

PRESENTACIÓN DEL MODELO Y ANÁLISIS DE RESULTADOS PARCIAL –APRENDIZAJE AUTOMÁTICO

Tercera entrega:

Presentación del Modelo y Análisis de Resultados

❖ Introducción

Este proyecto tiene como objetivo predecir la condición laboral de las personas en Argentina (empleado o desempleado) utilizando técnicas de aprendizaje automático, aplicadas sobre datos abiertos provistos por el INDEC a través de la Encuesta Permanente de Hogares (EPH). Identificar con precisión la situación laboral es fundamental para diseñar políticas públicas eficaces y focalizar programas sociales en poblaciones vulnerables.

❖ Origen de los datos

- Fuente: Encuesta Permanente de Hogares (EPH) – INDEC
- Trimestre: Cuarto trimestre de 2023 (T423)
- Unidades de análisis: Personas residentes en aglomerados urbanos
- Muestra seleccionada: 6.711 personas filtradas con estado laboral “ocupado” (1) o “desocupado” (2)

❖ Análisis Exploratorio

- Se realizaron visualizaciones para comprender mejor la distribución de las variables:

- Distribución de edades: los desempleados se concentran mayormente entre los 18 y 30 años, mientras que los empleados se distribuyen entre los 25 y 55 años.
- Proporción de clases: más del 95% de los casos corresponde a personas empleadas, lo que evidencia un importante desbalanceo de clases.
- Ingreso familiar e instrucción: a mayor nivel educativo, mayor proporción de personas empleadas.

❖ Pre-procesamiento

- Imputación de nulos con mediana (numéricas) y moda (categóricas) usando SimpleImputer.
- Codificación de variables categóricas con One-Hot Encoding.
- Escalado de variables numéricas con StandardScaler.
- División de datos en entrenamiento (70%) y prueba (30%).

❖ Modelos entrenados y Evaluación

Se entrenaron tres modelos clásicos de clasificación:

PRESENTACIÓN DEL MODELO Y ANÁLISIS DE RESULTADOS PARCIAL –APRENDIZAJE AUTOMÁTICO

Modelo	Accuracy	F1-Score (Desempleado)
Logistic Regression	96.81%	0.63
Decision Tree	96.29%	0.61
Random Forest	97.12%	0.67

El modelo Random Forest fue seleccionado como modelo final por su desempeño superior en precisión y balance entre clases.

❖ Evaluación de Tierra del Fuego

Se aplicó el modelo final al subconjunto correspondiente a Tierra del Fuego. Se incorporó la variable AGLOMERADO para mejorar la sensibilidad regional.

Clase	Reales	Predichos
Empleado	427	431
Desempleado	19	15

- Accuracy regional: 97.31%
- F1-score para desempleados: 0.65

Esto demuestra que el modelo puede adaptarse a contextos locales y mantener buen desempeño, incluso ante clases desbalanceadas.

❖ Variables más influyentes

- Edad
- Nivel educativo
- Rama de actividad
- Aglomerado
- Ingreso familiar total

Estas variables permiten entender mejor qué factores socioeconómicos influyen en la empleabilidad.

❖ Conclusión

El modelo desarrollado logra predecir con alta precisión la condición laboral de las personas, tanto a nivel nacional como regional. La inclusión de la variable “AGLOMERADO” permitió mejorar su adaptación local. Este proyecto aporta una base útil para extender este enfoque a otras regiones o integrarlo en decisiones de política pública.