sea, es todo juntos, ¿no? Nacimientos muertos, aunque como que sí. Pero no nos enfocamos mucho en eso porque muchos son ninguno, ninguno, ninguno. Más que todo tenemos como data que puede haber un antecedente, pero no pensamos usarlo porque no lo consideramos tan relevante.

A usted sí nos recomendaría usarlo para el planteamiento del modelo. Depende, depende de lo que quieras hacer. Podría ser otra idea.

Yo lo he mencionado al inicio, ¿no? Nos interesa... Nacimientos de madres adolescentes. Claro. Podrían calcular también, en vez del porcentaje de nacimientos sobre el total, porcentaje de nacimientos de adolescentes sobre el total de nacimientos.

O sea, si yo digo, mira, el próximo trimestre van a nacer en este distrito, en esta provincia, el 20% de los nacimientos de todo el distrito van a ser de jóvenes adolescentes, menores de 18 años. Podrían, en vez de pronosticar la tasa total, centrada en una población en particular, de jóvenes adolescentes. O podrían pronosticar, como yo les digo, nacidos no vivos.

O la tasa de abortos. Lo que quieran. O sea, tiene que tener una utilidad y una justificativa.

Yo creo que la natalidad estaría bien más que todo porque le ayuda al Estado a prepararse a lo que son recursos, lo que mencionó el profesor. Como manejarlo trimestral, tal vez, en vez de mensual, como que por paquete, cada trimestre. Sí, como menciona mi compañera, más que todo ahorita vamos a centrarlo, como usted lo menciona, el tema de manejarlo mensualmente y el tema de la población general en departamentos.

Sí, eso ahorita nos ha dado una mejor idea de cómo plantearlo. También coincide la idea de eso que han dicho. Ahora que estoy pensándolo, también podría ser interesante.

Podrían ser dos preguntas. Pronosticar la tasa total y la tasa de adolescentes. El curso de madero podría tener dos salidas.

Pronosticar la tasa total y la tasa relativa de la población de adolescentes que vienen de adolescente. Pero porque es otra dirección de los recursos del Estado. Pueden llegar a consumir otro tipo o más recursos porque no es una persona capaz y además, claro, no solo son recursos de salud, sino que es potencialmente una persona que temporalmente va a ser un costo mayor para el Estado, para su familia, porque no es productiva durante el tiempo.

Eso tiene otra connotación. Una consulta, profesor. Referente a como menciona hacer esta doble salida, más allá de que si sea correcto o no el planteamiento, con tema de tiempo, ¿usted cree que sería muy tedioso? Más que todo, no quisiéramos tampoco plantearlo en esta doble salida, pero luego al último momento estemos muy atareados.

No, no me da idea. No, claro, pero me refiero a que ¿usted lo ve muy tedioso realizar? Para mí no. Porque si tú lo haces con redes neuronales, es exactamente lo mismo.

Porque al final las redes neuronales te piden un valor, un vector. Es cuestión de acomodar la data. Para modelos más simples que tengamos estudiados al inicio, eso sí, hay que tener un poco más de cuidado.

Entonces, por ahora les recomendaría escoger un target. Solo un objetivo. Y háganlo bien, inspeccionen todo lo que quieran responder.

Si les da tiempo. Claro, o sea, si no lo ponen en el trabajo, no están obligados a hacerlo. Claro, creo que podría ser como un punto a considerar, que podemos agregar otros puntos en la predicción de la natalidad.

Y ya si nos da tiempo a predicción de la natalidad, ¿el cuánto porcentaje de esa natalidad va a corresponder a madres menores de edad? Porque lo que cambia, incluso si hacen un modelo con una sola salida, lo que cambiaría entre predecir la tasa total de natalidad o la tasa de natalidad de las madres adolescentes es básicamente la edad. El piper va a ser exactamente el mismo. Volver a correo con la data, cambiarle el target.

O sea, eso podrían hacer. Podrían hacer el resultado de la predicción total de natalidad y la predicción de edad de la tasa de adolescencia. Sí.

Profe, una consulta. En todo caso, como la entrega es el planteamiento recién, si es que hacemos el planteamiento de un target, pero para la segunda entrega quisiéramos colocar ese segundo target para tener esa doble salida, ¿se podría modificar? Porque entendería, al menos según lo que dice el informe, o sea, la rúbrica, es de que tengo que basarme en la primera entrega para presentar la segunda. O sea, si modificarle o agregarle cosas.

No sé si se podría. Me parece que lo vuelve a revisar. Sí, se vuelve a revisar.

Ah, ok, listo. La idea de la primera entrega es darte un feedback. Ah, ok.

Ha habido casos, pero no les recomiendo que cambien totalmente el tema, ¿no? Ah, claro. Porque no aprovechas todo lo que se te va a dar, ¿no? Pero sí se puede. O sea, hay margen, ¿no? Estamos creyendo.

Ah, sí, claro. Más que todo para evitar ese escenario de estas preguntas, ¿no? Sí, sí, sí. Entonces, eso sería mi pregunta.

Sí, profe, y una consulta. Como los datos son desde 2015 hasta 2025, ¿no habría ningún problema con la pandemia? Porque, por ejemplo, si dejamos esos años como para probar el modelo, capaz no sea tan exacto. Sí, o sea, es la realidad, ¿no? Claro.

Va a involucrar un periodo atípico. Dos, tres años atípicos. ¿Va a perturbar el modelo? Sí.

Podría preguntarse también, ¿no? ¿Cuánto influye la pandemia? Claro. ¿Colocan la data de la pandemia o la sacan? Hacen un experimento considerando los años de la pandemia y desconsiderando los años. Claro.

Pero, o sea, al principio hagan un modelo simple. O sea, vayan de lo simple a lo complejo. Claro.

Empiecen con el cumpleaños y nos perdemos ahí. Sí, sí. Entonces, podrían hacer experimentos de no considerar.

De repente eso mejora parte de su proceso. De hecho que eso se hace en varios estudios donde la pandemia ha influido fuertemente. Lo sacan para no perturbar.

Claro. Los modelos aprenden ciclicidad, ¿no? Si hay alguna cadencia anual, la van a aprender. Pero en pandemia se destruye eso.

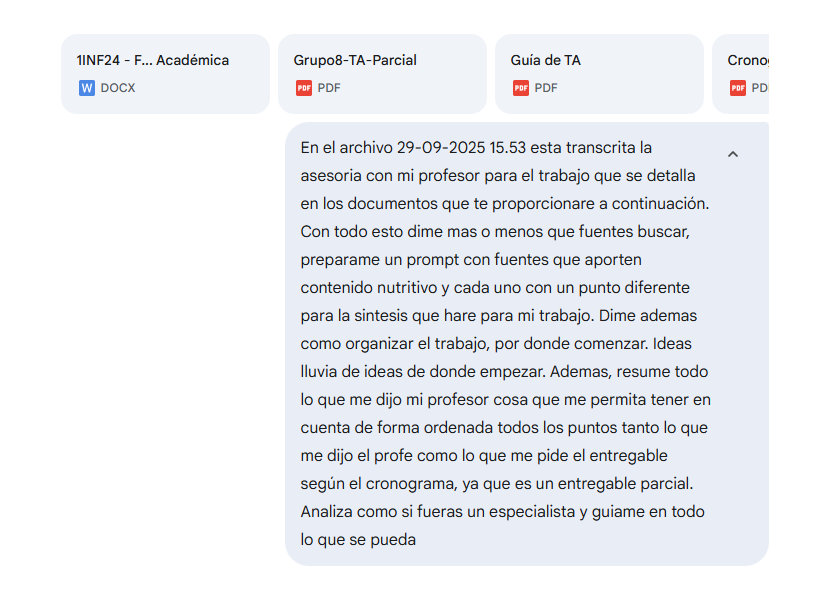
Claro, sí, sí. Es probable que aprendan cosas confusas que ya no se den ahora. Claro.

Parte del data pruning, ¿no? De la preparación de datos. Claro. Muchas gracias.

Muchas gracias por su tiempo. No, de que. Hasta la próxima.

(Transcrito por TurboScribe.ai. Actualizar a Ilimitado para eliminar este mensaje.)

# Prompt



¡Hola! Excelente iniciativa la tuya de buscar una guía detallada para tu Tarea Académica. He analizado a fondo todos los documentos que me proporcionaste: la transcripción de la asesoría con tu profesor, el cronograma del curso, la guía de la tarea y el formato del informe.

Actuaré como tu especialista guía para que puedas abordar este entregable parcial de la mejor manera posible. A continuación, te presento un plan de acción completo.

### **1. Resumen Clave de la Asesoría con tu Profesor**

Primero, sinteticemos los puntos más importantes que tu profesor mencionó. Tener esto claro es fundamental, ya que es la guía directa para tu proyecto.

* **Problema Central:** El objetivo es crear un modelo predictivo (proyectivo) para la tasa de natalidad en Perú. La utilidad principal es ayudar al Estado a prever y prepararse en términos de recursos, especialmente en el sistema de salud.
* **Alcance y Escala:**
  + **Geográfico:** Puedes trabajar con los 1,800 distritos del Perú, o para simplificar, enfocarte en una región representativa (costa, sierra y selva) o un solo departamento.
  + **Temporal:** El profesor desaconseja usar datos anuales por ser muy pocos (tendrías solo 10 puntos de 2015 a 2025). Recomienda una agregación  
     **mensual** o **trimestral** para tener una serie temporal más robusta.
* **Variables y Datos (Features):**
  + **Históricos:** La información principal para el modelo serán los valores previos de la tasa de natalidad (ej. los últimos 12 meses o trimestres).
  + **Locales/Exógenas:** Se recomienda fuertemente enriquecer el modelo con datos locales del distrito/provincia que influyen en la natalidad. Ejemplos: crecimiento económico, índices de pobreza, nivel educativo, índice de desarrollo humano, etc.
* **Variable Objetivo (Target a Predecir):**
  + **Opción Principal:** La tasa de natalidad general (nacimientos per cápita para que sea comparable entre distritos).
  + **Opciones Secundarias (Interesantes):**
    1. Pronosticar la tasa de natalidad específica de  
        **madres adolescentes**. Esto tiene implicaciones de política pública diferentes y muy relevantes.
    2. Pronosticar la tasa de nacimientos sin vida.
  + **Recomendación del Profesor:** Para el entregable parcial, céntrate en **un solo objetivo** (un target) para hacerlo bien. Más adelante, para el final, podrías añadir un segundo objetivo si el tiempo lo permite.
* **Preguntas de Investigación (Hipótesis):** Tu trabajo debe responder preguntas. El profesor sugirió algunas excelentes:
  + ¿Cuánto influyen las variables locales (socioeconómicas) en la precisión del pronóstico de la tasa de natalidad, en comparación con usar solo datos históricos?
  + ¿Qué tipo de modelo de Machine Learning (ej. regresión, redes neuronales) realiza el mejor pronóstico para estos datos?
  + ¿Cómo impactó la pandemia en los datos? ¿Es mejor incluir o excluir ese período (2020-2022) del entrenamiento?
* **Metodología y Experimentación:**
  + Realiza al menos dos experimentos: un modelo que solo use datos históricos y otro que incorpore las variables locales.
  + Usa una estrategia de validación temporal: entrena el modelo con datos hasta cierto año (ej. 2015-2022) y pruébalo en los años finales (ej. 2023-2025).
  + El 80% del éxito del trabajo está en el pre-procesamiento y la preparación de los datos.

### **2. Lluvia de Ideas y Plan de Acción: ¿Por Dónde Empezar?**

Basado en lo anterior y en los requisitos de tu entregable parcial, este es el camino que te recomiendo seguir:

**Paso 1: Definir el Alcance (Decisiones Clave)** Antes de escribir, decide y define:

1. **Variable Objetivo:** Comienza con la **tasa de natalidad general**. Es la idea principal y más sólida. Anota la idea de la natalidad adolescente para la sección de "trabajos futuros".
2. **Agregación Temporal:** Elige **trimestral**. Es un excelente punto medio que te dará suficientes datos (40 puntos por distrito) sin la alta volatilidad que podrían tener los datos mensuales.
3. **Agregación Geográfica:** Propón empezar con un **departamento representativo** que tenga diversidad (ej. Lima, que tiene distritos muy diferentes, o uno que combine costa y sierra). Esto hace el problema manejable.

**Paso 2: Formular la Pregunta de Investigación Central** Esta es la columna vertebral de tu informe. Basado en las sugerencias del profesor, una hipótesis fuerte sería:

*"La inclusión de variables socioeconómicas locales, como el Índice de Desarrollo Humano y la tasa de pobreza, mejora significativamente la precisión del pronóstico de la tasa de natalidad trimestral en los distritos de [Departamento Elegido], en comparación con un modelo que utiliza únicamente la serie temporal histórica de nacimientos."*

**Paso 3: Búsqueda de Fuentes Científicas (¡Clave para tu Nota!)** La rúbrica le da 6 de 20 puntos a esta sección. Necesitas encontrar y sintetizar al menos 3

*papers* relevantes. A continuación, te preparo los prompts.

**Paso 4: Estructurar y Redactar el Informe Parcial** Sigue el "Formato de Informe de Tarea Académica" al pie de la letra. Dedica tiempo a cada sección como te detallo más abajo. Recuerda que tienes un máximo de 4 páginas.

### **3. Prompts para la Búsqueda de Fuentes Científicas**

Usa estas búsquedas en Google Scholar, IEEE Xplore, Scopus o ScienceDirect. Te doy tres enfoques distintos para que tus fuentes sean variadas y nutritivas.

**Prompt 1: Enfoque en Modelos de Series Temporales y Machine Learning**

* **Búsqueda:** "birth rate forecasting" AND ("time series analysis" OR "machine learning" OR "LSTM" OR "SARIMA" OR "gradient boosting")
* **Objetivo:** Encontrar artículos que comparen diferentes algoritmos para predecir tasas de natalidad o datos demográficos similares.
* **¿Qué buscar en estos papers?:** Fíjate en **qué modelos usaron** (SARIMA, Prophet, LSTMs, etc.), **qué métricas de error** aplicaron (RMSE, MAE) y **cómo validaron sus resultados**. Esto te dará la base técnica para tu sección de *Metodología*.

**Prompt 2: Enfoque en los Factores Socioeconómicos (Variables Exógenas)**

* **Búsqueda:** ("socioeconomic determinants" OR "factors influencing") AND ("fertility rate" OR "birth rate") AND "developing countries"
* **Objetivo:** Encontrar estudios que identifiquen qué variables locales (pobreza, educación, salud, etc.) tienen el mayor impacto en las tasas de natalidad.
* **¿Qué buscar en estos papers?:** Identifica las **variables clave** que estos estudios mencionan. Esto justificará por qué eliges incluir el Índice de Desarrollo Humano, la desigualdad, etc., en tu modelo. Te servirá para la  
   *Introducción* (justificación del problema) y la *Metodología* (descripción de datos).

**Prompt 3: Enfoque en la Aplicación y Relevancia (Políticas Públicas)**

* **Búsqueda:** "population forecasting" AND "public health planning" AND "resource allocation" AND "machine learning"
* **Objetivo:** Encontrar artículos que conecten los modelos predictivos con la toma de decisiones gubernamentales.
* **¿Qué buscar en estos papers?:** Usa estos artículos para fortalecer la sección de *Introducción*, específicamente en **relevancia y justificativa**. Te darán argumentos sólidos sobre por qué tu trabajo es importante y cómo el Estado puede beneficiarse de él.

### **4. Guía Detallada para tu Entregable Parcial**

Aquí te explico qué poner en cada sección del informe, conectando todo lo anterior.

* **Título:** Algo claro y conciso. Ej: "Modelo Predictivo de la Tasa de Natalidad en Distritos Peruanos utilizando Machine Learning y Variables Socioeconómicas".
* **Resumen:** Escríbelo al final. Debe ser un párrafo corto que mencione el problema (necesidad de planificar recursos), tu objetivo (crear un modelo predictivo), tu metodología (usar modelos de regresión con datos históricos y locales) y la pregunta que buscas responder (si las variables locales mejoran la predicción).
* **1. Introducción:**
  + **Descripción del problema:** Explica por qué es vital para un gobierno anticipar los nacimientos (planificación de salud, vacunas, educación). Usa las fuentes del  
     **Prompt 3** para darle peso. Menciona la complejidad en un país diverso como Perú.
  + **Hipótesis y/o pregunta a abordar:** Coloca aquí la hipótesis que formulamos en el "Paso 2".
  + **Objetivos:**
    - *General:* Desarrollar y evaluar un modelo de Machine Learning para pronosticar la tasa de natalidad trimestral a nivel distrital en [Departamento Elegido].
    - *Específicos:*
      1. Comparar el desempeño de un modelo de pronóstico basado únicamente en datos históricos contra un modelo enriquecido con variables socioeconómicas.
      2. Identificar las variables locales con mayor impacto en la predicción de la tasa de natalidad.
      3. Evaluar la viabilidad del modelo como herramienta de apoyo a la planificación estatal.
* 2. Trabajos Relacionados:
  + Aquí sintetizas los 3 (o más) artículos que encontraste. Para cada uno, escribe un párrafo corto explicando:
    - El problema que abordaron.
    - La metodología o los modelos que usaron (fuentes del **Prompt 1**).
    - Las variables más relevantes que encontraron (fuentes del **Prompt 2**).
    - Su principal conclusión.
* 3. Metodología:
  + **Datos:** Describe tu fuente principal de datos de nacimientos (2015-2025). Menciona las fuentes de donde  
     *planeas* sacar los datos socioeconómicos (ej. INEI, PNUD para el IDH, etc.).
  + **Pre-procesamiento:** Explica que agregarás los datos por distrito y por trimestre. Menciona que normalizarás los nacimientos para obtener una tasa per cápita (por cada 1,000 o 10,000 habitantes) para hacerlos comparables. También, menciona que analizarás el efecto de la pandemia y considerarás excluir esos años como parte de un experimento.
  + **Enfoque Propuesto:** Describe los dos enfoques que vas a comparar, como sugirió el profesor:
    - **Modelo Base (Histórico):** Un modelo que usará como *features* los valores de la tasa de natalidad de trimestres anteriores (ej. las últimas 4 u 8 observaciones).
    - **Modelo Enriquecido (Histórico + Local):** El mismo modelo, pero añadiendo como *features* las variables socioeconómicas (IDH, pobreza, etc.).
  + Menciona los algoritmos que planeas probar. Dado el cronograma, para la semana 8 ya habrán visto algoritmos de regresión. Puedes proponer usar  
     **Regresión Lineal** como línea base y un modelo más avanzado como **Gradient Boosting (XGBoost)** o un **Random Forest**.
* **4. Experimentación y Resultados (¡Solo la Planificación!):**
  + En esta entrega no necesitas los resultados finales, sino  
     **describir cómo los obtendrás**.
  + **Setup experimental:** Explica tu estrategia de validación. "Se entrenarán los modelos con datos de 2015-T4 2022 y se evaluará su capacidad de pronóstico en los trimestres de 2023-T1 a 2025-T1".
  + **Métricas de evaluación:** Indica que usarás métricas estándar para regresión como el Error Cuadrático Medio (RMSE) y el Error Absoluto Medio (MAE) para comparar los modelos.
  + **Experimentos a realizar:** Lista claramente los experimentos que planeas hacer para responder tu hipótesis. Ej:
    - "Experimento 1: Evaluar el rendimiento del Modelo Base."
    - "Experimento 2: Evaluar el rendimiento del Modelo Enriquecido."
    - "Experimento 3: Comparar los resultados de ambos experimentos para determinar si la inclusión de variables locales aporta una mejora estadísticamente significativa."

Con esta estructura, no solo cumplirás con todos los requisitos del entregable parcial, sino que también tendrás una base extremadamente sólida para tu trabajo final. ¡Mucho éxito!

Prompts

¡Excelente! Armemos juntos esa sección de "Enfoque(s) propuesto" de manera sólida y formal, tal como lo pide la guía.

Aquí te presento una propuesta detallada que puedes adaptar directamente a tu informe. Incluye la notación formal, la descripción de entrada/salida y los detalles de la aplicación de los algoritmos, junto con las figuras para ilustrarlo.

### **3. Metodología**

#### **Enfoque(s) propuesto**

Para abordar el problema del pronóstico de la tasa de natalidad, se propone un enfoque comparativo basado en aprendizaje supervisado, específicamente en tareas de regresión. Se evaluarán dos modelos distintos para determinar el impacto de las variables socioeconómicas locales en la precisión del pronóstico.

**Descripción formal del problema**

El problema se define como la predicción de la tasa de natalidad en un área geográfica y un instante de tiempo futuros, basándose en información histórica y contextual.

Sea **D** el conjunto de todos los distritos a analizar (ej. D = {d₁, d₂, ..., dₙ}) y **T** la secuencia de períodos de tiempo trimestrales (T = {t₁, t₂, ..., tₘ}).

Nuestro objetivo es predecir la **tasa de natalidad** y para un distrito d ∈ D en el siguiente trimestre t+1. Definimos esta variable objetivo como:

* ŷ\_{d, t+1}: La tasa de natalidad predicha para el distrito d en el trimestre t+1.

Esta predicción se basará en un **vector de características (features)** X\_{d,t} que representa la información disponible para el distrito d hasta el trimestre t.

El problema se reduce a encontrar una función de mapeo ƒ tal que:

ŷ\_{d, t+1} = ƒ(X\_{d,t})

Donde ƒ es el modelo de Machine Learning entrenado. La función ƒ buscará minimizar una métrica de error, como el Error Cuadrático Medio (RMSE), entre los valores predichos (ŷ) y los valores reales (y).

Se proponen dos enfoques que difieren en la composición del vector de características X\_{d,t}.

**Enfoque 1: Modelo Base (Histórico)**

Este primer enfoque sirve como nuestra línea base (baseline) y asume que el comportamiento futuro de la tasa de natalidad puede ser explicado principalmente por sus valores pasados.

* **Descripción Entrada/Salida:**
  + **Entrada:** Un vector de características X\_{d,t} que consiste en una ventana de las últimas k tasas de natalidad observadas para el distrito d. X\_{d,t} = [y\_{d,t}, y\_{d,t-1}, ..., y\_{d,t-k+1}]
  + **Salida:** La tasa de natalidad predicha para el siguiente trimestre, ŷ\_{d, t+1}.

La Figura 1 ilustra el comportamiento de este modelo.

+------------------------------------------------------+

| |

| Entrada: |

| Serie Temporal Histórica (últimos k trimestres) |

| - Tasa de natalidad en t |

| - Tasa de natalidad en t-1 |

| - ... |

| - Tasa de natalidad en t-k+1 |

| (Para un distrito 'd') |

| |

+------------------------+-----------------------------+

|

▼

+------------------------------------------------------+

| |

| Modelo 1: ƒ\_base(X) |

| (Ej. Regresión Lineal, ARIMA) |

| Aprende patrones y tendencias de la serie temporal. |

| |

+------------------------+-----------------------------+

|

▼

+------------------------------------------------------+

| |

| Salida: |

| Predicción de la Tasa de Natalidad |

| para el trimestre t+1 |

| |

+------------------------------------------------------+

**Figura 1. Esquema del Modelo Base (Histórico).**

**Enfoque 2: Modelo Enriquecido (Histórico + Socioeconómico)**

Este segundo enfoque parte de la hipótesis de que, además de la tendencia histórica, existen factores externos locales que influyen directamente en la tasa de natalidad.

* **Descripción Entrada/Salida:**
  + **Entrada:** Un vector de características aumentado X\_{d,t} que contiene no solo las últimas k tasas de natalidad, sino también un conjunto de p variables socioeconómicas S para el distrito d en el trimestre t. S\_{d,t} = [s\_1, s\_2, ..., s\_p]\_{d,t} (Ej: Índice de Desarrollo Humano, tasa de pobreza, nivel educativo). X\_{d,t} = [y\_{d,t}, y\_{d,t-1}, ..., y\_{d,t-k+1}, s\_{1,d,t}, s\_{2,d,t}, ..., s\_{p,d,t}]
  + **Salida:** La tasa de natalidad predicha para el siguiente trimestre, ŷ\_{d, t+1}.

La Figura 2 ilustra el comportamiento de este modelo.

+------------------------------------------------------+

| |

| Entrada: |

| Datos Históricos + Datos Socioeconómicos |

| - Tasa de natalidad en t, t-1,... |

| - Índice de Desarrollo Humano en t |

| - Tasa de Pobreza en t |

| - ... (otras 'p' variables) |

| (Para un distrito 'd') |

| |

+------------------------+-----------------------------+

|

▼

+------------------------------------------------------+

| |

| Modelo 2: ƒ\_enriquecido(X) |

| (Ej. Gradient Boosting - XGBoost) |

| Aprende relaciones no lineales complejas entre |

| todas las variables de entrada. |

| |

+------------------------+-----------------------------+

|

▼

+------------------------------------------------------+

| |

| Salida: |

| Predicción de la Tasa de Natalidad |

| para el trimestre t+1 |

| |

+------------------------------------------------------+

**Figura 2. Esquema del Modelo Enriquecido (Histórico + Socioeconómico).**

**Algoritmos Aplicados al Problema**

Para implementar los enfoques descritos, se seleccionaron los siguientes algoritmos, adaptados a las particularidades de nuestro problema:

1. **Regresión Lineal (Para el Modelo Base):** Se utilizará como un benchmark simple y robusto. El modelo aprenderá un coeficiente (w) para cada una de las k observaciones históricas de la tasa de natalidad. La predicción se calculará como una suma ponderada de estas entradas, adaptándose específicamente para capturar la tendencia lineal de la serie temporal. La ecuación del modelo adaptado es: ŷ\_{d, t+1} = w₀ + w₁\*y\_{d,t} + w₂\*y\_{d,t-1} + ... + wₖ\*y\_{d,t-k+1}
2. **XGBoost (eXtreme Gradient Boosting) (Para el Modelo Enriquecido):** Se eligió este algoritmo por su alta capacidad para capturar relaciones complejas y no lineales entre las variables, lo cual es fundamental para nuestra hipótesis.
   * **Adaptación al Problema:** XGBoost se entrenará con el vector de características combinado. En lugar de aprender un solo modelo, XGBoost construirá un ensamble de árboles de decisión de forma secuencial. Cada nuevo árbol se especializará en corregir los errores de predicción del ensamble anterior. Esta capacidad es crucial para modelar cómo una combinación de factores (ej. una ligera baja en la tasa de natalidad histórica junto con un aumento en el nivel educativo local) puede impactar la tasa de natalidad futura.
   * **Análisis de Importancia de Características:** Una aplicación clave de XGBoost en este trabajo será utilizar su capacidad para calcular la "importancia de características" (feature importance). Tras entrenar el modelo, se analizará esta métrica para cuantificar qué variables (las históricas o las socioeconómicas) tienen mayor poder predictivo. Esto nos permitirá validar o refutar directamente nuestra hipótesis de investigación.

Esta fuente es **excelente** y **muy relevante** para tu trabajo. Es un *paper* científico que aborda un problema casi idéntico al tuyo (predicción de tasa de natalidad usando Machine Learning), lo que lo convierte en un pilar para tu sección de "Trabajos Relacionados" y una guía para tu metodología.

FUENTE 1

Aquí te detallo lo más útil que puedes extraer de este documento para tu objetivo de **predecir la tasa de natalidad por departamento**:

* **1. Justificación Sólida para tu Introducción:** La sección de "Introduction" (página 2) te da argumentos claros sobre **por qué es importante predecir las tasas de natalidad**. Habla de su impacto en la formulación de políticas, la planificación a largo plazo, la estabilidad fiscal y la anticipación de la demanda en varios sectores. Puedes citar y adaptar estas ideas para fortalecer la relevancia de tu propio proyecto.
* **2. Validación de tu Metodología:** El estudio compara **modelos de Machine Learning no lineales** (Random Forest, XGBoost) con **enfoques estadísticos tradicionales** (ARIMA, Regresión Lineal). Esto valida perfectamente el plan que te sugirió tu profesor: tener un modelo base y compararlo con uno más avanzado. Demuestra que tu enfoque es científicamente sólido.
* **3. Descripción Formal de Algoritmos:** ¡Esto es oro puro para tu sección de "Metodología"!
  + En la sección 2.4 (página 4), te da una descripción formal y hasta la **notación matemática de XGBoost** (ecuaciones 1 y 2). Puedes usar esto como base para describir el algoritmo que aplicarás a tus datos, adaptándolo a tu contexto.
  + También describe **Random Forest**, que es otro de los modelos que podrías comparar.
* **4. Métricas de Evaluación Claras:** La sección 2.5 (página 5) define las métricas que usan para medir el error: **RMSE** (Root Mean Square Error), **MAE** (Mean Absolute Error) y **SMAPE**. Estas son exactamente las métricas que debes usar en tu sección de "Experimentación" para comparar qué tan buenos son tus modelos.
* **5. Un Ejemplo de Diseño Experimental:** En la sección 4 ("Results", página 7), describen cómo hicieron su experimento:
  + Dividieron los datos en **entrenamiento y prueba (80-20%)**.
  + Usaron un enfoque de **ventana deslizante** (*rolling estimation window*).
  + Ajustaron los **hiperparámetros** de los modelos.
  + Esto te da un modelo a seguir para planificar y describir tus propios experimentos.
* **6. Ideas para el Pre-procesamiento de Datos:** En la sección 3 ("Data", página 6), mencionan que aplicaron una **transformación logarítmica** a los datos y realizaron pruebas de estacionariedad. Estas son técnicas de pre-procesamiento importantes que deberías considerar para tus datos departamentales de Perú.

En resumen, esta fuente no es solo "una referencia más". **Es un plano 🏛️.** Úsala como guía para estructurar tu informe, justificar tu problema, describir formalmente tu metodología y planificar tu experimentación. Citar este trabajo le dará mucha credibilidad y rigor científico a tu entregable.

# VALOR:

Analizando el cronograma y el sílabo que me enviaste, el algoritmo que más te conviene de la fuente es, sin duda, **XGBoost (Extreme Gradient Boosting)**.

Aquí te explico el porqué de manera detallada:

### **¿Por qué XGBoost es la mejor opción?**

1. **Alineación con el Cronograma:** Tu entregable parcial es en la **semana 8**.
   * En la **semana 7**, el curso cubre "**Algoritmos básicos para regresión**". El sílabo especifica "regresión lineal, k-NN".
   * En la **semana 8**, ven "**Algoritmos básicos para clasificación**", y el sílabo menciona explícitamente "**árbol de decisión**".
2. XGBoost es un algoritmo de *ensamble* muy potente que **se basa en árboles de decisión**. Por lo tanto, cuando presentes tu trabajo, podrás justificar tu elección diciendo que estás aplicando una versión avanzada y de alto rendimiento de un concepto fundamental visto en clase justo antes de la entrega.
3. **Rendimiento Superior:** La propia fuente que analizamos (fuente1.pdf) concluye que **XGBoost fue el modelo con mejor desempeño**, superando a la Regresión Lineal y a Random Forest. Usar el mejor algoritmo de tu paper de referencia como propuesta principal le da muchísima solidez a tu trabajo.
4. **Cumple con la Guía de la Tarea:** El objetivo de la Tarea Académica es "desarrollar y/o adaptar algoritmos de IA para resolver un problema". Proponer XGBoost no es solo "usar un algoritmo", es demostrar que puedes:
   * Identificar un método de vanguardia en la literatura científica.
   * Entender su fundamento (los árboles de decisión del curso).
   * Adaptarlo a tu problema específico (predicción de natalidad en departamentos de Perú).

### **Estrategia Recomendada para tu Informe 📝**

Para tu sección de "Metodología", te recomiendo proponer el siguiente plan, que está perfectamente alineado con el curso, tu profesor y la fuente:

* **Modelo Base (Línea Base):** Utiliza la **Regresión Lineal**. Es el algoritmo de regresión más claro que verán en la semana 7. Te servirá como un punto de comparación simple y robusto para demostrar la superioridad de tu modelo principal.
* **Modelo Propuesto (Avanzado):** Utiliza **XGBoost**. Descríbelo como un método de ensamble basado en árboles de decisión potenciados por gradiente (*gradient boosting*). Explica que lo elegiste por su excelente rendimiento en problemas similares, como lo demuestra tu fuente de referencia, y porque se fundamenta en conceptos vistos en el curso.

Esta estrategia es ganadora porque demuestra que dominas los conceptos básicos (Regresión Lineal) y que además tienes la iniciativa de investigar y aplicar técnicas más avanzadas y efectivas (XGBoost).

Old

Datasets

1. Matriculación y Trayectoria Estudiantil 2021-2024

<https://datosabiertos.gob.pe/dataset/matriculaci%C3%B3n-y-trayectoria-estudiantil-2021-2024>

1. Padrones de instituciones educativas (Reforma Magisterial)

<https://www.gob.pe/institucion/minedu/informes-publicaciones/6478096-padrones-de-instituciones-educativas-reforma-magisterial>

1. Servicios Educativos - ESCALE

<https://escale.minedu.gob.pe/>

<https://www.gob.pe/institucion/minedu/informes-publicaciones/6478096-padrones-de-instituciones-educativas-reforma-magisterial>

Publicaciones Científicas

[Student Dropout Prediction for University with High Precision and Recall](https://www.mdpi.com/2076-3417/13/10/6275) no

[(PDF) Predicting Student Dropout in Higher Education Using Machine Learning Techniques : A Predictive Model Using XGBoost Algorithm](https://www.researchgate.net/publication/386564297_Predicting_Student_Dropout_in_Higher_Education_Using_Machine_Learning_Techniques_A_Predictive_Model_Using_XGBoost_Algorithm) no