Trabajo Final de análisis de datos aplicando herramientas de Machine Learning

Bracaccini Agostina; Córdoba Julieta; Milanese Luisina

UNRaf, Universidad Nacional de Rafaela, Santa Fe, Argentina.
Aprendizaje Automático y Grandes Datos (IC)

Autor Corresp.: luisinamilanese03@gmail.com

Resumen

Este trabajo analiza la inserción laboral y los niveles salariales de más de 800.000 graduados universitarios argentinos mediante la aplicación de técnicas de *machine learning* a datos administrativos del Ministerio de Producción y Trabajo. El objetivo principal fue identificar patrones estructurales asociados al empleo formal, así como desigualdades salariales vinculadas con la disciplina, la región, el género y el tipo de institución universitaria. Para ello se desarrolló un proceso integral de análisis que incluyó limpieza del *dataset*, eliminación de *outliers*, codificación de variables categóricas mediante *One Hot Encoding* y normalización de variables numéricas mediante *StandardScaler*. Posteriormente se implementaron modelos supervisados de regresión y clasificación, junto con métodos no supervisados como *K-Means*.

Los resultados mostraron que los modelos de regresión presentan bajo poder predictivo debido a la alta variabilidad salarial y a factores económicos no registrados en el *dataset*. En contraste, los modelos de clasificación alcanzaron un desempeño moderado, permitiendo distinguir entre niveles salariales altos y bajos e identificar variables con mayor influencia. El análisis de clustering reveló perfiles diferenciados de graduados que reflejan brechas persistentes en el mercado laboral argentino. En conjunto, el estudio demuestra el valor del aprendizaje automático para comprender tendencias laborales y orientar decisiones institucionales.

Palabras clave: aprendizaje automático, inserción laboral, predicción salarial, educación superior, modelos supervisados.

1. Introducción

La relación entre la formación universitaria y las oportunidades laborales reales es un tema que preocupa tanto a quienes estudian como a quienes diseñan políticas educativas. En Argentina, el sistema universitario creció mucho en los últimos años, tanto en cantidad de estudiantes como en diversidad de carreras e instituciones. Pero ese crecimiento no siempre se traduce en mejores condiciones laborales, y siguen apareciendo diferencias importantes según la disciplina, la región, el género o el tipo de universidad. Por eso resulta necesario analizar con más detalle cómo se vinculan estos factores con la inserción laboral y los salarios de los graduados, especialmente en un contexto tan diverso y desigual como el argentino.

Si bien existen trabajos que ya aplicaron herramientas estadísticas o de *machine learning* a temas relacionados con empleo, salarios o rendimiento académico, la mayoría se apoya en encuestas o bases de datos pequeñas. Estos estudios suelen emplear modelos de regresión, algoritmos de clasificación como *Random Forest* o *Gradient Boosting*, o métodos no supervisados como K-Means. Sin embargo, al trabajar con muestras reducidas, muchas veces no logran captar desigualdades estructurales o variaciones regionales más profundas. Tampoco integran, en un mismo análisis, información educativa, sociodemográfica y salarial a gran escala.

Ahí aparece la brecha que este trabajo busca abordar: la falta de investigaciones basadas en datos masivos que permitan observar patrones laborales de manera más completa. El acceso a un conjunto de más de 800.000 registros del Ministerio de Producción y Trabajo abre la posibilidad de

examinar la inserción laboral y los salarios desde otro lugar, utilizando técnicas de *machine learning* que permiten detectar tendencias que no siempre se ven a simple vista. A partir de esto surge la pregunta que guía el estudio: ¿es posible identificar patrones de inserción laboral y niveles salariales usando únicamente la información académica y sociodemográfica disponible en este tipo de datos masivos? Analizar esta pregunta implica no sólo probar modelos, sino también interpretar qué desigualdades emergen cuando se mira el mercado laboral desde una perspectiva más amplia.

Para avanzar en esta dirección, el trabajo se desarrolla de manera progresiva. Primero, se presenta el análisis exploratorio de datos, donde se examinan las características principales del dataset y se abordan las preguntas iniciales que guiaron la primera parte del trabajo, permitiendo reconocer desigualdades, tendencias salariales y diferencias en la inserción laboral según disciplina, región, género y tipo de institución. Luego, se describen los modelos supervisados aplicados para estudiar la predicción salarial y la probabilidad de empleo, explicando cómo se entrenaron y evaluaron a partir de las variables que el propio EDA identificó como más relevantes. Seguido de esto, se introducen los métodos no supervisados que permitieron agrupar graduados según sus características compartidas, sin imponer etiquetas previas, y que ayudaron a observar patrones estructurales más amplios. Finalmente, se discuten los resultados de todas estas etapas y se reflexiona sobre el aporte que pueden tener estas herramientas para comprender las trayectorias laborales de los egresados y para pensar decisiones educativas más informadas y contextualizadas.

2. Metodología

La metodología de este trabajo combina procedimientos clásicos de ciencia de datos con técnicas de aprendizaje automático orientadas a comprender un fenómeno social complejo. No se trató únicamente de entrenar modelos, sino de recorrer un proceso iterativo que incluyó exploración, depuración, análisis preliminar, modelado y reflexión crítica sobre lo que los datos permitían concluir. Cada una de estas etapas fue ajustándose en función de los resultados intermedios, lo que permitió avanzar con una mirada progresiva y fundada tanto en la técnica como en la interpretación.

2.1. Análisis Exploratorio de Datos (EDA)

El punto de partida fue un análisis exploratorio de datos diseñado para conocer la estructura del dataset y responder un conjunto de preguntas iniciales que orientaron toda la investigación. El *objetivo* no era aún predecir, sino observar: identificar qué variables resultaban más relevantes, qué patrones surgían espontáneamente y cuáles eran las desigualdades más visibles dentro de los datos.

El *dataset* incluía más de 800.000 registros con información académica, demográfica y laboral de graduados universitarios argentinos. Esta magnitud hizo necesario un relevamiento inicial de la calidad de los datos: conteo de filas, tipos de variables, detección de valores faltantes y análisis de distribuciones. Durante esta etapa se identificaron salarios atípicos, edades improbables, categorías con muy poca representación y registros inconsistentes que requerían depuración. Se aplicaron criterios estadísticos (como el rango intercuartílico) para remover *outliers* y se imputaron valores faltantes utilizando la mediana para variables numéricas y la moda para categóricas.

Una parte importante del EDA se centró en explorar diferencias salariales según disciplina, región, género y tipo de institución. A través de histogramas, *boxplots* y comparaciones de grupos, se observaron brechas persistentes: los varones tendieron a concentrarse en rangos salariales más altos, mientras que las mujeres presentaron una distribución más homogénea, pero con valores menores en la mayoría de las áreas. También se identificaron desigualdades regionales, con la región pampeana y CABA acumulando los salarios más elevados y los mayores niveles de empleo formal.

Otro eje del análisis fue la inserción laboral. Se examinó qué disciplinas mostraban mayor proporción de graduados empleados, cómo influía el tipo de institución (pública o privada) y qué regiones concentraban más oportunidades laborales. En este sentido, áreas como Economía, Administración e Ingeniería exhibieron los porcentajes más altos de empleo, mientras que Educación y Artes concentraron los índices más bajos. Estas observaciones

iniciales permitieron anticipar cuáles variables tendrían impacto en los modelos de clasificación.

El EDA también permitió estudiar la relación entre la edad al egreso, los años de experiencia aproximados y el salario. Aunque existió una tendencia general a que los ingresos crecieran levemente con la edad, la influencia de la disciplina y la región resultó mucho más determinante. Finalmente, se analizó el tipo de título (pregrado, grado, posgrado), observándose que los posgrados tendían a mejorar los salarios, pero con variaciones significativas entre áreas.

En conjunto, esta primera parte del trabajo permitió identificar patrones estructurales y desigualdades profundas dentro de los datos. Más importante aún, orientó todo el modelado posterior: señaló qué variables eran relevantes, qué relaciones podían ser modeladas y qué limitaciones tenía el *dataset*.

2.2. Limpieza, preparación y pre-procesamiento

Con los hallazgos del EDA, se procedió a la preparación del *dataset* para el entrenamiento de modelos. Este paso fue crítico porque los algoritmos de aprendizaje automático requieren estructuras limpias, numéricas y coherentes.

Primero, se realizó la depuración: eliminación de registros con datos imposibles, recorte de *outliers* extremos y estandarización de campos. Se generaron variables derivadas, como la edad al momento del egreso, que aportaron información más precisa que la edad actual. Para las variables categóricas se aplicó codificación mediante *One Hot Encoding*, evitando imponer órdenes artificiales entre categorías. Las variables numéricas se escalaron mediante *StandardScaler,* especialmente importante para modelos sensibles a la magnitud de los valores, como *KNN, SVM o K-Means.*

Este pre-procesamiento permitió obtener una matriz de datos consistente, balanceada y preparada para el entrenamiento.

2.3. Modelos supervisados: predicción y clasificación

La siguiente etapa consistió en aplicar modelos supervisados para abordar dos problemas distintos: la predicción del salario y la predicción de la inserción laboral.

Predicción salarial (modelos de regresión)

Se comenzó con modelos lineales, como *Regresión Ridge*, para evaluar la relación inicial entre las variables y el salario. Estos primeros modelos obtuvieron valores bajos de R^2, lo que indicó que el salario era extremadamente difícil de predecir con las variables disponibles. Luego se avanzó a modelos más complejos, como *Random Forest Regressor, Gradient Boosting Regressor y HistGradientBoosting*, que permiten capturar relaciones no lineales. Aunque estos algoritmos redujeron levemente el error, el poder predictivo permaneció bajo. Este resultado llevó a transformar el problema en una tarea de clasificación por rangos salariales.

Clasificación salarial e inserción laboral

Se crearon categorías de salario (“bajo”, “medio”, “alto”) y luego una versión binaria (“alto” vs. “bajo”). También se definió la variable binaria “trabaja / no trabaja” para predecir la inserción laboral. Se aplicaron modelos como *Logistic Regression, Decision Tree, Random Forest y Gradient Boosting.* La Regresión Logística mostró el mejor equilibrio entre precisión y estabilidad, con métricas en torno al 56%.

2.4. Modelos no supervisados: *clustering*

Para observar patrones sin etiquetas previas, se aplicaron técnicas no supervisadas. El método principal fue K-Means, utilizando el método del codo para seleccionar k = 5 clusters. Estos grupos revelaron perfiles diferenciados según disciplina, género, región y salario promedio. El clustering puso de manifiesto desigualdades estructurales que complementaron la lectura obtenida de los modelos supervisados.

Para visualizar estos resultados se aplicó Análisis de Componentes Principales (PCA), que permitió reducir la dimensionalidad y representar los *clusters* en un plano. También se construyeron gráficos de burbujas que relacionaron edad promedio y salario por *cluster*, reforzando la interpretación cualitativa.

2.5. Criterios de evaluación y validación

La evaluación de los modelos se realizó con métricas específicas para cada tipo de problema.

- En regresión: MAE, RMSE, R^2.
- En clasificación: *accuracy, precision, recall y F1-score*.
- En *clustering*: análisis cualitativo de

cohesión, separación y composición de grupos.

Se aplicó validación cruzada (k=5) para medir la estabilidad de los resultados y evitar conclusiones dependientes de una sola partición. Además, se ajustaron hiperparámetros mediante *GridSearchCV* y se experimentó con distintos umbrales de decisión para observar cambios en el rendimiento.

3. Resultados

Los resultados obtenidos se presentan en dos grandes partes que reflejan el recorrido del trabajo: primero, los hallazgos provenientes del análisis exploratorio de datos (EDA), donde se abordaron las preguntas iniciales de la investigación y se identificaron patrones estructurales; y luego, los resultados de los modelos de aprendizaje automático, tanto supervisados como no supervisados, que permitieron profundizar esas tendencias y evaluar la capacidad de los algoritmos para capturar la complejidad del fenómeno laboral y salarial de los graduados universitarios en Argentina.

3.1. Resultados del Análisis Exploratorio de Datos

El análisis exploratorio permitió construir una primera mirada integral del *dataset* y responder a las preguntas planteadas en la etapa inicial del trabajo. Una de las observaciones más claras fue la persistencia de desigualdades salariales según género. En casi todas las disciplinas, los varones presentaron salarios más altos que las mujeres, aunque la magnitud de esa brecha varió entre áreas.

Las carreras económicas y administrativas mostraron los ingresos más elevados, mientras que Educación, Artes y algunas disciplinas de Ciencias Sociales concentraron los valores más bajos. Esta desigualdad no solo se manifestó en los valores centrales de los salarios, sino también en su dispersión: los varones tendieron a ocupar una mayor proporción de salarios altos, mientras que las mujeres se distribuyeron de manera más homogénea en rangos medios y bajos.

Al analizar el tipo de institución, si bien se observaron diferencias entre universidades públicas y privadas, estas fueron relativas cuando se introdujo la variable disciplina. Las universidades públicas mostraron, en promedio, una inserción laboral ligeramente superior; sin embargo, la disciplina explicó más variación que la gestión institucional. Esto sugiere que el campo profesional es un factor más determinante que el tipo de institución donde se obtuvo el título.

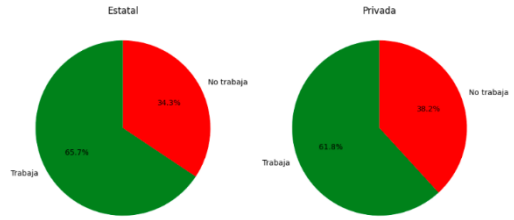

Ilustración 1: Gráfico de torta que representa la inserción laboral de graduados de universidades públicas vs. privadas

Las diferencias regionales también fueron evidentes. Las regiones pampeana y CABA concentraron los salarios más altos y los niveles más elevados de empleo formal. En contraste, el NOA (Noroeste Argentino) y el NEA (Noreste Argentino) mostraron porcentajes mayores de desocupación o inserción informal. Esta desigualdad territorial ya había sido anticipada por diferentes fuentes consultadas, pero el *dataset* permitió visualizarla con claridad cuantitativa. La región aparece como un condicionante fuerte del nivel de ingreso y de la probabilidad de obtener empleo registrado.

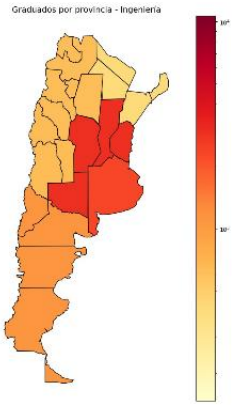

Ilustración 2: Mapa que muestra la concentración de graduados en Ingeniería, con mayor intensidad en las provincias de la región centro.

En relación con la edad, los resultados mostraron que los salarios tienden a aumentar levemente con la edad al egreso, aunque esta tendencia no fue uniforme. En varias disciplinas, especialmente las vinculadas al sector privado, la edad se asoció a mayores ingresos, posiblemente por experiencia acumulada. En otras áreas, como docencia o salud, este crecimiento salarial fue más moderado y no mostró una relación lineal.

También se exploró el tipo de título. Los graduados con posgrado tendieron a presentar salarios mayores que quienes alcanzaron solo un título de grado o

pregrado, aunque este efecto varió notablemente según la disciplina. En áreas donde los salarios están más regulados, como la docencia, el impacto del posgrado fue menor, mientras que en sectores competitivos del ámbito privado la diferencia fue más marcada.

En conjunto, esta primera parte permitió identificar patrones de desigualdad transversales: brechas por disciplina, por género, por región y por nivel formativo. Estas observaciones iniciales fueron esenciales para orientar el modelado posterior, ya que permitieron seleccionar las variables más relevantes y anticipar por qué ciertas relaciones podrían ser difíciles de predecir con modelos matemáticos.

3.2. Resultados de los modelos de regresión (predicción salarial)

La regresión salarial reveló inmediatamente la dificultad de predecir un valor numérico tan volátil como el salario utilizando solo las variables estructurales disponibles. Los modelos lineales mostraron un rendimiento muy bajo: *Ridge Regression* arrojó valores de R^2 cercanos a 0, e incluso negativos. Esto indicaba que el modelo no lograba explicar adecuadamente la variabilidad del salario, lo que era consistente con la enorme dispersión observada en el EDA.

La incorporación de modelos de mayor complejidad, incluyendo *Random Forest Regressor, Gradient Boosting y XGBoost*, permitió una leve reducción del error, pero aun así el desempeño continuó siendo limitado. El MAE se mantuvo alrededor de 15 millones de pesos anuales, mientras que el RMSE alcanzó valores entre 18 y 20 millones, y el R^2 siguió siendo muy bajo. Los modelos más sofisticados capturaron mejor las no linealidades, pero la información disponible resultó claramente insuficiente para predecir de manera precisa un salario individual.

3.3. Clasificación salarial

Debido a las limitaciones propias de la regresión, el problema se reformuló como una tarea de clasificación. Al dividir el salario en rangos, primero en tres categorías y luego en dos (“alto” y “bajo” según la mediana), fue posible estabilizar el comportamiento de los modelos. Esta transformación redujo la sensibilidad a la dispersión extrema de la variable y permitió detectar relaciones más robustas.

En esta etapa, los modelos supervisados mostraron desempeños más consistentes. *Logistic Regression* fue el modelo con mejor rendimiento general, con una precisión entre 55% y 57%.

3.4. Clasificación de inserción laboral

La predicción del empleo (divida en los grupos “Trabaja” y “No trabaja”) mostró un comportamiento más estable que la predicción salarial. Una vez codificadas y estandarizadas las variables, los modelos supervisados ofrecieron resultados moderados pero coherentes.

KNN y *Naive Bayes* brindaron desempeños relativamente bajos debido al ruido y la independencia asumida entre variables. *Decision Tree* mejoró ligeramente los resultados, pero fue la Regresión Logística la que mostró un rendimiento más sólido, alcanzando una precisión cercana al 69% y manteniendo ese comportamiento a lo largo de múltiples iteraciones y esquemas de validación. Los coeficientes de este modelo permitieron identificar las variables más influyentes: disciplina y región fueron nuevamente centrales, seguidas por tipo de título y género, mientras que la edad y la gestión institucional mostraron menor impacto.

El resultado más importante fue que los

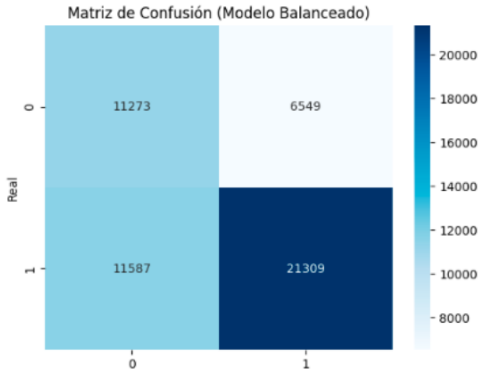

Ilustración 3:Muestra el desempeño de la clasificación entre “trabaja” y “no trabaja”, indicando los aciertos y errores para cada categoría.

modelos captaron patrones estructurales: las posibilidades de estar empleado no eran aleatorias, sino que dependían de forma consistente de combinaciones entre disciplina, región y género.

3.5. Resultados del *clustering*

El análisis no supervisado permitió una aproximación distinta: en lugar de predecir, se buscó descubrir. A partir de K-Means con K=5 —valor seleccionado por el método del codo— se identificaron cinco perfiles claramente diferenciados de graduados.

El primer grupo correspondió al *cluster 0*, formado mayoritariamente por mujeres

jóvenes de alrededor de 29 años, egresadas de carreras de Economía y Administración y provenientes del NEA, NOA y otras regiones del interior. Su inserción laboral rondó el 67,5%, con salarios ubicados en el tramo medio. El *cluster 1*, compuesto exclusivamente por varones jóvenes de CABA con títulos de grado en áreas económicas y administrativas, mostró salarios medios y una inserción laboral cercana al 66%, representando el perfil urbano profesional masculino más tradicional del mercado laboral argentino.

El *cluster 2* se destacó por agrupar en su mayoría a mujeres de mayor edad, con un promedio de 45 años al egreso, muchas de ellas provenientes de Buenos Aires o la región pampeana y vinculadas a carreras docentes, administrativas o del sector público. Presentaron la tasa de inserción más alta de todos los grupos (74,8%), probablemente debido a trayectorias laborales previas consolidadas antes de terminar la formación formal.

El *cluster 3*, el más numeroso de todo el análisis, reunió a mujeres jóvenes del conurbano bonaerense con títulos de grado en áreas administrativas. Su salario promedio se mantuvo en rangos medios, y su inserción laboral fue del 62,5%, reflejando trayectorias tempranas y mayor presencia en empleos de ingreso inicial.

Finalmente, el *cluster 4* agrupó mayormente a mujeres de entre 26 y 27 años con formación técnica o preuniversitaria en áreas paramédicas y auxiliares, principalmente en la región pampeana. Si bien mostraron una de las inserciones más altas del mercado laboral, especialmente por la demanda sostenida de profesionales de salud, sus salarios promedio se mantuvieron entre los más bajos del conjunto, lo que reproduce la histórica desvalorización económica del trabajo de cuidados.

En conjunto, los cinco *clusters* revelaron patrones coherentes: la disciplina y la región continuaron siendo los principales factores estructurantes de la inserción laboral, mientras que la edad y el género modificaban la brecha salarial.

3.6. Visualización de los grupos mediante PCA y análisis multivariado

El PCA permitió reducir la dimensionalidad y representar los *clusters* en un plano de dos componentes que explicaban la mayor parte de la varianza total. Los perfiles vinculados a áreas económicas jóvenes aparecieron cercanos entre sí, mientras que los perfiles de salud y las mujeres de egreso tardío se distanciaron, reflejando diferencias reales en los rasgos académicos y demográficos.

El gráfico de burbujas complementó la interpretación al mostrar la relación entre salario promedio y edad promedio por cluster. Allí se observó que, aunque existe una tendencia generalizada al aumento del salario con la edad, este crecimiento depende fuertemente del sector laboral. Las burbujas más grandes correspondieron a los grupos jóvenes, mientras que las más pequeñas representaron trayectorias tardías con salarios más altos pero mayor heterogeneidad interna.

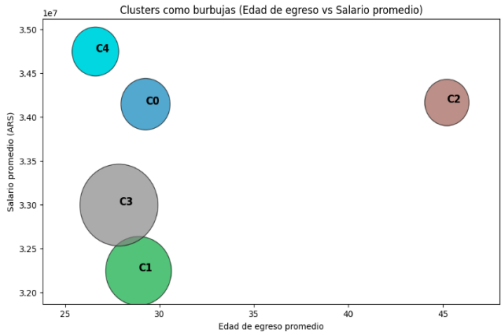

Ilustración 4: Representación en burbujas que muestra cómo varían los salarios medios y la edad de egreso entre los cinco clusters identificados.

3.7. Síntesis general de los resultados

Los resultados obtenidos permitieron reconstruir una mirada integral sobre las trayectorias sociolaborales de los graduados universitarios en Argentina. El EDA reveló desigualdades marcadas; los modelos supervisados mostraron que esas desigualdades tienen capacidad predictiva; y los modelos no supervisados permitieron agrupar a los graduados en perfiles coherentes con las estructuras del mercado laboral. En conjunto, los hallazgos muestran que la disciplina y la región son los ejes que más organizan las oportunidades laborales, mientras que el género, el tipo de institución y la edad de egreso modulan las posibilidades de ascenso salarial. La combinación de enfoques reveló que, si bien no es posible predecir un salario individual con precisión, sí es posible identificar patrones que atraviesan el sistema universitario y sus vínculos con el empleo.

4. Discusión

Los resultados obtenidos presentan patrones que coinciden en gran medida con hallazgos previos sobre el mercado laboral universitario argentino. Informes de organismos como CIPPEC, OEDE y la Secretaría de Políticas Universitarias ya habían señalado que la disciplina y la región son factores determinantes en la inserción laboral, mientras que el género y el tipo de institución funcionan como moduladores secundarios. Los datos de este trabajo reafirman ese comportamiento, mostrando brechas consistentes por área de estudio (especialmente entre carreras económicas y disciplinas vinculadas a la docencia o las humanidades) y marcadas desigualdades territoriales entre el AMBA y el interior del país.

Un resultado esperable fue la baja capacidad predictiva de los modelos de regresión para estimar salarios individuales. La dispersión del salario, observada en el EDA, anticipaba esta dificultad: el ingreso depende de variables no contempladas en el *dataset* (experiencia, sector económico, antigüedad, jornada laboral), por lo que los modelos no pudieron captar la complejidad del fenómeno. En cambio, resultados que podrían considerarse inesperados surgieron al analizar la inserción laboral por *cluster*. El grupo con mayor tasa de empleo no fueron los varones jóvenes de CABA (el perfil tradicionalmente privilegiado según estudios previos) sino las mujeres mayores del *cluster* 2, cuya elevada inserción probablemente se explique por trayectorias laborales ya consolidadas antes de finalizar sus estudios.

Por su parte los modelos supervisados, si bien no alcanzaron niveles altos de precisión, mostraron que las desigualdades estructurales tienen capacidad predictiva: factores como disciplina y región permiten anticipar con cierto grado de estabilidad quiénes acceden más fácilmente al empleo.

En conjunto, la discusión de los resultados indica que el análisis de datos masivos puede aportar evidencia consistente para comprender la estructura sociolaboral del país, pero también revela los límites de los modelos cuando las variables disponibles no reflejan la totalidad de los factores sociales que intervienen en la inserción laboral. Una recomendación para futuros trabajos sería incorporar información de experiencia laboral, sector productivo, salario horario y tipo de contratación, lo que permitiría construir modelos más robustos y una interpretación más completa del fenómeno salarial.

5. Conclusiones

El análisis realizado permitió construir una mirada amplia y fundamentada sobre las trayectorias laborales de los graduados universitarios en Argentina, combinando exploración estadística y modelado con técnicas supervisadas y no supervisadas. En primer lugar, el análisis exploratorio reveló patrones estructurales que ya habían sido señalados en estudios previos: fuertes desigualdades por disciplina, brechas salariales de género persistentes, diferencias regionales marcadas y un peso significativo del tipo de título. Estos resultados iniciales orientaron la selección de variables e hicieron evidente que la inserción laboral y el salario no se distribuyen de manera homogénea, sino que responden a estructuras sociales preexistentes.

En cuanto a los modelos supervisados, la predicción del salario resultó altamente compleja. Aun con algoritmos no lineales, los errores se mantuvieron elevados y la capacidad explicativa fue baja. Esto confirmó que el salario individual depende de factores ausentes en el *dataset* como experiencia previa, tipo de empleo, sector productivo, antigüedad o dedicación horaria y que no puede modelarse con precisión solo a partir de variables estructurales. Los modelos de clasificación, en cambio, mostraron patrones más estables: la Regresión Logística alcanzó desempeños cercanos al 69%, permitiendo identificar que disciplina y región son las variables más influyentes sobre la probabilidad de estar empleado.

El *clustering* aportó una perspectiva complementaria. Al agrupar a los graduados sin etiquetas previas, emergieron cinco perfiles laborales coherentes y diferenciados, organizados principalmente por disciplina, región, género y edad de egreso. Estos grupos permitieron sintetizar la heterogeneidad del sistema universitario argentino y mostraron que las trayectorias no se explican solo por la formación recibida, sino también por contextos territoriales, momentos del ciclo vital y dinámicas históricas del mercado laboral.

En términos de implicaciones, los resultados sugieren que la información disponible en bases institucionales permite

detectar patrones estructurales, pero es insuficiente para realizar predicciones individuales precisas. Esto tiene consecuencias tanto para la investigación como para la toma de decisiones: las herramientas de aprendizaje automático pueden ayudar a describir desigualdades, identificar perfiles y orientar políticas, pero no deben interpretarse como sistemas deterministas o predictivos en sentido estricto. Los modelos funcionan mejor como instrumentos de lectura del sistema que como mecanismos de predicción exacta.

Finalmente, este estudio deja abiertas varias líneas posibles para trabajos futuros. Incorporar variables vinculadas a la experiencia laboral, al tipo de contrato, al sector económico o a la dedicación horaria permitiría mejorar considerablemente la capacidad de predicción salarial. También sería útil profundizar el análisis territorial, explorando las diferencias al interior de cada región, y avanzar en modelos más complejos o técnicas de aprendizaje profundo siempre y cuando existan datos más ricos. En conjunto, los resultados muestran que el análisis de datos masivos es una herramienta valiosa para comprender las desigualdades laborales de los graduados, pero su potencial depende directamente de la calidad y diversidad de la información disponible.

- **Referencias**

[1] T. Amr, *Hands-On Machine Learning with Scikit-Learn: A Practical Guide to Implementing Supervised and Unsupervised Algorithms in Python*. Birmingham, UK: Packt Publishing, 2020.
[2] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. 2nd ed. Sebastopol, CA: O'Reilly Media, 2022.
[3] C. Huyen, *Designing Machine Learning Systems*. 1st ed. Sebastopol, CA: O'Reilly Media, 2022.
[4] A. Jung, *Machine Learning: The Basics*. Cambridge, UK: Cambridge University Press, 2022.
[5] CIPPEC, *La inserción laboral de los jóvenes en Argentina: brechas, desafíos y políticas*. Buenos Aires: CIPPEC, 2022.
[6] SPU – Secretaría de Políticas Universitarias, *Anuario de Estadísticas Universitarias*. Buenos Aires: Ministerio de Educación de la Nación, 2021.
[7] OEDE – Ministerio de Trabajo, Empleo y Seguridad Social, *Informes sobre empleo y dinámica laboral*. Buenos Aires: MTEySS, 2022.
[8] Scikit-Learn Developers (2023), *Scikit-Learn Documentation* [Online]. Available: https://scikit-learn.org/
[9] XGBoost Developers (2023), *XGBoost Documentation* [Online]. Available: https://xgboost.readthedocs.io/