UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

(Universidad del Perú, DECANA DE AMÉRICA)

FACULTAD DE CIENCIAS MATEMÁTICAS

CURSO: ALGORÍTMICA Y FUNDAMENTOS DE PROGRAMACIÓN

SEMANA 11

Estado del arte

DOCENTE

Oscar Benito Pacheco

ALUMNO

Ayzanoa Solano, Joao Carlos

2025 – I

Plantilla del Estado del Arte Articulo 1

Author (s): Zoraida Mamani Rodríguez.
Títle of paper: Proceso de machine learning para determinar la demanda social de puestos de empleo de profesionales de TI.
Journal: Industrial Data
Volumen(numero):25(2)
pag – pag (year): 275-287(2022)

Estado del arte que hace el autor

La autora introduce el machine learning como una disciplina de la inteligencia artificial que se apoya en la computación científica, matemáticas y estadística, a través de técnicas automatizadas como la clasificación, regresión y clustering. En particular, se destaca el clustering como una técnica no supervisada adecuada para etapas exploratorias de grandes volúmenes de información. Dentro de este marco, el trabajo plantea que la demanda social de profesionales de Tecnologías de la Información (TI) puede analizarse de manera automatizada a partir de los datos obtenidos de portales de empleo, donde se expresan las necesidades del mercado laboral.

El concepto de demanda social se relaciona con los requerimientos de productos o servicios esperados de los egresados universitarios, y está alineado con las políticas de aseguramiento de calidad de la educación superior. En este sentido, el estudio propone un proceso de machine learning con enfoque no supervisado para extraer perfiles ocupacionales de TI, construir un modelo multidimensional de dichos perfiles, aplicar técnicas de clustering para agrupar los empleos según similitudes, y finalmente analizar los resultados obtenidos. La propuesta se contextualiza dentro de un marco académico y nacional orientado a actualizar catálogos ocupacionales y contribuir a la mejora de la oferta educativa.

Motivación del autor (críticas del autor a otros trabajos)

El trabajo reconoce investigaciones previas relevantes que abordan la relación entre perfiles laborales y técnicas de machine learning. Qin et al. (2018) utilizan modelos semánticos y técnicas de aprendizaje supervisado (regresión logística, árboles de decisión, bosques aleatorios, gradient boosting) para mejorar el ajuste entre candidatos y puestos. Boselli et al. (2018) también utilizan enfoques supervisados con SVM y redes neuronales, incluyendo clasificadores como n-grams para la detección de habilidades. Estos trabajos dependen de datos previamente etiquetados, lo cual puede limitar la escalabilidad de sus soluciones.

Por otro lado, Lynch (2017) critica la subjetividad en la definición de perfiles laborales y aplica aprendizaje supervisado para predecir puestos basados en descripciones. Marrara et al. (2017) proponen un enfoque basado en modelos lingüísticos para identificar nuevas ocupaciones. Finalmente, Vinel et al. (2019) exploran el uso de clustering con diversos modelos de representación de texto (TF-IDF, word2vec, BERT, etc.) y técnicas de agrupamiento como HDBSCAN y k-means, concluyendo que k-means es eficaz si se especifica el número de clusters.

La autora, sin embargo, cuestiona la dependencia de modelos supervisados y propone una metodología más flexible mediante aprendizaje no supervisado, con énfasis en la aplicabilidad directa al contexto peruano. Esta motivación impulsa la necesidad de soluciones automatizadas, adaptables y replicables, que minimicen la intervención manual y permitan una mayor objetividad en la detección de demandas laborales.

Descripción del aporte del autor

El aporte principal del trabajo radica en el diseño e implementación de un proceso de machine learning no supervisado que permite identificar la demanda social de profesionales de TI. Este proceso se basa en dos subprocesos principales:

1. **Extracción de datos desde portales de empleo** mediante técnicas de webscraping.
2. **Entrenamiento de un modelo de clustering k-means** para agrupar perfiles laborales.

Se recolectaron 8640 anuncios de empleo desde ocho plataformas digitales, entre ellas Google Jobs Search, Buscojobs, Freelancer, LinkedIn, Indeed y Computrabajo. La extracción se realizó mediante scripts en Python, empleando la librería BeautifulSoup para navegar el DOM de los portales y almacenar los datos en una base de datos relacional.

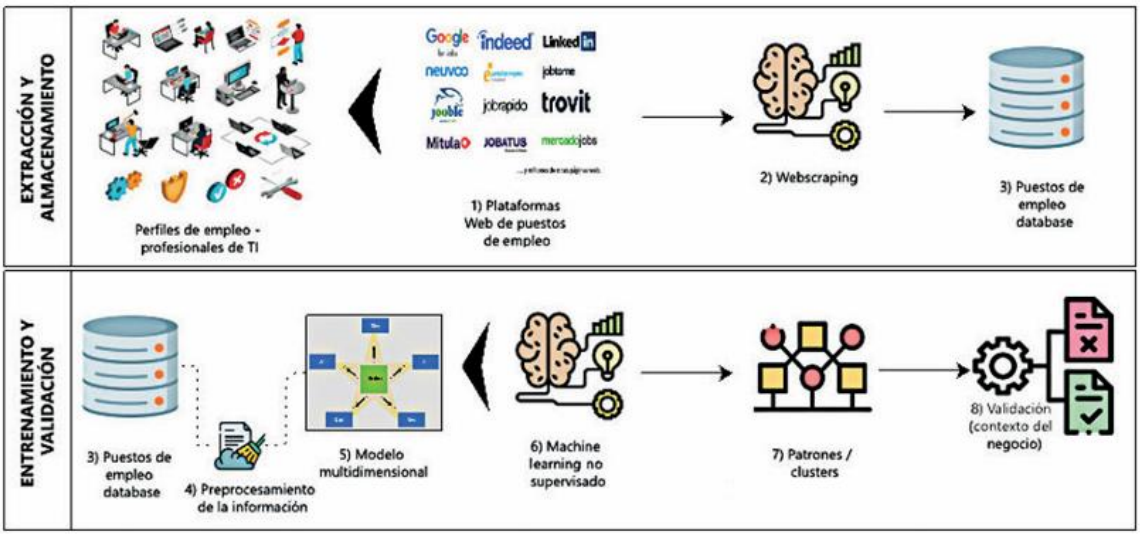

Figura 1. Proceso de *machine learning* no supervisado.

Fuente: Elaboración propia.

Figura 1. Proceso de machine learning propuesto por la autora

Posteriormente, se realizó el **preprocesamiento** de datos mediante normalización, tokenización y eliminación de ruido textual. Luego, se diseñó un **modelo multidimensional tipo copo de nieve** con 14 dimensiones relevantes como: perfil, función, habilidades, competencias, formación, beneficios, entre otros. Esta estructura permitió el análisis desde diversas perspectivas con dashboards en Power BI.

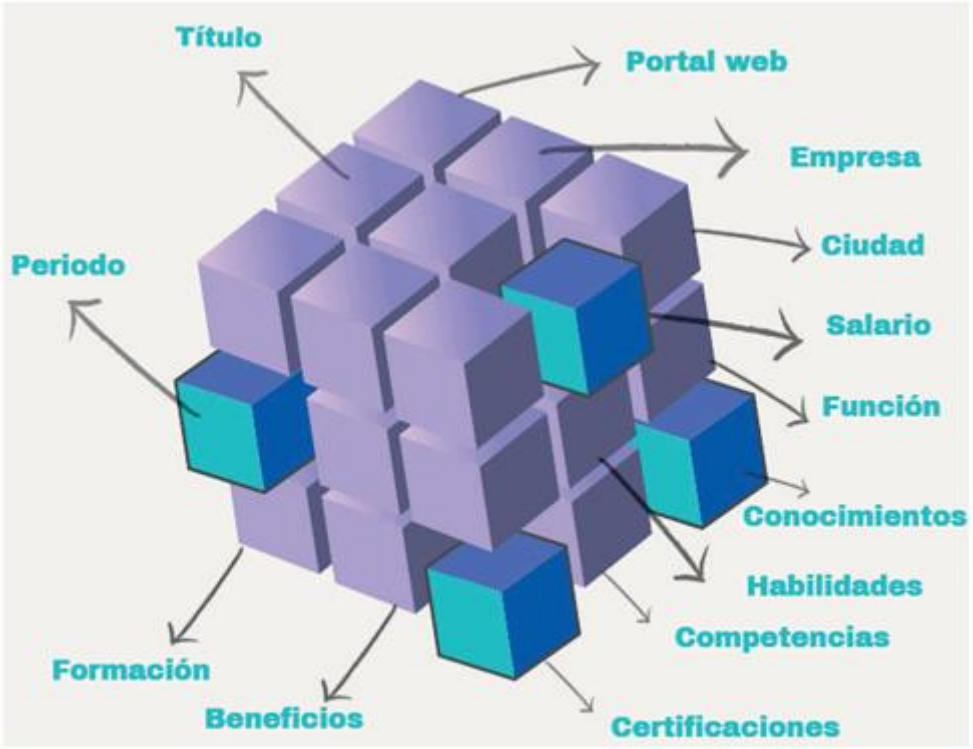

Figura 2. Modelo multidimensional.

Fuente: Elaboración propia.

Figura 2. Modelo multidimensional propuesto (fuente: autora)

El clustering se implementó mediante el algoritmo **k-means++** en el software Weka, seleccionando el número óptimo de clusters con el método del codo. Se ejecutaron modelos con datasets de 14 y 6 dimensiones. Los resultados principales fueron:

Dataset	Clusters	Iteraciones	SSE (Error)
14 dim.	15	6	34,696.82
6 dim.	15	4	16,027.00

Tabla 1. Resultados del clustering k-means con Weka

Los clústeres identificaron perfiles como "developer" y "fullstack developer" como los más comunes, habilidades clave como "comunicación efectiva" y "conocimiento en Java", y beneficios frecuentes como "estabilidad laboral". Los portales más usados fueron Google Jobs y Buscojobs, con una fuerte concentración geográfica en Lima.

Observaciones y/o críticas suyas

El estudio de Mamani Rodríguez ofrece una contribución significativa en la aplicación de machine learning al análisis laboral. Destaca por su enfoque no supervisado, lo cual es adecuado para escenarios donde no existe información etiquetada o categorizada previamente. El uso de un modelo multidimensional junto con dashboards interactivos facilita el análisis desde distintas dimensiones, lo que puede resultar útil para tomadores de decisiones.

No obstante, se podría enriquecer el modelo con análisis semántico profundo de los perfiles laborales mediante embeddings modernos (p. ej., BERT), que podrían capturar relaciones más complejas entre conceptos laborales. Asimismo, una validación con expertos del sector permitiría contrastar la calidad de los clústeres y su interpretabilidad práctica. Finalmente, el enfoque podría escalarse a otras profesiones para robustecer su utilidad como herramienta nacional de monitoreo de tendencias laborales.

Plantilla del Estado del Arte Articulo 2

Autor: **Gehad ElSharkawy, Yehia Helmy, Engy Yehia**
Título del artículo: Employability Prediction of Information Technology Graduates Using Machine Learning Algorithms
Revista: International Journal of Advanced Computer Science and Applications (IJACSA)
Volumen: Vol. 13, No. 10, 2022
Páginas: 359-366

Estado del arte que hace el autor

El artículo parte del reconocimiento de que el desajuste entre la educación superior y las necesidades del mercado laboral es una de las principales causas del desempleo juvenil en Egipto. En ese contexto, se destaca el valor del capital humano y la necesidad de formar profesionales con habilidades alineadas a las demandas actuales.

La literatura previa muestra un interés creciente en utilizar algoritmos de aprendizaje automático (machine learning, ML) para analizar la empleabilidad de los egresados. Sin embargo, los estudios existentes han estado limitados por el uso de pocas variables o enfoques parciales. Además, se resalta que la mayoría de los estudios se han centrado en modelos supervisados aplicados a bases de datos institucionales, sin integrar información de empleadores ni considerar todas las dimensiones que influyen en la empleabilidad (habilidades blandas, técnicas, actividades co-curriculares y características demográficas).

El estudio también contextualiza el uso del ML como una herramienta que permite predecir patrones a partir de grandes volúmenes de datos, a diferencia de los algoritmos tradicionales, lo que hace viable anticiparse a los factores clave que condicionan la inserción laboral de los egresados en un mercado dinámico.

Motivación del autor (críticas del autor a otros trabajos)

Los autores analizan numerosos estudios previos sobre predicción de empleabilidad, evidenciando sus limitaciones. En trabajos como los de Othman et al. (2021), se emplearon algoritmos como DT, ANN y SVM para predecir la empleabilidad, pero con conjuntos de datos limitados y sin considerar la experiencia laboral directa. Otros estudios utilizaron bases de datos institucionales sin integrar encuestas a empleadores.

Se observa que, aunque se utilizaron modelos como redes neuronales, SVM, KNN, NB, LR, RF y DT, las variables analizadas se reducen a aspectos técnicos y demográficos, excluyendo elementos clave como habilidades interpersonales, co-curriculares o conocimientos en demanda actual (por ejemplo, IA, IoT o ciberseguridad). Además, muchos modelos carecen de validación práctica con el mercado laboral.

Por ello, los autores plantean como críticas la reducida diversidad de características analizadas, la falta de contextualización nacional (caso Egipto) y la escasa inclusión de la voz de los empleadores en los modelos. Estas falencias motivan el desarrollo de un enfoque más integral, con encuestas a egresados y empleadores y el uso combinado de diversos algoritmos.

Descripción del aporte del autor

El estudio propone un modelo predictivo de empleabilidad de egresados de TI egipcios basado en algoritmos de ML y un conjunto de datos recolectado por los propios autores. Se aplicaron cinco algoritmos de clasificación binaria:

- Árbol de Decisión (DT)
- Bayesiano Gaussiano (NB)
- Regresión Logística (LR)
- Random Forest (RF)
- SVM

El conjunto de datos proviene de encuestas a egresados de programas de TI y a empleadores, e incluye 296 registros clasificados en dos clases: "calificado" y "no calificado".

Metodología general (Figura 1 del artículo):

1. Recolección de datos mediante encuestas.
2. Preprocesamiento y limpieza (valores nulos, ruido, normalización).
3. División en conjuntos de entrenamiento (80%) y prueba (20%).
4. Aplicación de los 5 algoritmos.
5. Evaluación mediante matriz de confusión: Precisión, Recall, F1.

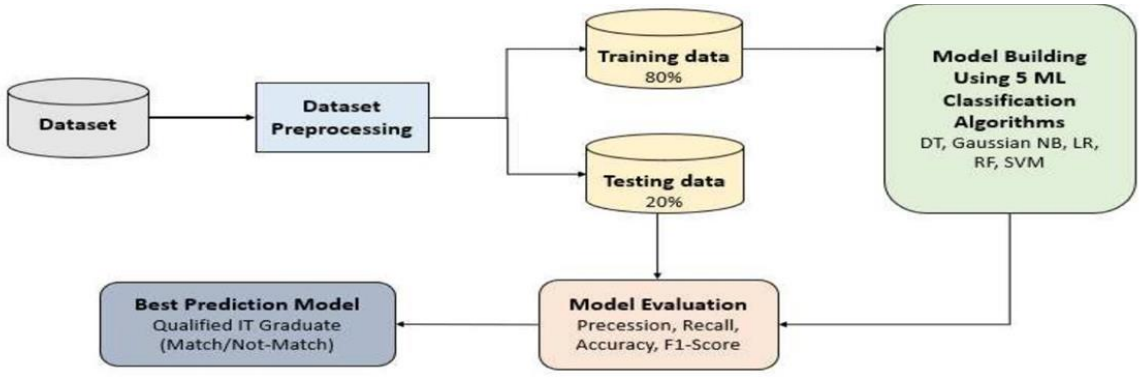

Fig. 1. La metodología de la investigación.

Categorías analizadas (Tabla II):

- Habilidades técnicas
- Habilidades blandas
- Habilidades en demanda (IA, IoT, ciencia de datos, etc.)
- Características demográficas y formación previa

Resultados clave:

- La mayor parte de los encuestados no califica como "empleable" según el modelo.
- El algoritmo de árboles de decisión (DT) logró la mayor precisión con un 98%, seguido de SVM (98%) y NB (69%).

Gráficos destacados:

- Matriz de correlación de variables.

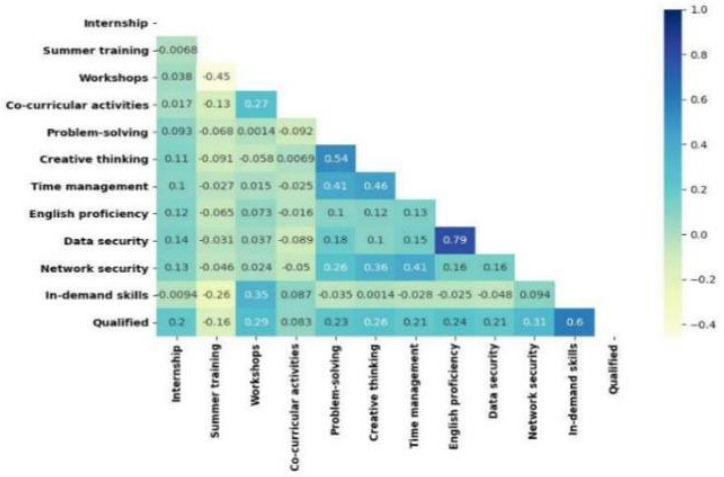

Fig. 2. Matriz de correlación de características seleccionadas.

- Distribución de graduados calificados vs no calificados (Figura 3).

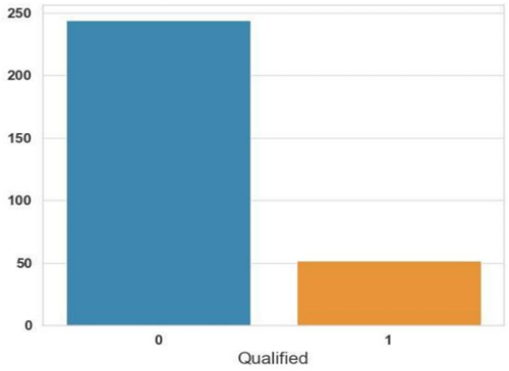

Fig. 3. Recuento de clase de empleabilidad (calificado/no calificado).

- Distribución por tipo de entrenamiento recibido (Figura 4).

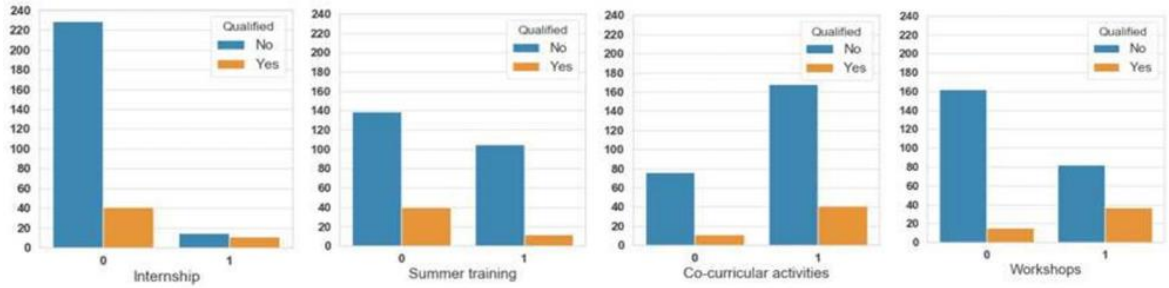

Fig. 4. Distribución de los encuestados en función de formaciones.

- Distribución de habilidades blandas y técnicas.

El estudio demuestra que ciertos factores como pasantías, talleres, habilidades blandas y dominio del inglés tienen fuerte correlación con la empleabilidad, y propone este enfoque como apoyo para reformar los planes de estudio.

Observaciones y/o críticas suyas

Este artículo presenta una investigación aplicada de alto valor sobre la empleabilidad de egresados en TI usando técnicas de ML, con una clara ventaja por el uso de datos primarios (encuestas a empleadores y egresados). Su estructura metodológica es sólida y el enfoque multicriterio permite una mejor evaluación de las competencias demandadas.

Sin embargo, la muestra de 296 registros podría ampliarse para mejorar la generalización. También sería relevante probar técnicas de aprendizaje no supervisado o aprendizaje profundo para comparar resultados. Finalmente, el modelo podría complementarse con un dashboard interactivo para uso institucional que permita a las universidades monitorear en tiempo real la preparación de sus egresados.

Plantilla del Estado del Arte Articulo 3

Author (s): Nik Dawson, Marian-Andrei Rizoiu, Benjamin Johnston, Mary-Anne Williams.
Título of paper: Predicting Skill Shortages in Labor Markets: A Machine Learning Approach.
Journal: Proceedings of the IEEE International Conference on Big Data (BigData 2020)
Volumen(numero):N/A
pag – pag (year): (2020)

Estado del arte que hace el autor

Los autores abordan un problema crítico en los mercados laborales globales: la escasez de habilidades. Este fenómeno genera impactos negativos como pérdida de productividad, oportunidades económicas desperdiciadas y desventajas competitivas para países y empresas. A través de un enfoque basado en ciencia de datos, el artículo explora tres grandes problemas abiertos: (1) identificar qué habilidades son más importantes en ocupaciones con escasez, (2) predecir dichas escaseces a futuro y (3) determinar qué variables son más predictivas.

Las investigaciones previas han empleado métodos como conteo de frecuencias en anuncios de empleo o encuestas a empleadores, pero estas técnicas tienen limitaciones para diferenciar habilidades comunes de las especializadas. Además, ha existido una falta de consenso sobre qué variables mejor predicen cambios en la clasificación de escasez de una ocupación, y poca exploración del uso de modelos predictivos avanzados.

El estudio se enfoca en el mercado laboral australiano, utilizando una combinación inédita de datos: más de 7.6 millones de anuncios de empleo en línea y 20 medidas oficiales de fuerza laboral de la Oficina Australiana de Estadísticas. Se analiza un conjunto de 132 ocupaciones estandarizadas de 2012 a 2018, y se utiliza el modelo de clasificación XGBoost para predecir cambios anuales en la escasez de habilidades, con un desempeño F1 de hasta 83%.

Motivación del autor (críticas del autor a otros trabajos)

A pesar de múltiples estudios sobre la escasez de habilidades, muchas investigaciones previas se han basado en métodos indirectos o poco representativos. Por ejemplo, las encuestas a empleadores suelen reflejar percepciones subjetivas, influenciadas por factores internos de la empresa como salarios poco competitivos o ubicaciones desfavorables.

Otros estudios han usado medidas indirectas como crecimiento de salarios, horas trabajadas o tasas de vacantes, pero estos indicadores no siempre se combinan con modelos predictivos robustos. Además, muchos trabajos usan frecuencias simples de aparición de habilidades en anuncios de empleo, sin considerar su prevalencia general, lo que distorsiona la importancia real de ciertas habilidades dentro de una ocupación.

Los autores critican esta falta de rigurosidad metodológica y plantean un enfoque más robusto: combinar datos de demanda (anuncios de empleo) y oferta laboral (estadísticas oficiales) bajo un marco de aprendizaje automático. También destacan la importancia de incorporar características temporales (como el estado de escasez en años previos) para mejorar la predicción.

En resumen, la motivación del artículo radica en superar las limitaciones de enfoques anteriores mediante un análisis más profundo, cuantitativo y basado en datos reales del mercado laboral.

Descripción del aporte del autor

1. Diseño metodológico

Los autores construyen un marco predictivo usando el modelo XGBoost, que combina datos de demanda laboral (anuncios de empleo) y oferta laboral (encuestas oficiales) para predecir la escasez de habilidades en ocupaciones específicas un año hacia el futuro. Para ello:

- Se recolectaron 7,697,568 anuncios de empleo de Australia entre 2012 y 2018.
- Se incorporaron 20 variables del mercado laboral (por ejemplo: personas empleadas, desempleadas, horas trabajadas).
- Se utilizaron datos del Departamento Federal de Educación y Empleo de Australia para validar el modelo (clasificación binaria: "En escasez" o "No en escasez").

2. Análisis de habilidades

Se comparan dos métodos para determinar la importancia de habilidades en ocupaciones en escasez:

- **Frecuencia de aparición**: método tradicional que solo cuenta la cantidad de veces que una habilidad aparece.
- **Ventaja Comparativa Revelada (RCA)**: ajusta por la frecuencia global de habilidades, destacando las más específicas y emergentes.

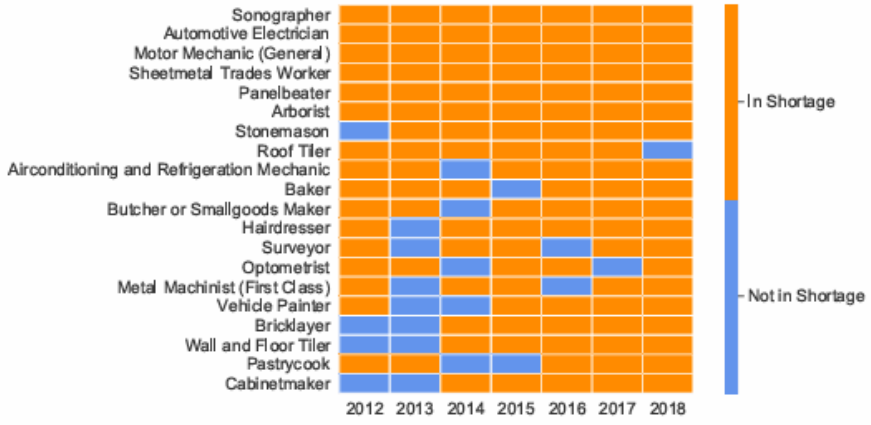

Fig. 2: Top occupations most *In Shortage* at the ANZSCO 6-digit occupational level.

3. Resultados de predicción

- El modelo con mejores resultados incluye características autorregresivas (estado de escasez previo).
- F1 macro promedio alcanzado: **83%**.
- Aun sin datos históricos (modelo sin autoregresivos), el desempeño fue sólido (F1 ≈ 72%), lo cual lo hace replicable en otros países.
- Predicciones de cambios de estatus (ocupaciones que pasan de "No en escasez" a "En escasez") fueron más difíciles, pero el modelo funcionó mejor con datos de demanda y oferta laboral.

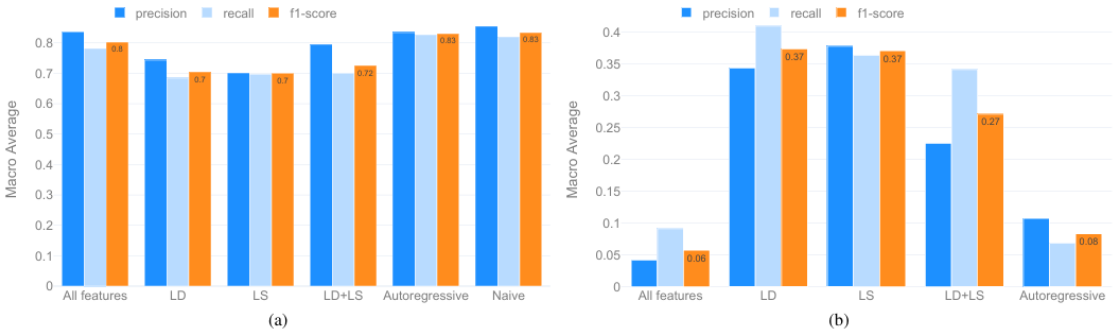

Fig. 4: **Skills Shortage prediction results**: **(a)** While the prediction results are highly auto-regressive, Labor Demand and Labor Supply features alone (and combined) perform almost as well for predicting occupational shortages; **(b)** Labor Demand and Labor Supply features perform better than other features at predicting shortage status changes of occupations.

4. Importancia de variables

Entre las variables más predictivas se encontraron:

- Horas trabajadas (6 de las 20 más importantes).
- Educación requerida.
- Experiencia requerida.
- Salario mediano.

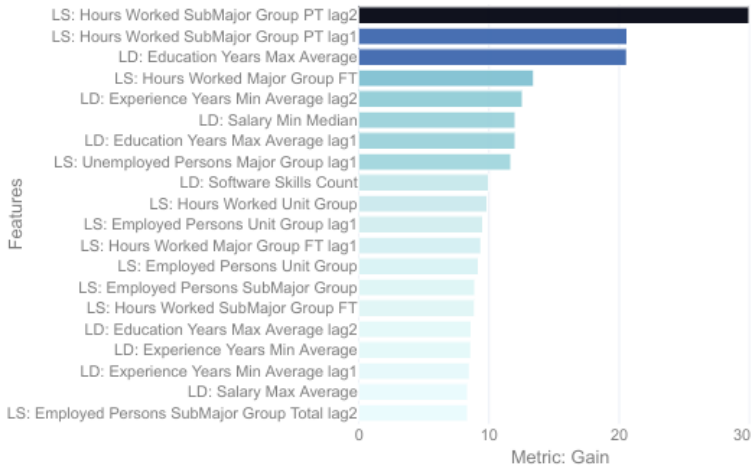

Fig. 5: Feature importance of Labor Demand and Labor Supply feature model.

Observaciones y/o críticas suyas

El trabajo representa un avance significativo en el estudio de la escasez de habilidades. Su enfoque basado en datos reales, tanto de oferta como de demanda, y el uso de aprendizaje automático permite obtener predicciones más precisas y aplicables en políticas públicas.

Sin embargo, el estudio presenta algunas limitaciones importantes:

- **Dependencia de datos oficiales**: El modelo se entrena con datos históricos de Australia. No todos los países cuentan con una fuente similar, lo que limita su generalización.
- **Sesgo ocupacional**: La mayoría de las ocupaciones analizadas pertenecen a los grupos de "Técnicos" y "Profesionales", lo que podría reducir la validez de los resultados para otros sectores.
- **Dificultad para predecir cambios de estado**: Aunque los modelos son buenos prediciendo si una ocupación permanecerá igual, tienen menor precisión cuando hay un cambio en el estado de escasez, precisamente donde más valor podrían aportar.

A pesar de ello, la propuesta tiene un gran valor práctico. Puede ser útil para gobiernos que buscan ajustar programas educativos, políticas migratorias o intervenciones laborales. Para replicarlo en otros contextos, sería clave recopilar datos equivalentes y adaptar los modelos a las condiciones locales.