Proyecto Final Aprendizaje Automático 2025

Abstract - Employment Status Classification Model

Accurately understanding employment status is a key step toward developing effective labor strategies and social support programs. This project presents a classification model designed to predict individuals' employment status based on demographic and socioeconomic characteristics. The dataset was sourced from the National Institute of Statistics and Censuses (INDEC), specifically from the Argentine Permanent Household Survey. After data cleaning and preprocessing, several algorithms were evaluated, with Random Forest demonstrating the best performance. It achieved 97% accuracy and significantly improved the classification of employed versus unemployed individuals. The model was trained on nationwide data and later applied to a regional subset from Tierra del Fuego. To improve accuracy at the local level, the variable "AGLOMERADO" was added during training, which helped reduce the previous overestimation of unemployment. Moreover, inactive individuals—who are outside the labor force but maybe at risk of social vulnerability—were considered as potentially vulnerable and relevant for future refinement. Although further optimization is possible, the current model offers a reliable tool for analyzing employment patterns and informing public policy with a data-driven approach.

Resumen - Modelo de Clasificación del Estado Laboral

Comprender con precisión el estado laboral es un paso clave para desarrollar estrategias laborales efectivas y programas sociales de apoyo. Este proyecto presenta un modelo de clasificación diseñado para predecir el estado laboral de las personas utilizando variables demográficas y socioeconómicas. El conjunto de datos proviene del Instituto Nacional de Estadística y Censos (INDEC), específicamente de la Encuesta Permanente de Hogares de Argentina. Luego del proceso de limpieza y preparación de los datos, se evaluaron varios algoritmos, siendo Random Forest el

que presentó el mejor desempeño. Alcanzó una precisión del 97% y mejoró notablemente la clasificación entre personas empleadas y desempleadas. El modelo fue entrenado con datos a nivel nacional y luego aplicado a un subconjunto correspondiente a Tierra del Fuego. Para mejorar la precisión regional, se incorporó la variable “AGLOMERADO” durante el entrenamiento, lo que permitió reducir la sobreestimación del desempleo observada en pruebas anteriores. Además, se consideró la inclusión de personas inactivas—quienes están fuera de la fuerza laboral pero podrían estar en situación de vulnerabilidad social—como grupo vulnerable relevante para futuras mejoras. Aunque aún puede optimizarse, el modelo actual ofrece una herramienta sólida para analizar condiciones laborales y apoyar decisiones de política pública basadas en datos.