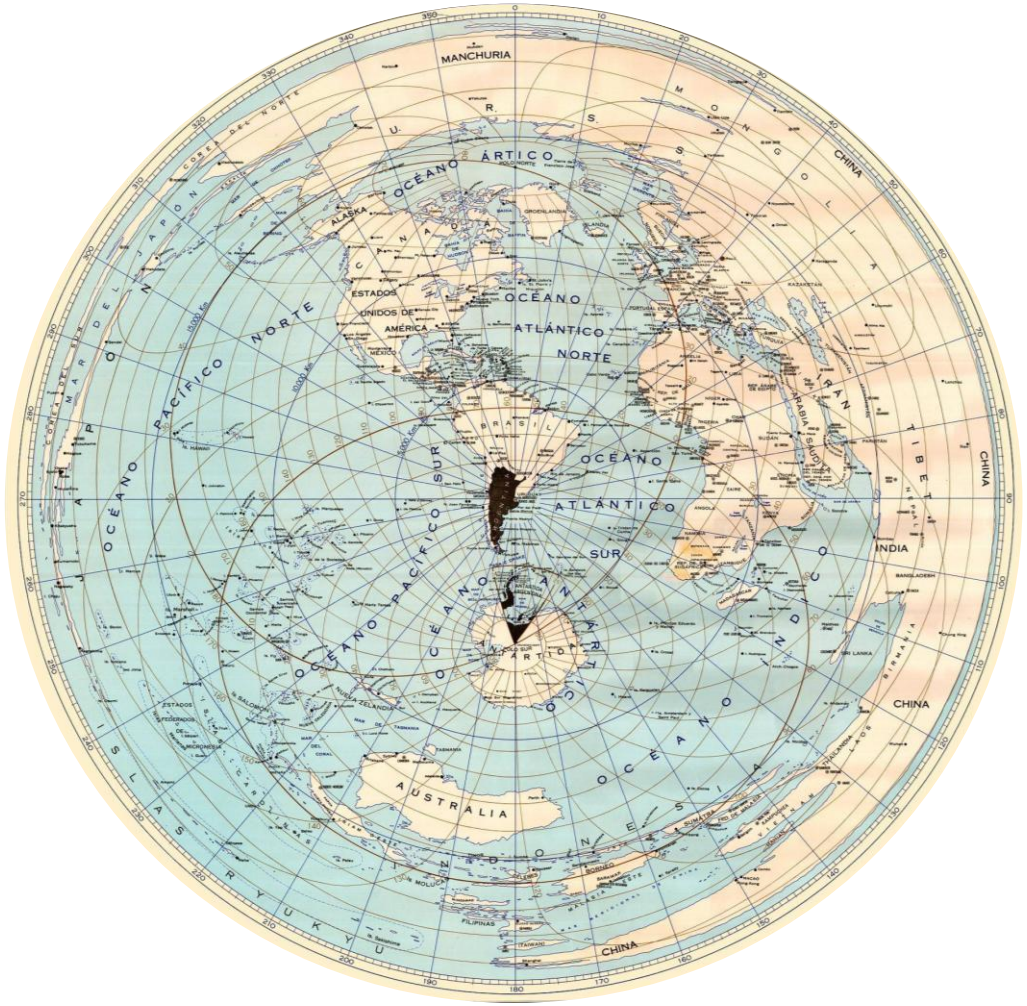
Análisis de Establecimientos Productivos Laborales de Argentina

Trabajo Final de Ciencia de Datos para Economía y Negocios

- Estudiante: Valentina Gutiérrez Curátola
- Registro: 906.474
- Docente: Nicolás Sidicaro

Contenido

- ✓ Introducción y Objetivos.
- ✓ Descripción de los datos utilizados.
- ✓ Metodología Aplicada.
- ✓ Análisis Exploratorio de Datos (“EDA”).
- ✓ Resultados Principales
- ✓ Modelos de Machine Learning.
- ✓ Conclusiones.
- ✓ Limitaciones y Trabajo Futuro.



Introducción y Objetivos

¿Existen factores que influyen en la participación femenina laboral?

- ✓ Este proyecto analiza 1.37 millones de establecimientos productivos argentinos.
- ✓ Se aplican técnicas de ciencia de datos para identificar patrones de género, geografía y actividad.
- ✓ Los hallazgos aportan una mirada crítica sobre la estructura productiva y los determinantes de la participación femenina en el mercado laboral.

Introducción y Objetivos

Objetivo del proyecto:

- ✓ Explorar patrones de participación femenina por sector y región.
- ✓ Desarrollar modelos de clasificación y predicción.
- ✓ Generar recomendaciones para política pública y negocios.

Introducción y Objetivos

Hallazgos Principales

- ✓ **Determinismo sectorial:** El tipo de industria predice la composición de género mejor que la ubicación geográfica.
- ✓ **Paradoja exportadora:** Los sectores con mayor capacidad exportadora tienen menor participación femenina.
- ✓ **Liderazgo patagónico:** La región sur muestra mayor inclusión de género que el centro tradicional.
- ✓ **Concentración de servicios:** El 82% de establecimientos se concentra en 3 sectores principales:

    1. Sectores Tradicionales (construcción, agricultura, transporte)
    2. Servicios Especializados (salud, educación, servicios profesionales).
    3. Sectores Productivos (industria y tecnología).

Descripción de los datos utilizados

Datos Originales

- ✓ Los datos fueron obtenidos del Centro de Estudios para la Producción (CEP XXI).
- ✓ Se obtuvieron 1.408.470 registros correspondientes a establecimientos productivos relevados durante los años 2021 y 2022.
- ✓ Para el análisis se consideraron 24 variables que incluyen información empresarial, geográfica y sectorial.

Descripción de los datos utilizados

Proceso de Limpieza

- ✓ Se conservaron el 100% de los registros vinculados a actividades, departamentos y combinaciones departamento-actividad.

- ✓ En cuanto a los establecimientos, se logró conservar el 97% de los registros originales, alcanzando un total de 1.366.827 casos finales.

Etapa	Registros Iniciales	Registros Finales	Retención
Actividades	951	951	100%
Departamentos	527	527	100%
Departamento-Actividad	316.697	316.697	100%
Establecimientos	1.408.470	1.366.827	97,04%

- ✓ Este proceso aseguró la calidad y representatividad de la muestra sin sacrificar volumen de datos relevante.

Descripción de los datos utilizados

Proceso de Limpieza

- ✓ Los datos provienen de un portal oficial, lo que garantiza su validez y estructura estandarizada.
- ✓ Al ser generados por organismos públicos con metodologías consistentes, llegaron en buen estado de limpieza, con identificadores claros y sin registros relevantes faltantes.
- ✓ Esto se refleja en las altas tasas de retención con 100% en actividades, departamentos y combinaciones, y 97% en establecimientos.

Descripción de los datos utilizados

Transformación de Datos

1. Se ajustó la validación del código de actividad económica (CLAE6) para aceptar tanto códigos de cinco como de seis dígitos, garantizando así una mayor cobertura de actividades.

2. Se eliminaron filtros geográficos que resultaban demasiado estrictos, lo que permitió incluir más regiones en el análisis.

3. Se priorizó la conservación de datos de alta calidad, siendo estos los más completos y consistentes, logrando una retención del 97% de los registros tras la depuración.

Metodología Aplicada

¿Cómo analizamos los datos?

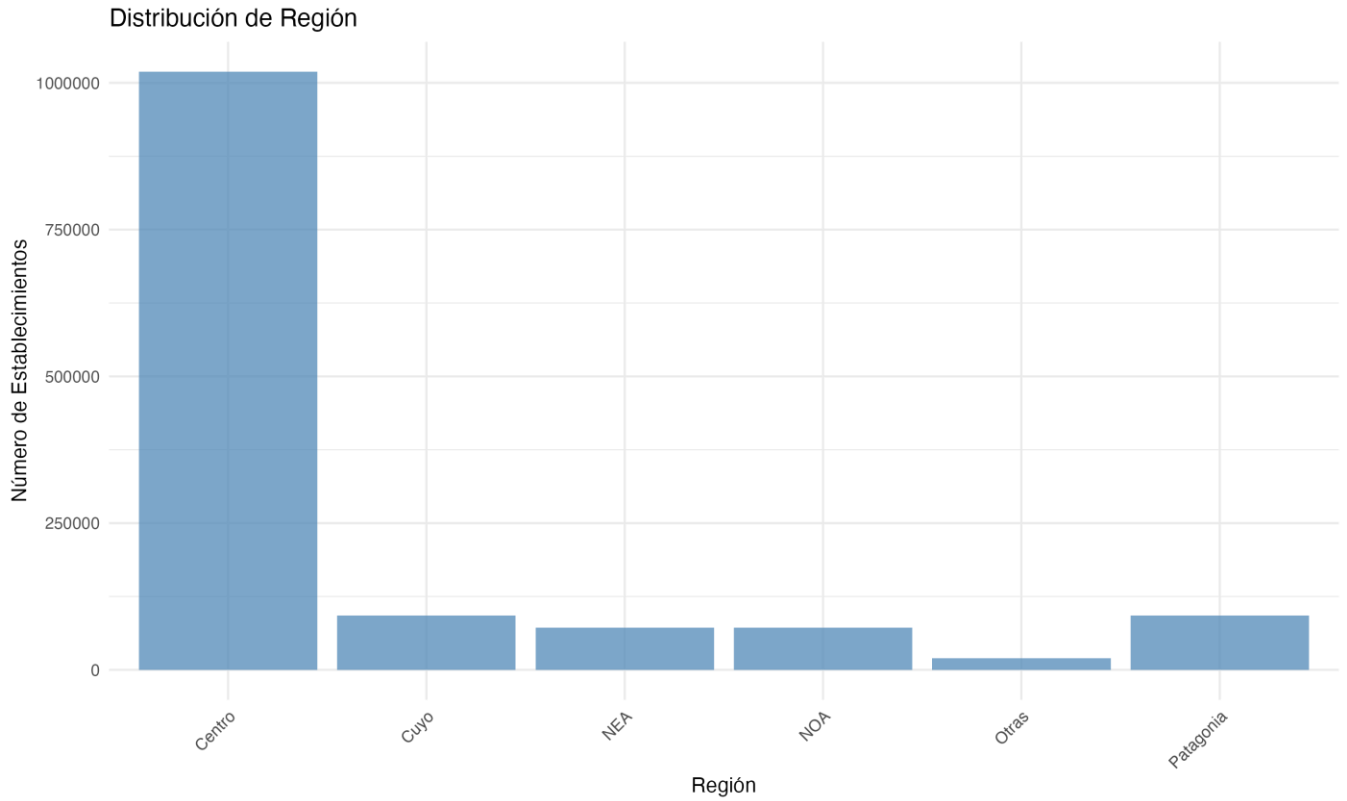
1. Exploración
   - ¿Dónde están las empresas?
   - ¿Qué sectores son más grandes?
   - ¿Cuáles tienen más mujeres?
2. Identificación
   - ¿Qué empresas exportan?
   - ¿Qué sectores son más grandes?
   - ¿Cuáles tienen más mujeres?
3. Agrupación
   - ¿Se pueden agrupar sectores similares?
   - ¿Qué tienen en común?
4. Validación
   - ¿Los patrones son confiables?
   - ¿Se pueden usar para predecir?

¿Por qué este enfoque?

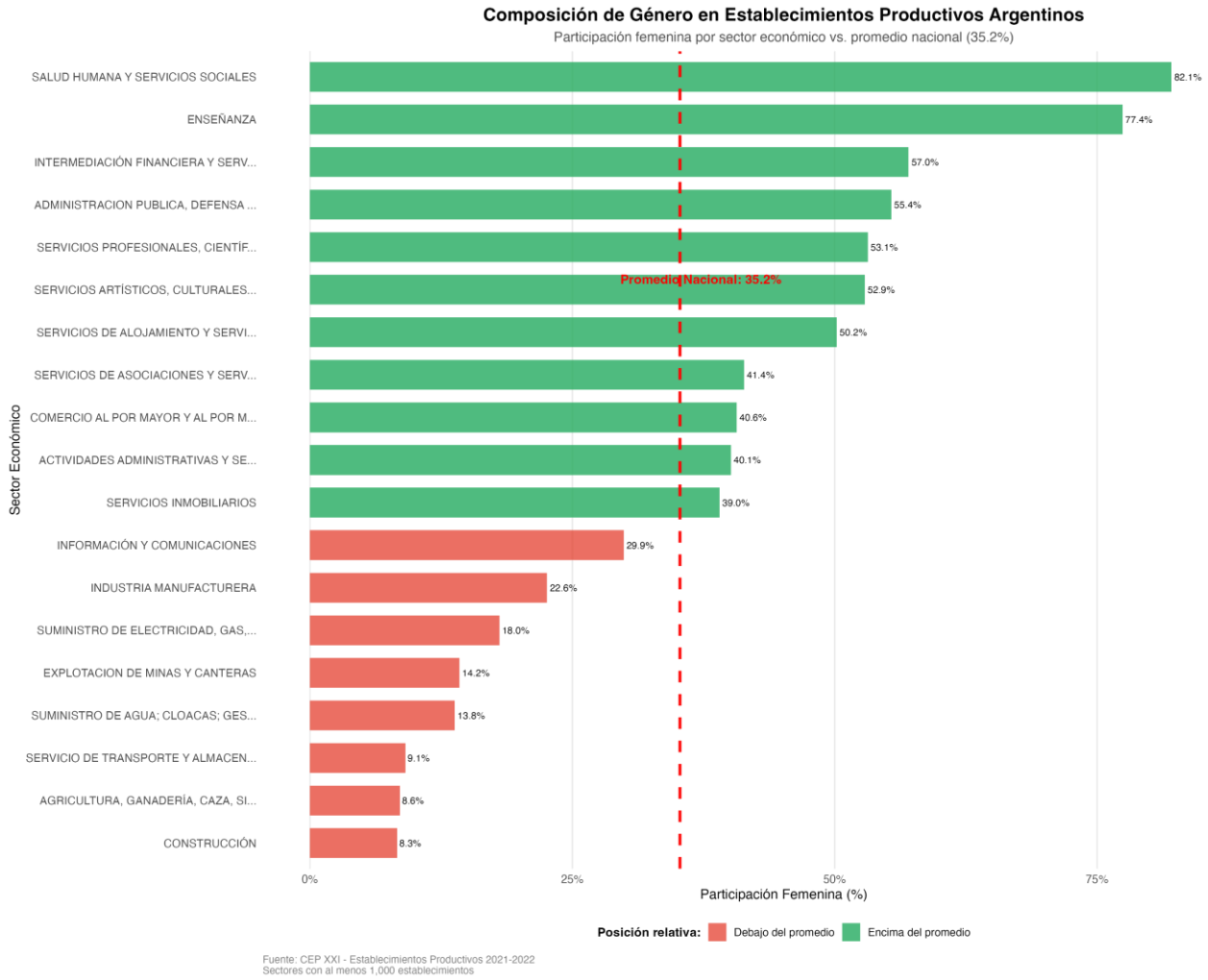
✓ Dado el volumen y complejidad de los datos (1.37 millones de empresas), fue necesario aplicar métodos sistemáticos para identificar patrones y evitar conclusiones erróneas.

EDA: Concentración Geográfica

- ✓ Centro: 74,5% de establecimientos (Buenos Aires, Córdoba, Santa Fe).
- ✓ Distribución equilibrada: Resto del país 25,5%.

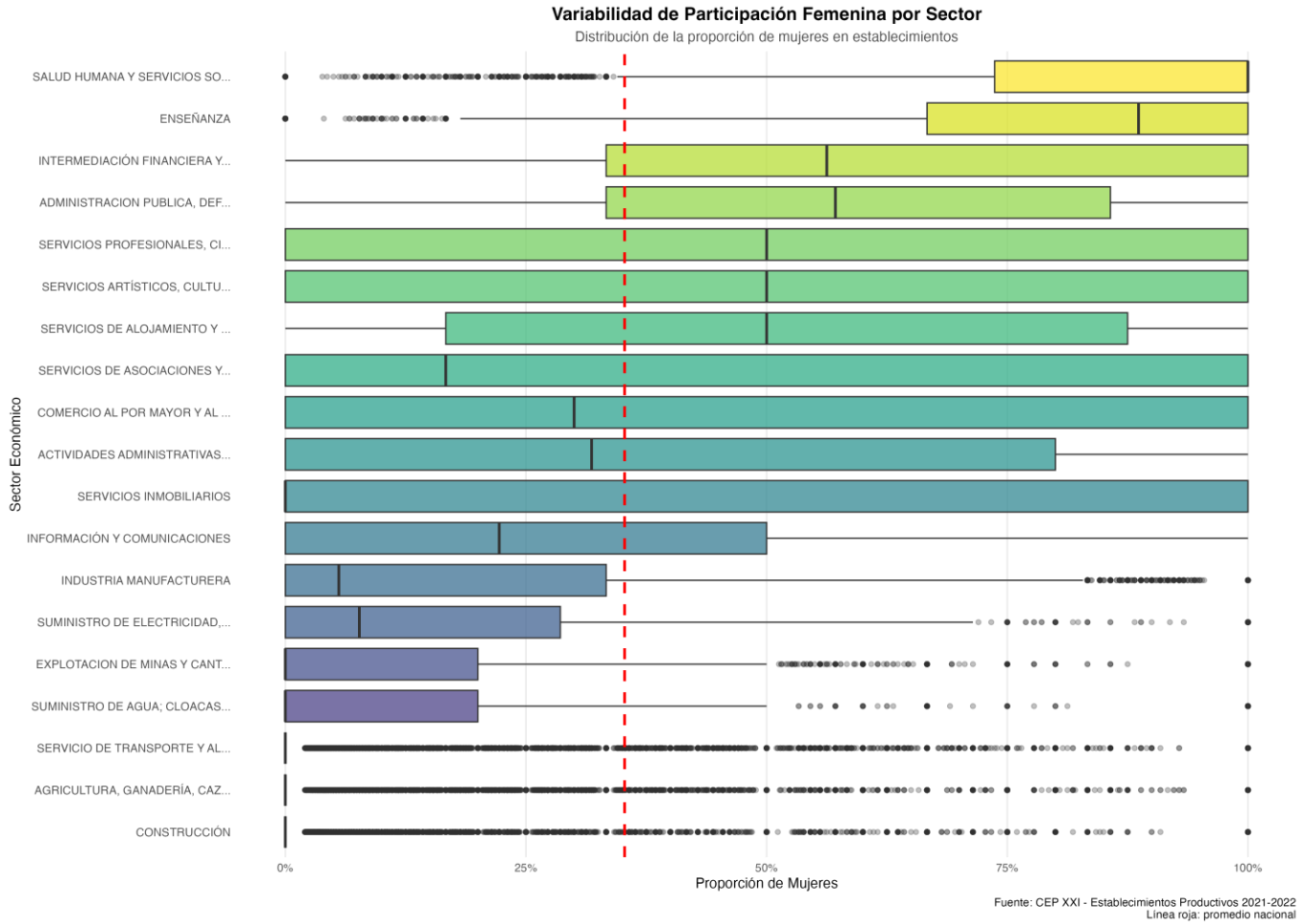


EDA: Composición de Género



Promedio nacional: 35,2% de participación femenina.

EDA: Composición de Género



Variabilidad alta: Desde 8% (construcción) hasta 82% (salud).

EDA: Patrones Sectoriales - Proceso de Clusterización

Sectores "Feminizados" (>50% mujeres)

- ✓ Salud: 82% mujeres, 0,5% exportadores.
- ✓ Educación: 77% mujeres, mínima exportación.
- ✓ Patrón identificado (cluster): sectores orientados al mercado doméstico y con alta calificación requerida.

EDA: Patrones Sectoriales - Proceso de Clusterización

Sectores "Masculinizados" (<20% mujeres)

- ✓ Construcción: 8% mujeres, baja exportación.
- ✓ Agricultura: 9% mujeres, exportación moderada.
- ✓ Patrón identificado (cluster): actividades intensivas en trabajo físico, asociadas a sectores tradicionales.

EDA: Patrones Sectoriales - Proceso de Clusterización

Sectores Exportadores (>5% exportan)

- ✓ Industria manufacturera: 23% mujeres, 12% exportadores.

Resultados Principales: Análisis Sectorial

Principales Sectores por Volumen

- ✓ El sector con mayor cantidad de establecimientos es el comercio, que representa el 28,4% del total (387.785 casos).
- ✓ Le siguen los servicios a asociaciones, con un 10,1%, y la agricultura, con un 10,0%.
- ✓ La industria manufacturera también tiene un peso significativo, con un 9,8% de los establecimientos.
- ✓ El transporte completa el top cinco, con un 6,8%.

EDA: Sorpresas Geográficas

Patagonia Líder en Inclusión

- ✓ Santa Cruz: 41,4% participación femenina
- ✓ Neuquén, Chubut: >40% participación
- ✓ Contradicción: Región menos poblada, mayor inclusión

Centro Tradicional

- ✓ Buenos Aires: 34,9% (bajo el promedio nacional)
- ✓ Reflexión: ¿Concentración no implica inclusión?

Resultados Principales: Análisis Sectorial

Participación Femenina por Sector

- ✓ En los sectores de salud y enseñanza, la participación femenina supera el 75%, siendo considerados espacios altamente feminizados.
- ✓ En los servicios financieros, el porcentaje alcanza el 57%, mostrando una presencia importante de mujeres en servicios especializados.
- ✓ En contraste, en sectores tradicionalmente masculinos como industria, construcción y agricultura, la participación femenina es mucho menor (22,6%, 8,3% y 8,6%, respectivamente).
- ✓ Esta diferencia evidencia una fuerte segmentación de género en el mercado laboral por sector.

Resultados Principales: Análisis Geográfico

Participación Femenina por Provincia

- ✓ Las provincias de la Patagonia lideran en inclusión de género, con Santa Cruz (41,4%), Chubut (40,4%) y Neuquén (40,0%) encabezando el ranking nacional.
- ✓ Tierra del Fuego (39,3%) y Río Negro (38,9%) completan el top 5, consolidando a la región patagónica como referente en participación femenina (promedio regional: 39,0%).
- ✓ En contraste, el NEA presenta los menores niveles de inclusión (27,8% promedio), mientras que la región Centro se ubica en un punto intermedio con un 34,9%.
- ✓ Estos patrones regionales reflejan desigualdades geográficas en la inclusión de género dentro del mercado laboral.

Modelos de Machine Learning ("ML")

Dos líneas de análisis con algoritmos de clasificación y segmentación

1. **Hipótesis de negocio**
   - ✓ Predicción de capacidad exportadora.
   - ✓ Influencia geográfica sobre la participación femenina.
   - ✓ Identificación de patrones distintivos en el sector salud.

2. **Comparación de modelos**
   - ✓ Predicción de alta participación femenina (≥50%).
   - ✓ Evaluación de regresión logística vs. random forest.
   - ✓ Métricas consideradas: Accuracy, AUC, sensibilidad, F1-score.

ML - Hipótesis 1: ¿Podemos predecir qué establecimientos serán exportadores?

✓ Objetivo: Predecir exportación a partir de características sectoriales, geográficas y de género.

✓ Modelo: Random Forest (300 árboles, 5-fold CV)

✓ Resultados:
- Accuracy: 78.8%
- Sensitivity: 69.6%
- AUC: 0.785
- Balanced Accuracy: 74.4%

✓ Variables predictivas clave:
- Sector industria (más exportador)
- Tamaño de empresa
- Cantidad de empleo
- Proporción de mujeres

✓ Interpretación:
- Tamaño e industria son predictores clave de capacidad exportadora.
- Los modelos predicen con buena precisión a pesar del fuerte desbalance (3.2% exportadores).

ML - Hipótesis 2: ¿La participación femenina varía geográficamente de forma predecible?

✓ Objetivo: Determinar si la ubicación predice niveles de participación femenina altos.

✓ Modelo: Regresión Logística (5-fold CV)

✓ Resultados:

- Accuracy: 55.2%
- AUC: 0.568

✓ Variables significativas:

- Longitud: Más al Este ligeramente mayor participación femenina
- Diversidad sectorial: Mayor diversidad en el departamento contribuye a mayor participación femenina
- Región centro: se asocia a menor participación femenina, en promedio

✓ Interpretación:

- Influencia geográfica limitada.
- Efectos significativos pero marginales → el sector sigue siendo el factor dominante.

ML - Hipótesis 3: ¿El sector salud presenta patrones únicos identificables?

✓ Objetivo: Detectar patrones distintivos del sector salud.

✓ Modelo: Random Forest (200 árboles)

✓ Resultados:
- Accuracy: 74.0%
- Sensitivity: 85.6%
- AUC: 0.869

✓ Variables predictivas clave:
- Proporción de mujeres
- Quintil exportador (negativo)
- Empleo

✓ Interpretación:
- El sector salud es altamente identificable por su fuerte perfil de género (82% mujeres), baja exportación y perfil geográfico estable.

ML - Conclusiones del modelado por hipótesis

Conclusiones generales:

- ✓ Sector económico es más determinante que la geografía.
- ✓ ML permite validar hipótesis de política pública y estrategia empresarial.
- ✓ Sectores tradicionalmente masculinos son más exportadores, pero menos inclusivos.

Hipótesis	Modelo	Accuracy	AUC	Conclusión
H1	Random Forest	78.8%	0.785	Predictivo
H2	Regresión Logística	55.2%	0.568	Limitado
H3	Random Forest	74.0%	0.869	Distintivo

ML - Comparación de Modelos: Predicción de Alta Participación Femenina

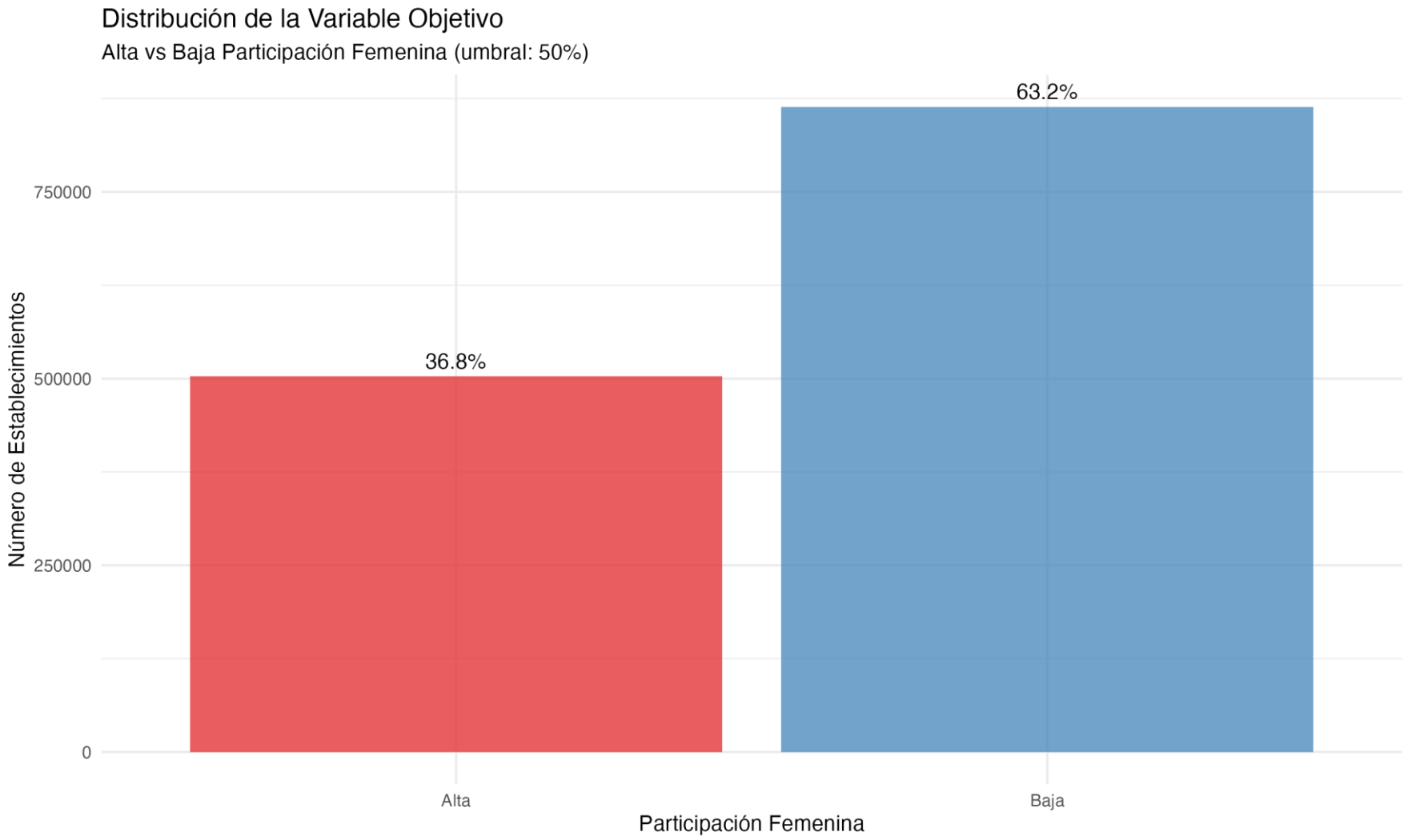
- ✓ Objetivo: Predecir si un establecimiento tiene alta participación femenina (≥50%) en base a sus características (sector, región, tamaño, etc.)

- ✓ Modelos Comparados:
  - Regresión Logística – Modelo interpretable, rápido
  - Random Forest – Modelo no lineal, captura interacciones

- ✓ Métrica de referencia (baseline):
  - Predecir siempre la clase mayoritaria
  - Accuracy = 63.2%

- ✓ Decisión de Corte
  - Alta participación = proporción de mujeres ≥ 50%

ML - Comparación de Modelos: Predicción de Alta Participación Femenina



ML - Resultados y Comparación de Modelos

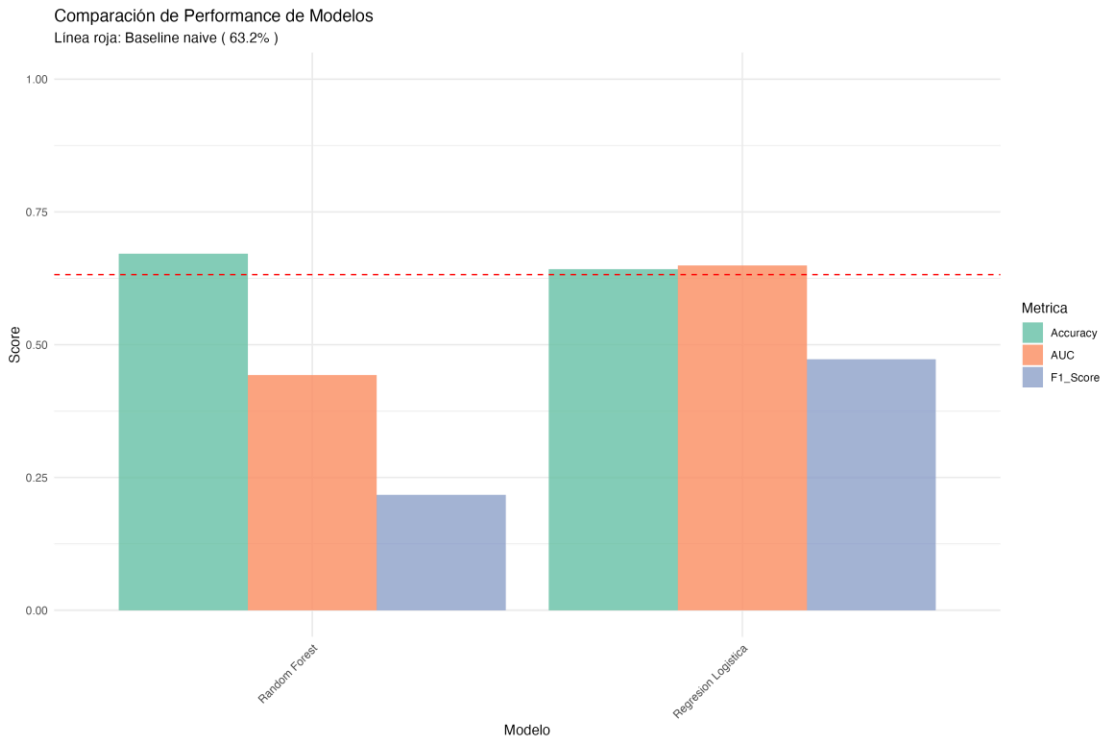
✓ Desempeño en conjunto de prueba.

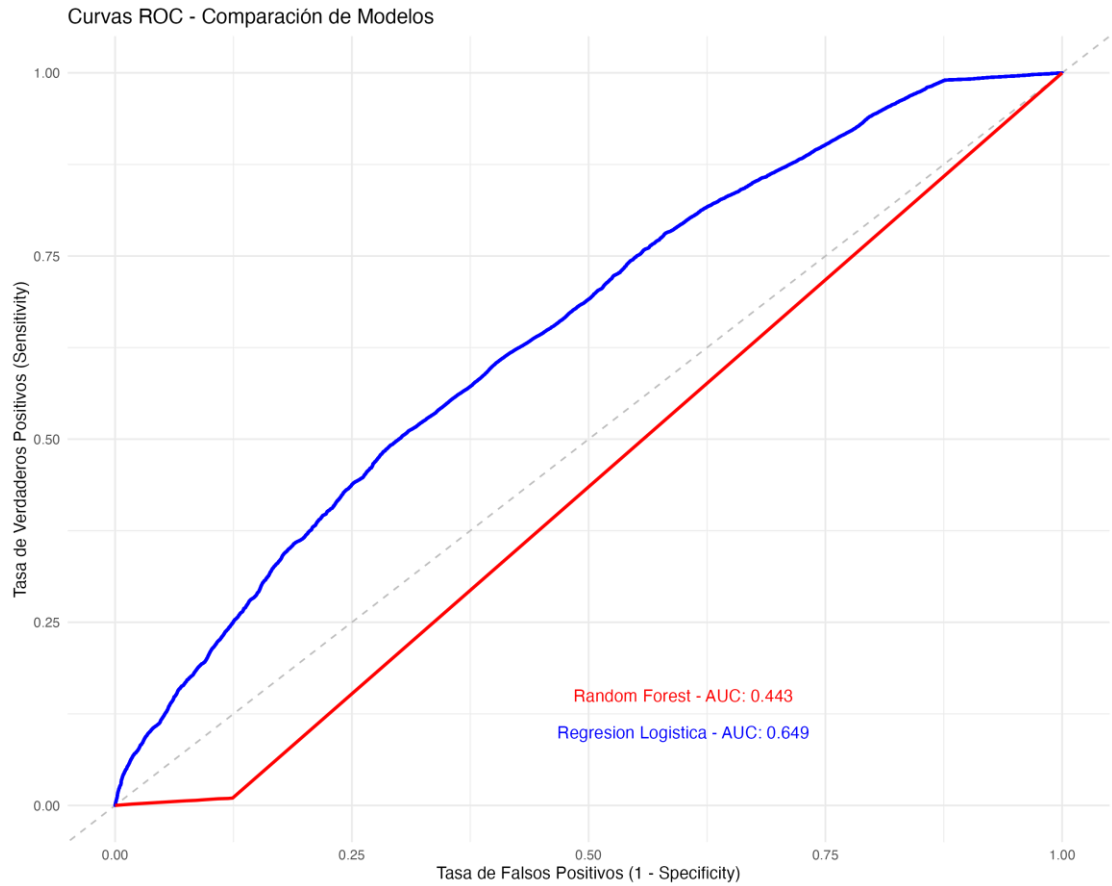
Métrica	Reg. Logística	Random Forest
Accuracy	64.2%	**67.1%**
AUC	**0.649**	0.443
F1-Score	**0.473**	0.217
Sensitivity	**43.6%**	12.4% ⚠

✓ Conclusión

- Regresión Logística es preferida: mejor balance entre precisión, sensibilidad y comprensión.
- Random Forest tuvo mejor accuracy general pero muy baja capacidad para detectar casos positivos.

ML - Resultados y Comparación de Modelos



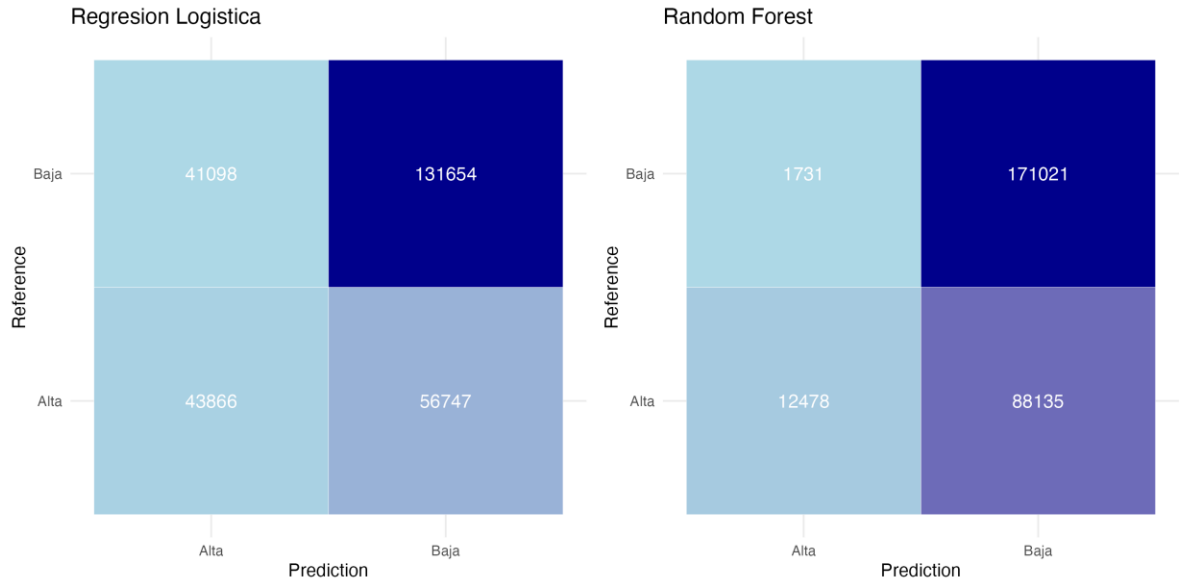


ML - Análisis de Errores

Principales Errores Observados (matrices de confusión)

✓ Random Forest clasifica casi todo como "Baja", con sensibilidad muy baja.

✓ Regresión Logística mejora la detección de "Alta", pero aún con margen de mejora.



ML - Oportunidades de Mejora

- ✓ Balance de Clases
  - La variable objetivo está desbalanceada (63% baja, 37% alta).
  - Podrían explorarse técnicas más avanzadas de resampling o SMOTE.

- ✓ Inclusión de más variables
  - Incorporar variables adicionales como tipo jurídico, fecha de creación, o redes de proveedores podría mejorar la predicción.
- ✓ Nuevos Algoritmos
  - Evaluar métodos como Gradient Boosting (XGBoost) o SVM para capturar mejor patrones no lineales.

- ✓ Evaluación contextual
  - Ajustar métricas al caso de uso: si el objetivo es identificar con seguridad los establecimientos con alta participación, priorizar sensibilidad y recall.

- ✓ Optimización técnica
  - El entrenamiento de Random Forest fue costoso (30 minutos). Puede mejorarse con reducción de dimensionalidad o sampling más agresivo.

Conclusiones

¿Qué aprendimos?

- ✓ Sobre exportaciones
  - Se puede predecir qué empresas tienen potencial exportador.
  - Políticas focalizadas son más efectivas que universales.
- ✓ Sobre empleo femenino
  - El sector importa más que la ubicación geográfica.
  - Existe un dilema entre exportar e incluir mujeres.
- ✓ Sobre geografía
  - La ubicación influye pero no determina.
  - La Patagonia sorprende ubicándose como líder en inclusión.
- ✓ Sobre sectores
  - Existen 3 grupos muy diferentes de actividades económicas.
  - Cada grupo necesita políticas específicas.

Se podrían diseñar políticas más precisas y efectivas usando estos hallazgos como herramientas.

Conclusiones

¿Qué aprendimos?

- ✓ El uso de herramientas de ciencia de datos, como la extracción y análisis de datos, el aprendizaje supervisado y no supervisado, permite descubrir patrones, modelizar la realidad y tomar decisiones informadas.

Limitaciones y Trabajo Futuro

¿Qué no incluye este estudio?

✓ Evolución en el tiempo:
- Solo analizamos 2021-2022.
- No sabemos cómo cambian los patrones año a año.

✓ Otras variables importantes:
- Salarios y productividad.
- Factores culturales regionales.
- Impacto de políticas específicas.

✓ Relaciones causales:
- Identificamos patrones, no causas definitivas.
- ¿Los sectores determinan el género, o viceversa?

Limitaciones y Trabajo Futuro

Otras consideraciones metodológicas

✓ Cobertura restringida al empleo formal:

- Solo incorpora trabajadores en relación de dependencia registrados en SIPA y AFIP.
- Excluye empleo informal, cuentapropistas, casas particulares y microestablecimientos estacionales.
- Subestima la densidad productiva real en zonas con alta informalidad

✓ Clasificación sectorial:

- Algunas empresas informan según versiones antiguas de CLAE por lo que pueden generarse errores de clasificación.
- Se suprimen actividades que no pudieron vincularse al CLAE 2010 ocasionando omisiones sectoriales puntuales en microsectores o actividades emergentes.

✓ Comparabilidad y Temporalidad:

- Es una versión exploratoria por lo que al no ser final, carece de robustez para comparaciones históricas o estudios a largo plazo

Limitaciones y Trabajo Futuro

Próximos Pasos de Investigación

- ✓ Seguimiento temporal: Ver evolución 2019-2025.
- ✓ Relevamiento de otras variables: Integrar datos de salarios y productividad y factores culturales regionales.
- ✓ Estudios causales: Evaluar impacto de políticas específicas.