Exploración y Predicción de Salarios: un Enfoque de Minería de Datos

Autores
Juan_Ardila, Sebastián_Morales
e-mail: jaardilap_1@uqvirtual.edu.co, smoralesd@uqvirtual.edu.co
Universidad del Quindío
Noviembre de 2023

Resumen

Este artículo aborda el análisis de un dataset centrado en la variable de salario mediante técnicas de minería de datos. Desde la exploración inicial hasta la implementación de estrategias predictivas y técnicas de clustering, se ha llevado a cabo un análisis exhaustivo. La visualización de datos reveló patrones y relaciones, seguida de una limpieza de datos integral. La predicción de salarios se exploró mediante técnicas de clasificación y balanceo de datos, y las características clave se identificaron con SelectKBest. Finalmente, el clustering proporcionó insights sobre la agrupación de datos. Este trabajo destaca la aplicabilidad y riqueza de la minería de datos en la comprensión de conjuntos de datos complejos.

Abstract

This paper delves into the analysis of a dataset focusing on the salary variable using data mining techniques. From initial exploration to the implementation of predictive strategies and clustering techniques, a comprehensive analysis has been conducted. Data visualization unveiled patterns and relationships, followed by thorough data cleaning. Salary prediction was explored through classification techniques and data balancing, and key features were identified using SelectKBest. Finally, clustering provided insights into data grouping. This work highlights the applicability and richness of data mining in understanding complex datasets.

Palabras clave:
Minería de Datos, Predicción de Salarios, Limpieza de Datos, Clasificación, Clustering.

Key words:
Data Mining, Salary Prediction, Data Cleaning, Classification, Clustering.

Contexto

Área de Conocimiento: Minería de Datos y Ciencia de Datos.
Temática: Análisis Predictivo y Exploratorio de Datos.
Institución: Universidad del Quindío.

1. Introducción

En el contexto de la asignatura de Descubrimiento del Conocimiento, hemos abordado un extenso análisis de un dataset llamado Adult que tiene como variable principal el salario. Este estudio abarca desde la exploración inicial de relaciones hasta técnicas avanzadas de predicción y clustering. Cada paso ha sido fundamental para extraer conocimientos significativos y revelar patrones ocultos en la complejidad de los datos. El uso de Python y sus librerías especializadas, como pandas y seaborn, desempeñó un papel crucial en todas las fases del análisis, permitiendo una manipulación eficiente de datos y una visualización clara de resultados.

2. Estado del arte

El estudio titulado "**A Project on Salary Prediction Using Regression Techniques**" [1], se enfocó en predecir salarios futuros mediante técnicas de regresión, basándose en una limpieza exhaustiva de datos históricos sobre crecimiento salarial. La recopilación del dataset se llevó a cabo desde múltiples fuentes, incluyendo datos de empresas y encuestas a empleados, para asegurar una representación amplia del panorama salarial. Se emplearon algoritmos de Regresión Lineal y Polinómica, modelando relaciones entre variables y generando gráficos predictivos que estiman salarios en diversas posiciones. Una validación meticulosa contrastó los resultados con índices de exactitud y confiabilidad, utilizando herramientas estadísticas para respaldar la solidez de las predicciones. Estos gráficos ofrecieron estimaciones salariales en distintos campos laborales, destacando cómo estas técnicas de regresión respaldadas por el aprendizaje automático ofrecen una base sólida para proyectar el crecimiento salarial en diversos contextos laborales.

El estudio presentado en el artículo "**Statistical Machine Learning Regression Models for Salary Prediction**" [2], ofrece una perspectiva completa de la predicción salarial en la economía saudita, empleando técnicas de regresión y aprendizaje automático. Su enfoque se centra en un marco integral que utiliza características ocupacionales y organizacionales para predecir el salario medio en actividades económicas y grupos ocupacionales principales. Al emplear cinco algoritmos de aprendizaje automático supervisado, se lograron mejoras significativas en la precisión de las predicciones. El modelo de regresión de proceso gaussiano bayesiano y las redes neuronales artificiales destacaron en la predicción del salario medio sobre actividades económicas y grupos ocupacionales respectivamente, mostrando mejoras sustanciales en los coeficientes de determinación y reduciendo significativamente los errores. Este marco propuesto permite estimar niveles salariales anuales en diferentes contextos laborales con datos limitados, considerando tanto características organizacionales como ocupacionales. Además, destaca cómo estas técnicas respaldadas por el aprendizaje automático ofrecen una sólida base para proyectar el crecimiento salarial en diversos campos laborales, presentando una metodología novedosa adaptable a diferentes mercados laborales internacionales.

El informe "**Salary Prediction in Data Science Field Using Specialized Skills and Job Benefits - A Literature Review**" [3], resalta la necesidad de una aproximación multifacética para predecir salarios en ciencia de datos. Autores como Chen, Sun, Thakuriah, Das, Barik, Mukherjee, y Dutta, Halder, Dasgupta evidencian enfoques diversos en la recolección y limpieza de datos, empleando conjuntos variados y técnicas específicas de preprocesamiento. Utilizaron métodos como CNN, Bidirectional-GRU-CNN, KNN y modelos de regresión, validados con precisión, MSE y R-cuadrado, permitiendo evaluar su eficacia. Aunque no se realizó una validación explícita, las conclusiones se basan en la comparación de resultados, resaltando la importancia de considerar habilidades especializadas y beneficios laborales para mejorar la precisión en las predicciones salariales en ciencia de datos.

En el informe analizado, contenido en el artículo"**An Ensemble Machine Learning Approach for Classifying Job Positions"** [4], se enfocó en la clasificación de diversas posiciones laborales mediante técnicas de aprendizaje automático, utilizando datos obtenidos de Glassdoor a través de web scraping, resultando en un conjunto de 955 instancias. Siguiendo el proceso CRISP-DM, se realizaron acciones de preparación de datos, como la eliminación de delimitadores inesperados, manejo de valores faltantes y transformación de atributos continuos. Se emplearon algoritmos como Adaboost, Random Forest, Gradient Boosting, XG Boost y Extra Trees Classifiers, evaluándolos con métricas como precisión, recall, F1-score, MSE, RMSE y el coeficiente de correlación de Matthews. Los ensembles heterogéneos mostraron un rendimiento superior, alcanzando una precisión cercana al 100% mediante votación suave, aunque se reconoció la limitación del tamaño del conjunto de datos y se propuso el uso de datos más amplios para futuras investigaciones.

3. Metodología

La metodología utilizada se dividió en varias etapas para abordar de manera integral el análisis del dataset centrado en la variable de salario.

3.1. Exploración de Datos:

Análisis descriptivo de las columnas del dataset. Visualización de relaciones mediante gráficos como cajas y bigotes, diagramas de barras e histogramas.

3.2. Limpieza y Preprocesamiento:

Eliminación de columnas no relevantes. Agrupación de categorías para mejorar la representación. Tratamiento de datos nulos y dispersos.

3.3. Predicción de Salarios:

Partición del dataset en conjuntos de entrenamiento y prueba (70/30) a manera de evitar un sobreajuste en las predicciones. Aplicación de técnicas de clasificación, como Decision Trees, Random Forest, Logistic Regression, y SVM. Balanceo de datos para abordar desafíos de clases desequilibradas. Evaluación de modelos mediante matrices de confusión y métricas como la exactitud.

3.4. Selección de Características:

Utilización de SelectKBest para identificar las características más relevantes.

3.5. Clustering:

Aplicación de KMeans para agrupar datos. Evaluación de la cantidad óptima de clusters mediante métricas como la inercia.

3.6. Visualización de Resultados:

Representación gráfica de centroides obtenidos del clustering. Análisis de las diferencias entre clusters y sus características distintivas.

Esta metodología integral permitió abordar cada fase del análisis de datos de manera efectiva, desde la exploración

inicial hasta la obtención de insights a través de técnicas predictivas y de agrupación.

4. Exploración de datos

4.1. Dataset

A continuación se muestra una tabla para resumir las características del dataset Adult utilizado antes de la aplicación de las técnicas mencionadas. Contando con la variable salario, que es la clase del dataset o dato que se busca predecir, y demás variables tanto categóricas como numéricas. Este conjunto de datos se ha utilizado ampliamente para la práctica y evaluación de algoritmos de clasificación en problemas de ingresos. La tarea típica es predecir si un individuo tiene un ingreso superior a $50,000 o no, convirtiéndolo en un problema de clasificación binaria.

I. Tabla Diccionario de Datos

DICCIONARIO DE DATOS		
Variable	**Tipo**	**Media o moda**
edad	Entero	38.245855
tipo_empleo	Polinomial	Private
codigo	Entero	189778.366512
nivel_educativo	Polinomial	High School
experiencia	Entero	10.080679
estado_civil	Polinomial	Married-civ-spouse
ocupacion	Polinomial	Adm-clerical
raza	Polinomial	White
genero	Binomial	Male
ganancias	Entero	0
perdidas	Entero	0
horas_semanales	Entero	40.437456
pais	Polinomial	United-States
salario	Binomial	<=50K

En las siguientes dos figuras se han graficado dos box-plot, para observar un comportamiento de los datos, en este caso un posible relación entre la variable edad y la clase salario. En resumen, las conclusiones obtenidas sugirieron que las edades varían entre las dos categorías de salario, con una tendencia a mayores edades en el grupo con salarios más altos.

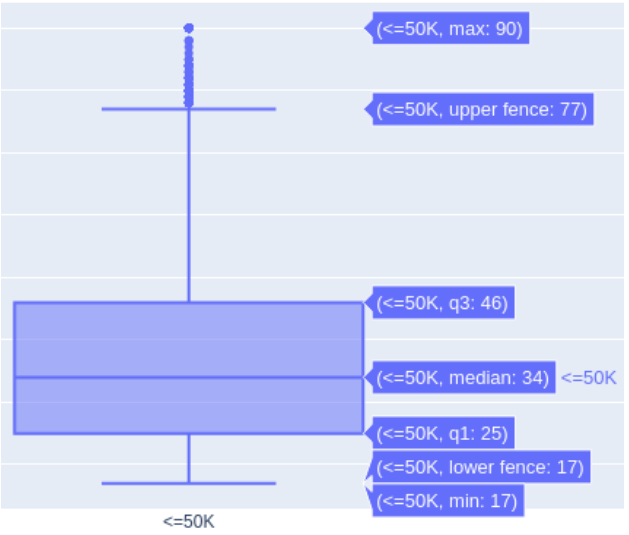

Fig. 1 Box-plot edad y salario menor que 50k

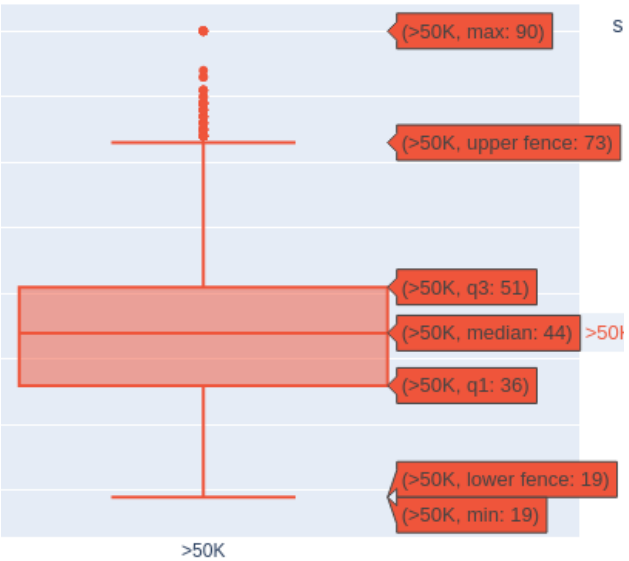

Fig. 2 Box-plot edad y salario mayor que 50k

Identificación de datos nulos. En la siguiente imágen se pudo observar que sólo tres columnas contienen datos faltantes: tipo_empleo, ocupacion y pais, dónde cada una tiene un porcentaje de datos nulos 5.6%, 5.6% y 1.8% respectivamente.

```
Valores nulos por columna

edad                   0
tipo_empleo         1836
codigo                 0
nivel_educativo        0
experiencia            0
estado_civil           0
ocupacion           1843
relacion_familiar      0
raza                   0
genero                 0
ganancias              0
perdidas               0
horas_semanales        0
pais                 583
salario                0
```

Fig. 3 Identificación de datos nulos

Se identificó la necesidad de simplificar la variable "nivel_educativo" debido a la presencia inicial de diversas categorías como se puede observar en el siguiente gráfico. Este proceso fue crucial para mejorar la comprensión del conjunto de datos y facilitar análisis posteriores. Las categorías se organizaron de manera minuciosa en grupos, reflejando la progresión natural de la educación, desde la educación básica hasta niveles universitarios avanzados. Esta categorización permitió un análisis más coherente y significativo del nivel educativo.

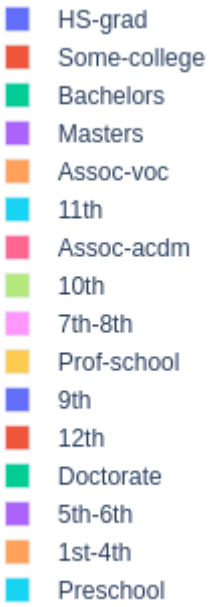

Fig. 4 Diversas categorías de la variable nivel_educativo

Se implementó una tabla de contingencia para examinar la relación entre las variables categóricas "relacion_familiar" y "estado_civil". Los resultados de este análisis revelaron similitudes significativas entre ambas variables, sugiriendo la posibilidad de que una pudiera reemplazar a la otra de manera efectiva. Dicha relación se observa en la siguiente tabla.

II. Tabla de Contingencia Estado Civil vs Relación Familiar

Se enfocó en analizar la distribución de salarios según el género en el conjunto de datos. Se observó una distinción notable entre los individuos masculinos (Male) que ganan más de $50K y aquellos que ganan menos. Específicamente, se registraron 6662 hombres con salarios superiores a $50K, mientras que la cantidad de mujeres que ganan menos de $50K fue de 9592. Esta disparidad sugiere que el género puede desempeñar un papel crucial en la distribución de ingresos entre la población estudiada.

Además, se presenta un gráfico que respalda visualmente esta observación, proporcionando una representación gráfica más intuitiva y detallada de las diferencias en la distribución salarial entre los géneros.

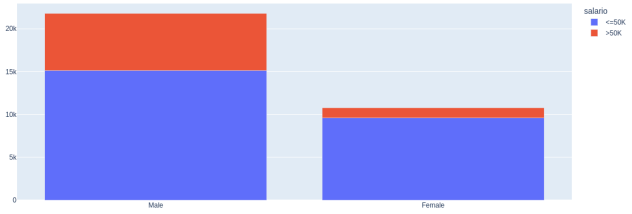

Fig. 5 Distribución salarial entre géneros

5. Limpieza y procesamiento

El equipo se centró en mejorar la calidad de las variables en el conjunto de datos. Para el problema de los datos nulos presentados se calcularon los promedios de las variables para dos grupos según el salario y, posteriormente, se corrigieron los valores atípicos. Estas acciones aseguran una representación precisa y coherente de las variables, aspecto esencial para análisis realizados posteriormente.

En cuanto a la variable “nivel__educativo”, se ejecutó la decisión de agrupar las categorías, implementando la simplificación planeada en la exploración. Este paso fue parte de una estrategia más amplia para limpiar y preparar

los datos para análisis avanzados. La reorganización de las categorías, como se observa en las figuras 5 y 6 se llevó a cabo con atención a la naturaleza progresiva de la educación, asegurando una representación más clara y efectiva de esta variable crucial en el conjunto de datos.

```
'Preschool': 'Basic Education',
'1st-4th': 'Basic Education',
'5th-6th': 'Basic Education',
'7th-8th': 'Basic Education',
'9th': 'Basic Education',
'10th': 'Basic Education',
'11th': 'Basic Education',
'12th': 'Basic Education',
'HS-grad': 'High School',
'Some-college': 'Post-Secondary Education',
'Assoc-acdm': 'Post-Secondary Education',
'Assoc-voc': 'Post-Secondary Education',
'Bachelors': 'Advanced Education',
'Masters': 'Advanced Education',
'Doctorate': 'Advanced Education',
'Prof-school': 'Advanced Education'
```

Fig. 6 Fragmento de código reagrupación de las categorías

Fig. 7 Listado de las categorías reagrupadas

En vista de la relación identificada anteriormente en la tabla II, se tomó la decisión de eliminar la variable "relacion_familiar". Es relevante mencionar que este proceso de selección se realizó estratégicamente considerando que la variable seleccionada ya contenía dos valores, "Husband" y "Wife", los cuales también estaban representados en otra columna del conjunto de datos. Estos valores específicos pueden ser fácilmente derivados de la variable "genero". Esta estrategia de limpieza no solo simplifica la estructura del conjunto de datos, sino que también preserva la información esencial, facilitando análisis posteriores.

Para la aplicación de métodos de clasificación y balanceo fue pertinente la codificación de las variables categóricas del dataset. Como se muestra a continuación.

```
df['tipo_empleo'] = label_encoding.fit_transform(df['tipo_empleo'].astype(str))
df['nivel_educativo'] = label_encoding.fit_transform(df['nivel_educativo'].astype(str))
df['estado_civil'] = label_encoding.fit_transform(df['estado_civil'].astype(str))
df['ocupacion'] = label_encoding.fit_transform(df['ocupacion'].astype(str))
df['raza'] = label_encoding.fit_transform(df['raza'].astype(str))
df['genero'] = label_encoding.fit_transform(df['genero'].astype(str))
df['pais'] = label_encoding.fit_transform(df['pais'].astype(str))
df['salario'] = label_encoding.fit_transform(df['salario'].astype(str))
```

Fig. 8 Fragmento de código codificación de variables categóricas.

En el siguiente gráfico de barras se nota de manera visual el desbalance presentado en la clase 'salario', dónde el valor 0 fue asignado para “<50” y 1 para “>50” gracias a la codificación.

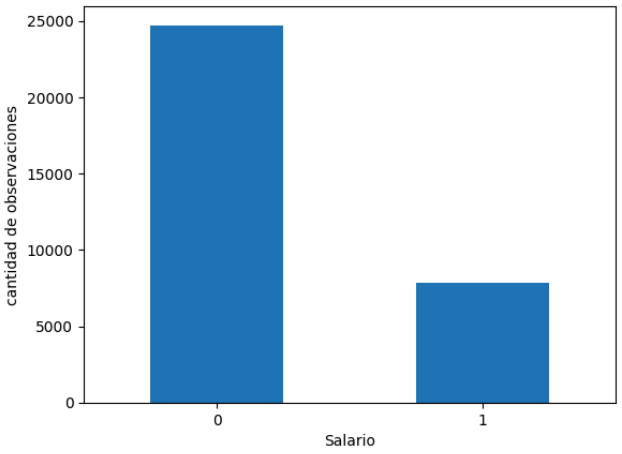

Fig. 9 Clase desbalanceada

El conjunto de datos se dividió en conjuntos de entrenamiento y prueba en una proporción de 70/30. Esta estrategia permite entrenar el modelo con el 70% de los datos y evaluar su rendimiento en el 30% restante. La división garantiza una representación adecuada de las clases en ambos conjuntos, evitando sesgos en el rendimiento del modelo. A continuación, se presenta un fragmento de código que ilustra cómo se realizó esta partición:

```
[ ]  # se toman todas las columnas menos salario, que es la clase
     X = df.drop('salario', axis=1)
     # se selecciona la columna de la clase
     y = df['salario']
     # Split dataset into training set and test set
     X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.3)
```

Fig. 10 Particionado del dataset

Durante la fase de desarrollo y entrenamiento del modelo de clasificación, se llevaron a cabo experimentos para abordar el desequilibrio de clases en el conjunto de datos. Se exploraron diversos enfoques de balanceo para mejorar la capacidad del modelo de clasificar las categorías "<=50K" y ">50K". Los métodos considerados incluyeron:

Penalización para Compensar: Se aplicaron técnicas de penalización para otorgar mayor peso a la clase minoritaria. Aunque mejoró la exactitud de esta clase, no se obtuvieron resultados globalmente satisfactorios.

Subsampling y Oversampling: Se experimentó con subsampling (muestreo aleatorio de la clase mayoritaria) y oversampling (duplicación de ejemplos de la clase minoritaria), sin alcanzar los resultados deseados.

Resampling con SMOTE-Tomek: La técnica SMOTE-Tomek, que combina SMOTE con la eliminación de ejemplos Tomek, se aplicó para equilibrar las clases y eliminar ejemplos redundantes. Aunque mejoró la clasificación de la clase minoritaria, persistieron limitaciones en las métricas globales.

Ensamble de Modelos con Balanceo: Se adoptó una estrategia de ensamble de modelos, específicamente el "Balanced Random Forest Classifier." Este modelo

aborda automáticamente el desequilibrio de clases y demostró ser la elección más prometedora.

Resultados del Modelo Final con Balanced Random Forest: Tras implementar el "Balanced Random Forest Classifier", el modelo alcanzó un rendimiento considerado satisfactorio. Las métricas de evaluación en el conjunto de prueba fueron:

Exactitud (Accuracy): 0.81

Puntuación F1 promedio (Macro AVG): 0.78

Estos resultados reflejan un equilibrio exitoso entre la exactitud y la recuperación en la clasificación de ambas clases, "<=50K" y ">50K," con una exactitud general del 81%. La puntuación F1 promedio de 0.78 indica una ponderación razonable entre exactitud y recuperación. A continuación se puede apreciar la matriz de confusión obtenida con el modelo seleccionado:

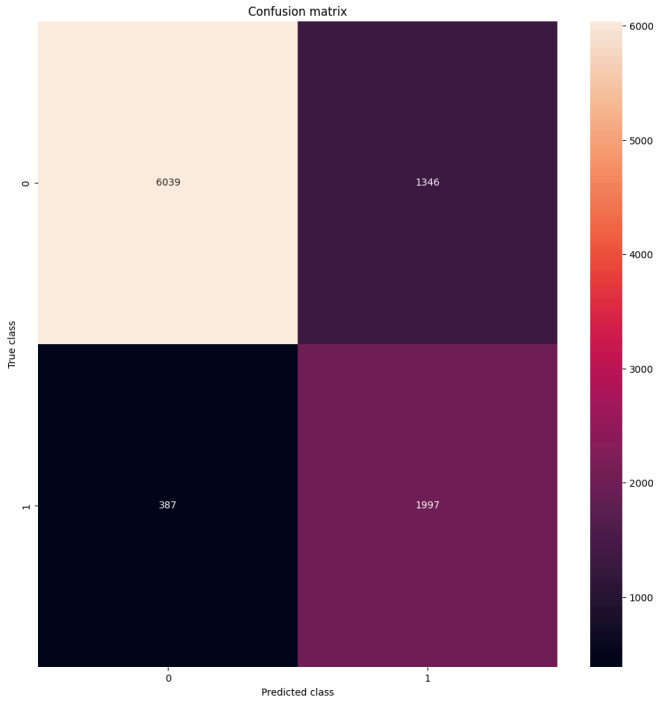

Fig. 11 Matriz de confusión para el modelo Balanced Random Forest

Se aplicaron diversas técnicas de clasificación, incluyendo Decision Tree, Random Forest, Logistic Regression y SVM. Los resultados y métricas obtenidos de estos modelos llevaron a las siguientes conclusiones:

Random Forest: Encabezando el ranking con el puntaje F1 más alto (0.7939), el modelo Random Forest exhibe un equilibrio óptimo entre exactitud y recall. Con una exactitud del 86%, demuestra efectividad en predecir salarios superiores o inferiores a $50,000.

Decision Tree: Aunque eficaz con un puntaje de F1 de 0.7878, el modelo de árbol de decisión se sitúa en segundo lugar. Aunque ligeramente menos preciso que Random Forest, aún muestra habilidad para predecir con alta exactitud las clases de salario.

Logistic Regression: Con un F1 Score de 0.7810, la regresión logística ocupa el tercer lugar. Aunque muestra una exactitud decente, se observa un bajo rendimiento en recall, indicando posibles dificultades para identificar casos de salario alto.

SVM (Support Vector Machine): Registrando el F1 Score más bajo (0.7429), el modelo SVM exhibe un rendimiento más bajo en comparación con los demás modelos, sugiriendo que podría no ser la mejor elección en este contexto.

Además, se abordaron problemas de overfitting y underfitting mediante la selección adecuada de hiperparámetros y técnicas como la validación cruzada, garantizando la capacidad de generalización a datos no vistos.

Luego se obtuvieron las 5 mejores características por medio del método de SelectKBest para proceder a aplicar clustering. Listadas así:

1. edad
2. nivel_educativo
3. experiencia
4. ganancias
5. horas_semanales

El siguiente gráfico utiliza la técnica del codo para proporcionar una aproximación al valor de k que podría seleccionarse para obtener los centroides de los clusters. La representación visual revela que el codo resultante se encuentra igualado o muy cercano al valor de 2 en el eje x. Este punto sugiere un posible punto de inflexión donde aumentar el número de clusters podría no proporcionar beneficios significativos en la reducción de la varianza intracluster. La técnica del codo, al identificar este punto, brinda una valiosa perspectiva sobre la elección adecuada de k para optimizar la formación de clusters en el conjunto de datos analizado.

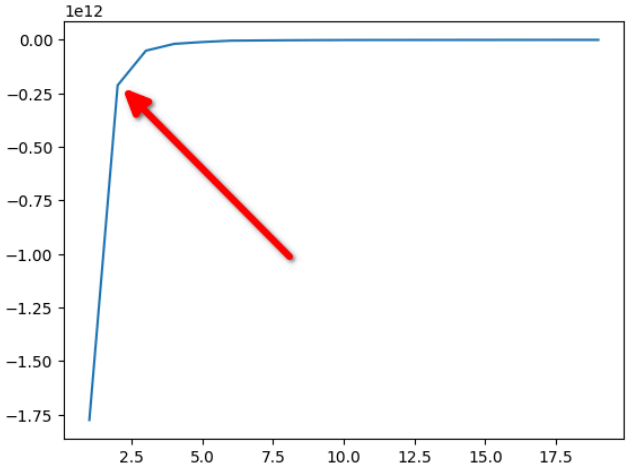

Fig. 12 Técnica del codo

Se llevó a cabo un ajuste de hiperparámetros para identificar el número óptimo de clusters en el conjunto de datos. El análisis se realizó considerando un rango desde

2 hasta 9 clusters. La evaluación se basó en la varianza intracluster y su relación con el número de clusters. El resultado reveló que el número óptimo de clusters es 2, lo que indica un equilibrio adecuado entre la cohesión interna de los clusters y la minimización de la dispersión intracluster. Este parámetro optimizado se utilizó posteriormente en el algoritmo KMeans para obtener centroides significativos y representativos de la estructura subyacente en los datos.

```
parametro:  {'n_clusters': 2} puntaje:  0.9877943161158996
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default va
  warnings.warn(
parametro:  {'n_clusters': 3} puntaje:  0.951709488933217
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default va
  warnings.warn(
parametro:  {'n_clusters': 4} puntaje:  0.9574307420537522
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default va
  warnings.warn(
parametro:  {'n_clusters': 5} puntaje:  0.9596618375148007
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default va
  warnings.warn(
parametro:  {'n_clusters': 6} puntaje:  0.9690933729522463
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default va
  warnings.warn(
parametro:  {'n_clusters': 7} puntaje:  0.9699549667838677
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default va
  warnings.warn(
parametro:  {'n_clusters': 8} puntaje:  0.9687117515821262
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default va
  warnings.warn(
parametro:  {'n_clusters': 9} puntaje:  0.9704016391064616
{'n_clusters': 2}
```

Fig. 13 Mejor parámetro de clusters

Se aplicó el algoritmo KMeans con un número de clusters igual a 2, considerando la información obtenida del ajuste de hiperparámetros. A continuación, se presentan los puntos que representan los centroides resultantes. Estos centroides juegan un papel crucial en la definición y caracterización de los grupos identificados en el conjunto de datos, proporcionando información valiosa sobre las tendencias y relaciones presentes en las variables seleccionadas.

```
Centroides
[[3.82080048e+01 1.67822974e+00 1.00667551e+01 5.92231436e+02
  4.03915190e+01]
 [4.59591175e+01 6.10062893e-01 1.29182390e+01 9.99990000e+04
  4.97987421e+01]]
```

Fig. 14 Puntos de los centroides obtenidos

El análisis detallado de los centroides obtenidos del clustering revela diferencias significativas entre los dos clusters identificados:

Cluster 1 (Centroide 0 - Azul): Este grupo está caracterizado por individuos más jóvenes, con una edad promedio de 38.21 años. Aunque poseen menos experiencia y un nivel educativo moderado, muestran equilibrio en las ganancias y horas semanales de trabajo, con ingresos promedio de $592.23 y una carga laboral promedio de 40.39 horas semanales.

Cluster 2 (Centroide 1 - Naranja): En contraste, el segundo cluster está compuesto por individuos mayores, con una edad promedio de 45.96 años. A pesar de tener una mayor experiencia, presentan un nivel educativo más bajo en comparación con el primer cluster. Sorprendentemente, este grupo muestra ganancias significativamente más altas, con un ingreso promedio de $99,999.00, y dedican más tiempo al trabajo, con un promedio de 49.80 horas semanales.

En resumen, el primer cluster refleja a individuos más jóvenes con equilibrio en sus ingresos y horas de trabajo, mientras que el segundo cluster representa a individuos mayores con mayores ingresos, probablemente debido a una mayor experiencia y dedicación laboral como es posible visualizar en el gráfico a continuación. Estas diferencias proporcionan percepciones valiosas para comprender la diversidad en la población estudiada.

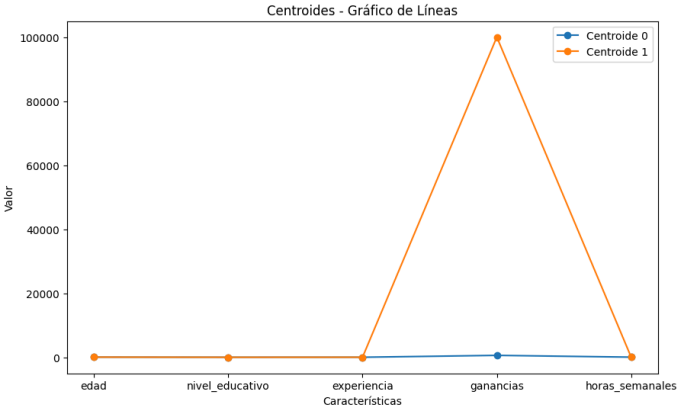

Fig. 15 Gráfico de centroides

6. Resultados de clasificación

III. Tabla Resultados Modelos de Clasificación

Modelo	F1 Score	Exactitud
Random Forest	0.7939	86%
Decision Tree	0.7878	81%
Logistic Regression	0.7810	80%
SVM	0.7429	79%

7. Conclusiones

Con base en el extenso análisis de datos realizado en este proyecto de minería de datos sobre el conjunto de datos de empleados y salarios, se pueden derivar varias conclusiones significativas:

- La exploración inicial reveló patrones y relaciones clave entre las variables, proporcionando una visión profunda del conjunto de datos.
- La limpieza de datos, que incluyó la eliminación de variables redundantes y la corrección de valores atípicos, contribuyó a la calidad general de los datos.
- La aplicación de SelectKBest permitió identificar las características más relevantes para predecir los salarios, destacando la importancia de variables como la edad, la educación y la experiencia.

- El proceso de balanceo, utilizando técnicas como Balanced Random Forest Classifier, demostró ser esencial para abordar el desequilibrio de clases.
- El modelo Random Forest destaca como la mejor opción para la clasificación de salarios en este conjunto de datos, ofreciendo el F1 Score más alto y una exactitud considerable. Superando a otros modelos como Decision Tree, Logistic Regression y SVM.
- La aplicación de KMeans reveló dos clusters distintos en función de características clave, como la edad, la educación y las ganancias. Dónde las ganancias se demostraron como un factor diferenciador sumamente importante.
- Los clusters proporcionaron percepciones valiosas sobre la diversidad en la población estudiada, diferenciando entre individuos más jóvenes con ingresos equilibrados y mayores con mayores ganancias.
- La técnica del codo logró acertar junto con el ajuste de hiperparámetros al proporcionar el valor óptimo para el número de clusters, fundamentando la elección de dos clusters para el análisis de centroides.
- La representación gráfica de datos, como el gráfico de líneas de centroides, mejoró la interpretación y comunicación de resultados complejos.

En conjunto, este proyecto destaca la importancia de un enfoque integral, desde la exploración inicial hasta la aplicación de técnicas avanzadas, para extraer conocimientos significativos de conjuntos de datos complejos. Los resultados obtenidos proporcionan una base sólida para comprender las relaciones y patrones en el ámbito laboral y salarial.

8. Referencias

[1] PULENDRA KUMAR YADAV, RIKESH KUMAR. "A Project on Salary Prediction Using Regression Techniques". SSRN. 2023. Disponible en: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3990877 (Accedido el 18 de noviembre de 2023).

[2] Matbouli, Yasser T., and Suliman M. Alghamdi. "Statistical Machine Learning Regression Models for Salary Prediction Featuring Economy Wide Activities and Occupations." Information 13, no. 10 (2022): 495. [En línea] https://www.mdpi.com/2078-2489/13/10/495

[3] Tee Zhen Quan y Mafas Raheem. "Salary Prediction in Data Science Field Using Specialized Skills and Job Benefits – A Literature Review." Journal of Applied Technology and Innovation, vol. 6, no. 3, 2022, pp. 70. Disponible en: https://www.researchgate.net/publication/362280362_Salary_Prediction_in_Data_Science_Field_Using_Specialized_Skills_and_Job_Benefits_-A_Literature_Review

[4] Ayaz Kh. Mohammed, Abdullahi Aliyu Danlami, Dindar I. Saeed, Abdulmalik Ahmad Lawan, Adamu Hussaini, Ramadhan Kh. Mohammed. "An Ensemble Machine Learning Approach for Classifying Job Positions." Academic Journal of Nawroz University (AJNU), vol. 12, no. 3, 2023, pp. 547. Disponible en: https://www.researchgate.net/publication/373530245_An_Ensemble_Machine_Learning_Approach_for_Classifying_Job_Positions