Futura, Data & Analítica Avanzada

Data Science For Business
Trabajo Práctico

CASO 5
Modelo de clasificación

Autores:
- Farfan Florián Erik David
- Lozada Perez María del Cielo

Fecha: 13 de noviembre de 2024

Indice

1. Introducción

La alta rotación de personal en nuestra empresa representa una pérdida significativa de conocimiento y un impacto negativo en nuestra productividad. Con el objetivo de revertir esta tendencia y fortalecer nuestra capacidad de retención del talento, llevaremos a cabo un análisis exhaustivo de los datos de HRAnalytics.csv. A través de técnicas de aprendizaje automático, identificaremos los factores clave que influyen en la decisión de los empleados de permanecer en la organización y construiremos un modelo predictivo para identificar a aquellos con mayor potencial de crecimiento. Al anticipar las necesidades de nuestros colaboradores y diseñar programas de desarrollo personalizados, podremos mejorar la satisfacción laboral, reducir la rotación y fomentar una cultura organizacional más sólida.

Para desarrollar este modelo, se emplearán técnicas de aprendizaje automático, como modelos bagging y boosting. Se realizará una exploración exhaustiva de los datos para identificar las variables más relevantes y se evaluará el rendimiento del modelo utilizando métricas como precisión y recall.

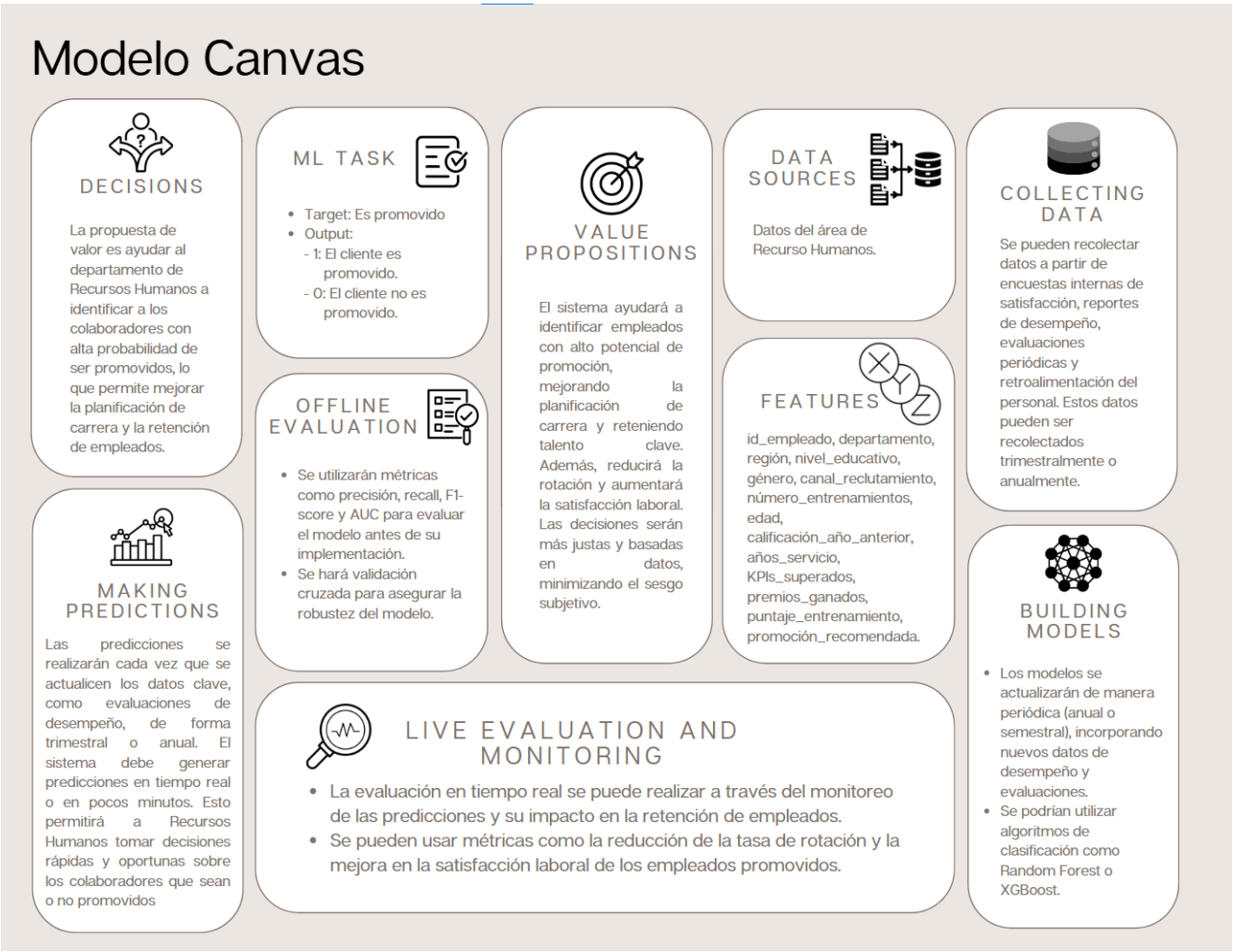
2. Planteamiento del problema

2.1. Problema y objetivo

El problema que enfrenta la empresa es la alta rotación de personal, que por ejemplo en el área de Analytics supera el 20 % anual, lo cual afecta la continuidad de los procesos, los proyectos, y la transferencia de conocimientos dentro de esa área. Los colaboradores que se han ido mencionan como causa principal la falta de oportunidades de promoción y desarrollo de carrera en la empresa. Objetivo del negocio: Identificar a los colaboradores con alta probabilidad de ser promovidos en la empresa. De esta forma, el área de Recursos Humanos podría intervenir de manera eficaz para mejorar la retención de personal clave y reducir la rotación.

Para desarrollar este modelo, se emplearán técnicas de aprendizaje automático, como modelos bagging y boosting. Se realizará una exploración exhaustiva de los datos para identificar las variables más relevantes y se evaluará el rendimiento del modelo utilizando métricas como precisión y recall.

3. Machine Learning Canvas



4. Tipo de variables

Las variables que se nos disponibiliza son:

- **employee id**: ID único del empleado.
- **department**: Departamento de empleado.
- **region**: Región de empleo (desordenada).
- **education**: Nivel de educación.
- **gender**: Género del empleado.
- **recruitment channel**: Canal de contratación de empleados.

- **no of trainings**: Número de entrenamientos completados el año anterior sobre habilidades blandas, habilidades técnicas, etc.
- **age**: Edad del empleado.
- **previous year_rating**: Calificación del empleado del año anterior.
- **length of service**: Duración del servicio en años.
- **KPIs_met >80 %**: Si el porcentaje de KPI (indicadores clave de rendimiento) es mayor a 80 %, entonces es 1, sino es 0.
- **awards won?**: Si los premios ganados durante el año anterior, entonces 1, sino 0.
- **avg training score**: Puntaje promedio en evaluaciones de entrenamiento actuales.
- **is promoted**: Recomendado para ser promovido.

Variables numéricas	Variables binarias	Variables categóricas
age avg training score	KPIs met >80 % awards won? is promoted	no of trainings previous year rating department region education gender recruitment channel

5. Supuesto de completitud de la información (missings)

La base de datos brindada no cumple con el supuesto de completitud de información, las variables previous year rating y education presentan el 7.5 % y 4.4 % de missings respectivamente.

	Total	Percent
previous_year_rating	4124	7.524449
education	2409	4.395344

previous year rating: Analizando esta variable, nos dimos cuenta que justo las observaciones que tienen missings coincidían con los colaboradores que tienen un año de experiencia.

```
datos[datos['previous_year_rating'].isnull()]['length_of_service'].value_counts()
```

```
length_of_service
1    4124
Name: count, dtype: int64
```

Esto tiene sentido, ya que la variable es la calificación del empleado un año anterior y como no se tiene información está como vacía. Por esa razón, imputarlos con un valor numérico no es la solución, ya que la variable como tal no debe tener valor. Por lo tanto, lo más adecuado sería crear una nueva categoría llamada sin rating.

previous_year_rating	
3.0	18618
5.0	11741
4.0	9877
1.0	6223
2.0	4225
sin_rating	4124

educación: Por el lado de la variable educación no se encontró ningún patrón del porqué de sus missing. Imputar al colaborador el nivel de ecuación más frecuente no tendría sentido, estar´ıamos llenándolo de información errónea. Una opción sería pedirles a recursos humanos que nos ayuden con esa información, y la otra, agregar una nueva categoría. Optamos por la segunda opción.

education	
Bachelor's	36669
Master's & above	14925
no_identificado	2409
Below Secondary	805

6. Resumen global del total de variables de estudio

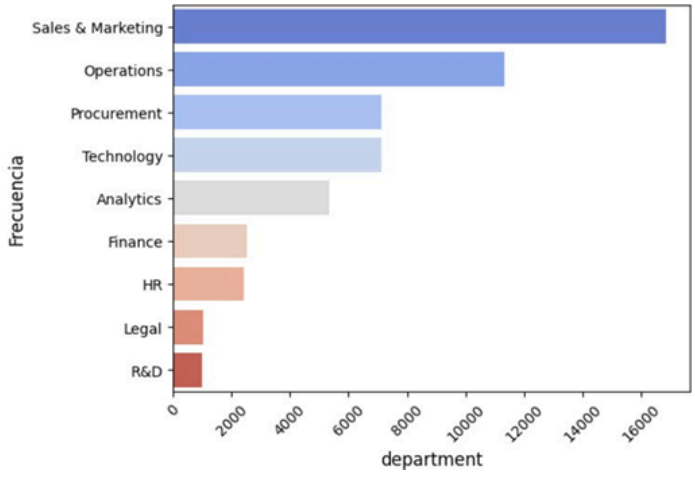
6.1 Nivel Univariado

a) Variables categóricas y binarias

	count	unique	top	freq
department	54808	9	Sales & Marketing	16840
region	54808	34	region_2	12343
education	54808	4	Bachelor's	36669
gender	54808	2	m	38496
recruitment_channel	54808	3	other	30446
no_of_trainings	54808	10	1	44378
previous_year_rating	54808.0	6.0	3.0	18618.0
KPIs_met >80%	54808	2	0	35517
awards_won?	54808	2	0	53538
is_promoted	54808	2	0	50140

- **department**: Hay 54,808 registros en total, distribuidos entre 9 departamentos únicos. El departamento más frecuente es Sales & Marketing, con 16,840 ocurrencias.

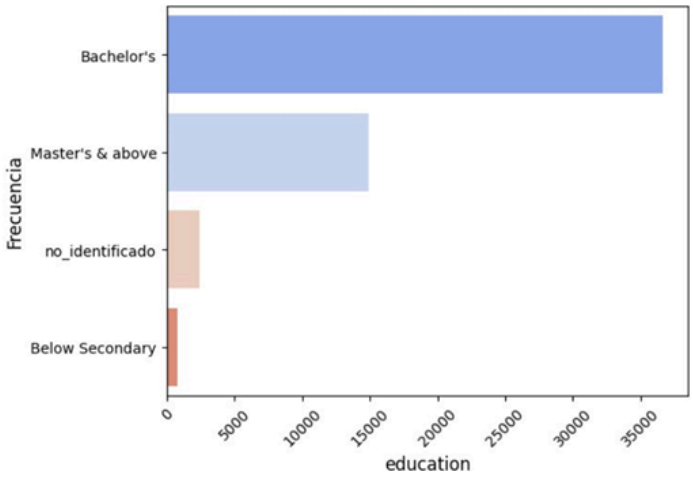
```
#-----------------------------------------------------------------#
Frecuencia y pocentaje de la variable department
#-----------------------------------------------------------------#
         department  count  %porcentaje
0  Sales & Marketing  16840    30.725442
1         Operations  11348    20.705007
2        Procurement   7138    13.023646
3         Technology   7138    13.023646
4          Analytics   5352     9.764998
5            Finance   2536     4.627062
6                 HR   2418     4.411765
7              Legal   1039     1.895709
8                R&D    999     1.822727
```



- **región**: Existen 34 regiones únicas en el conjunto de datos, con un total de 54,808 registros. La región más común es la región 2, que aparece 12,343 veces.

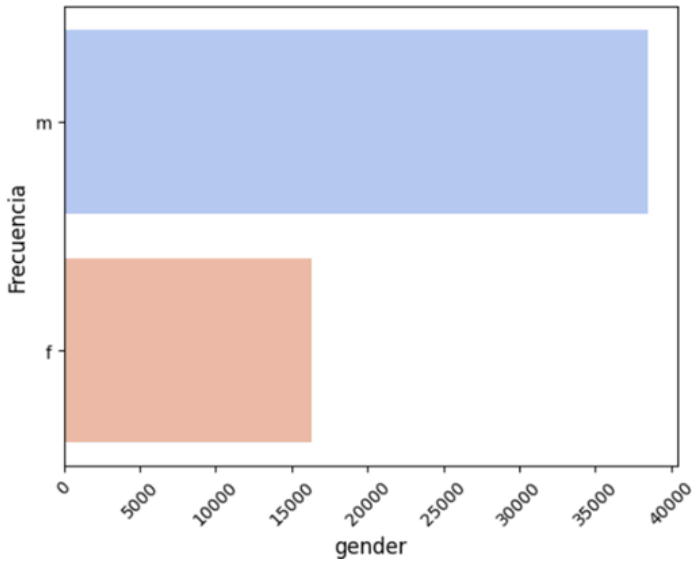
- **education**: Esta variable presenta información sobre el nivel educativo de 54,808 indivi- duos, con 4 niveles educativos distintos. El nivel más frecuente es Bachelor 's, que aparece 36,669 veces.

```
#-------------------------------------------------------------#
Frecuencia y pocentaje de la variable education
#-------------------------------------------------------------#
          education  count  %porcentaje
0         Bachelor's  36669    66.904467
1  Master's & above  14925    27.231426
2    no_identificado   2409     4.395344
3    Below Secondary    805     1.468764
```



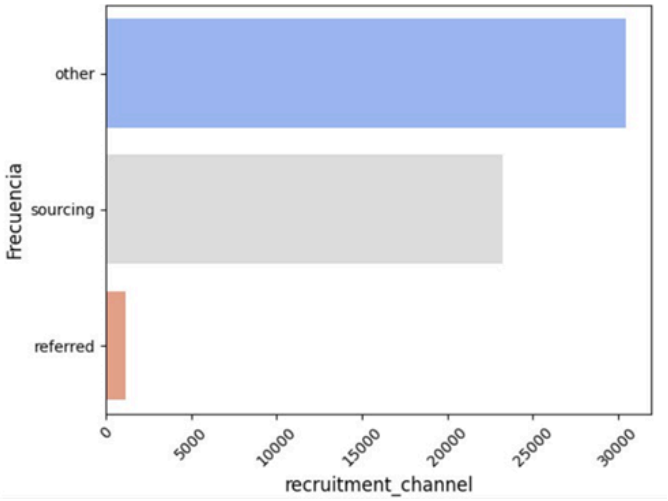
- **gender**: Se registran dos géneros en el conjunto de datos, con un total de 54,808 observaciones. El género más común es m (masculino), con 38,496 registros.

```
#-----------------------------------------------------#
Frecuencia y pocentaje de la variable gender
#-----------------------------------------------------#
  gender  count  %porcentaje
0      m  38496    70.237921
1      f  16312    29.762079
```



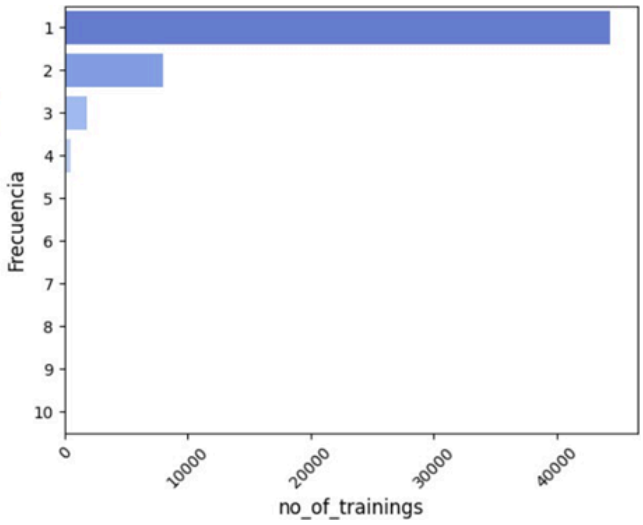
- **recruitment** channel: Hay 3 canales de reclutamiento identificados entre los 54,808 registros, siendo el canal más común, con 30,446 ocurrencias.

```
#-----------------------------------------------------------------#
Frecuencia y pocentaje de la variable recruitment_channel
#-----------------------------------------------------------------#
  recruitment_channel  count  %porcentaje
0               other  30446    55.550285
1           sourcing  23220    42.366078
2            referred   1142     2.083637
```



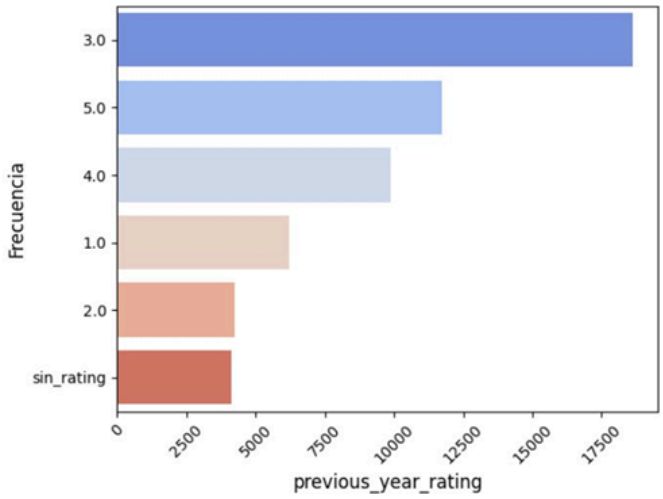
- **no of trainings**: Esta variable describe la cantidad de entrenamientos completados, con 10 valores únicos registrados. El valor más común es 1, con 44,378 observaciones.

```
#-----------------------------------------------------------------#
Frecuencia y pocentaje de la variable no_of_trainings
#-----------------------------------------------------------------#
  no_of_trainings  count  %porcentaje
0               1  44378    80.969931
1               2   7987    14.572690
2               3   1776     3.240403
3               4    468     0.853890
4               5    128     0.233543
5               6     44     0.080280
6               7     12     0.021895
7               8      5     0.009123
8               9      5     0.009123
9              10      5     0.009123
```



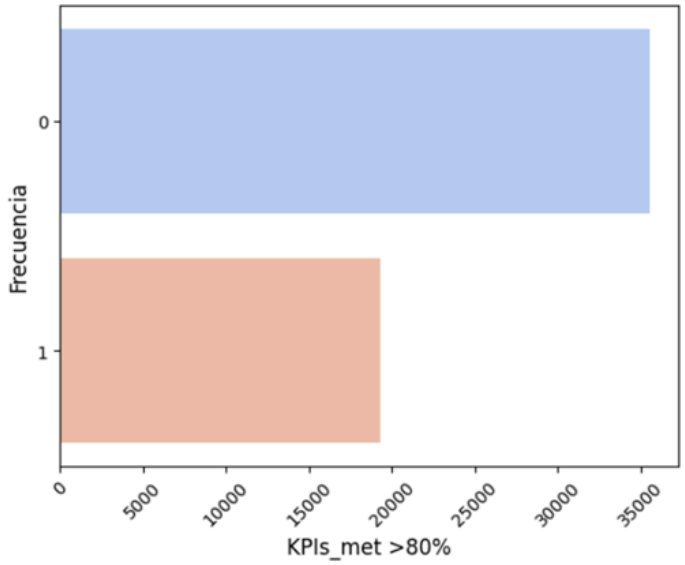
- **previous year rating**: La calificación del año anterior está presente para los 54,808 registros, con 6 calificaciones únicas. La calificación más común es 3.0, que aparece en 18,618 casos.

```
#-------------------------------------------------------------#
Frecuencia y pocentaje de la variable previous_year_rating
#-------------------------------------------------------------#
  previous_year_rating  count  %porcentaje
0                  3.0  18618    33.969494
1                  5.0  11741    21.422055
2                  4.0   9877    18.021092
3                  1.0   6223    11.354182
4                  2.0   4225     7.708729
5          sin_rating   4124     7.524449
```



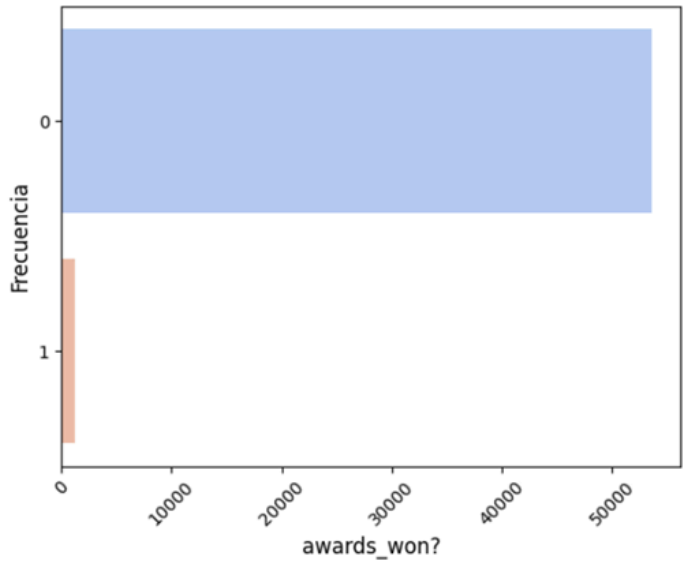
- **KPIs met ¿80 %**: Indica si los empleados alcanzaron más del 80 % de sus KPIs. Hay 2 valores posibles, la mayor´ıa de los empleados (35,517) no lo lograron.

```
#-------------------------------------------------------#
Frecuencia y pocentaje de la variable KPIs_met >80%
#-------------------------------------------------------#
   KPIs_met >80%  count  %porcentaje
0              0  35517    64.802584
1              1  19291    35.197416
```



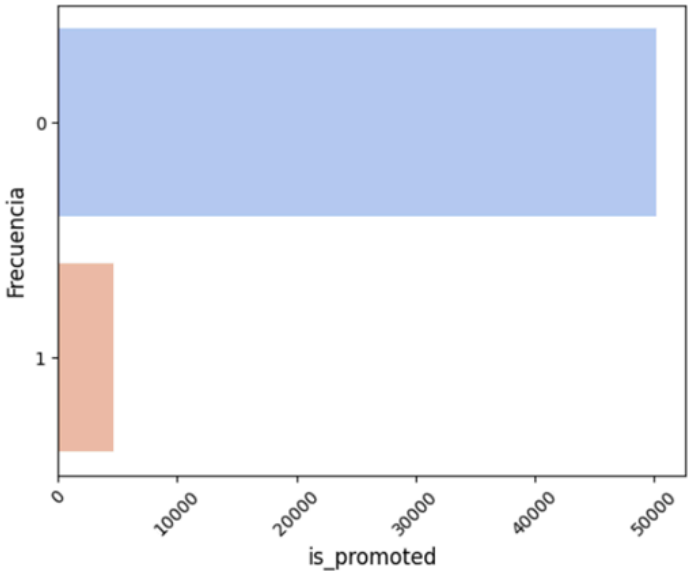
- **awards won?**: Señala si los trabajadores ganaron algún premio. Tiene 2 valores posibles. La mayor´ıa (53,538 trabajadores) no ganaron premios.

```
#---------------------------------------------------------#
Frecuencia y pocentaje de la variable awards_won?
#---------------------------------------------------------#
   awards_won?  count  %porcentaje
0            0  53538     97.68282
1            1   1270      2.31718
```



- **is promoted**: Indica si los trabajadores fueron promovidos. Hay 2 valores posibles, y la mayor´ıa (50,140 trabajadores) no fueron promovidos.

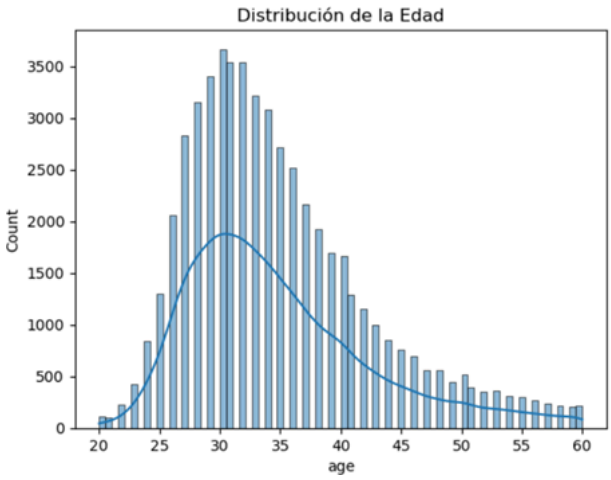
```
#---------------------------------------------------------#
Frecuencia y pocentaje de la variable is_promoted
#---------------------------------------------------------#
   is_promoted  count  %porcentaje
0            0  50140    91.482995
1            1   4668     8.517005
```

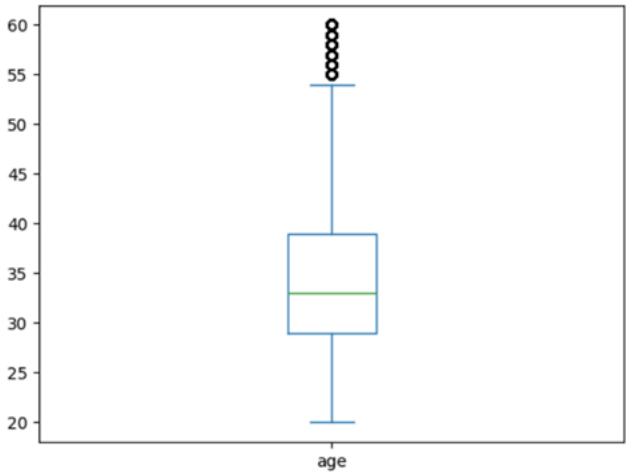


b) Variables numéricas

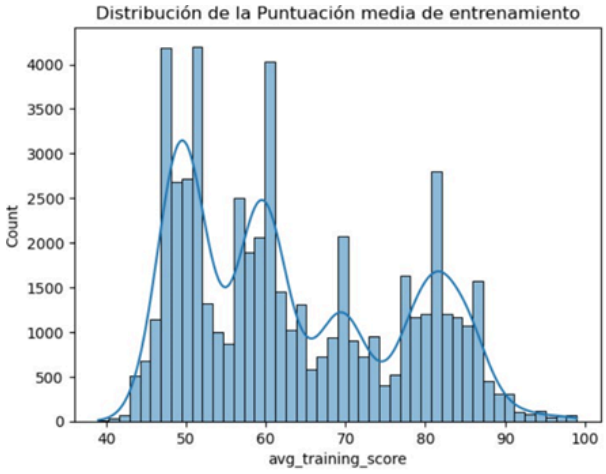
	count	mean	std	min	25%	50%	75%	max
age	54808.0	34.803915	7.660169	20.0	29.0	33.0	39.0	60.0
avg_training_score	54808.0	63.386750	13.371559	39.0	51.0	60.0	76.0	99.0

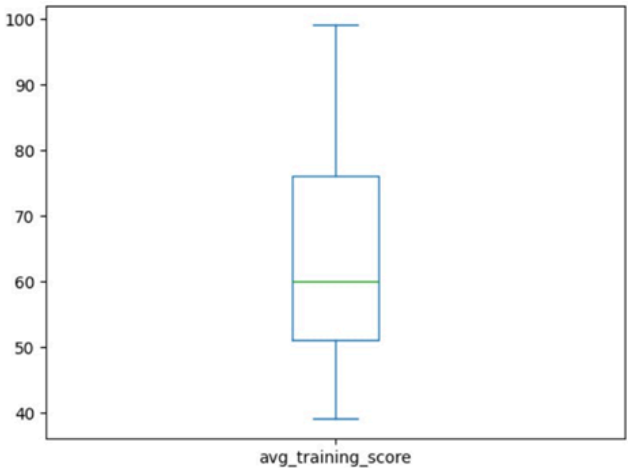
- Age:
  - La edad promedio es de 34.8 años, con una variabilidad de 7.66 años (desviación estándar), lo que indica que las edades de los empleados están bastante dispersas.
  - La edad mínima registrada es 20 años y la máxima es 60.
  - El 25 % de los empleados tiene 29 años o menos, la mediana (el 50 %) es de 33 años, y el 75 % tiene 39 años o menos. Esto sugiere que la mayor´ıa de los empleados tienen entre 29 y 39 años.
  - La distribución es unimodal, lo que significa que tiene un solo pico. El gráfico tiene un sesgo hacia la derecha (asimetría positiva), lo que indica que la mayor parte de los datos se concentran en edades más jóvenes, pero hay una cola que se extiende hacia edades mayores.





- **avg training score:**
  - El puntaje promedio es de 63.39, con una desviación estándar de 13.37, lo que indica una moderada dispersión en los puntajes.
  - El puntaje mínimo es 39 y el máximo es 99.
  - El 25 % de los empleados obtuvo un puntaje de 51 o menos, la mediana es de 60, y el 75 % de los empleados alcanzó un puntaje de 76 o menos. Esto sugiere que la mayor´ıa de los empleados tuvieron puntajes entre 51 y 76.
  - Esta distribución es multimodal, ya que tiene varios picos en distintas puntuaciones. Los grupos principales de promedios de puntuación parecen estar alrededor de los 50, 60, 70 y 80 puntos.





7. Análisis de asociación de variables independientes con la variable objetivo

Como se pudo observar la gran mayor´ıa de variables son categóricas o binarias y el target es binario, es por ello que el análisis de asociaci´on de ellas respecto al target se har´a mediante cruces de tal manera que se podr´a identificar si el target se discrimina en sus categor´ıas.
Por el lado de las numéricas igualmente se hará un análisis gráfico e identificamos cómo se distribuyen respecto al target.

- Variables cualitativas y binarias

```
#---------------------------------------#
Cruce 'is_promoted' con 'no_of_trainings':
#---------------------------------------#
is_promoted               0         1
no_of_trainings
1                 91.189328  8.810672
2                 92.425191  7.574809
3                 93.130631  6.869369
4                 94.444444  5.555556
5                 97.656250  2.343750
6                 95.454545  4.545455
7                100.000000  0.000000
8                100.000000  0.000000
9                100.000000  0.000000
10               100.000000  0.000000
All               91.482995  8.517005
```

Se puede observar que la variable no of trainings no discrimina bien al target, en todas las categorías la proporción de colaboradores que son promovidos es baja o incluso 0 en comparación a los que no lo son.

```
#---------------------------------------#
Cruce 'is_promoted' con 'KPIs_met >80%':
#---------------------------------------#
is_promoted            0          1
KPIs_met >80%
0              96.041332   3.958668
1              83.090560  16.909440
All            91.482995   8.517005
```

Por el lado de la variable KPIs met >80 % podemos observar una clara discriminaci´on del target en sus categorías, ya que cuando el indicador de rendimiento es menor a 80 % solo el 1 % es promovido, y en caso contrario el 16 % es promovido. Es una diferencia considerable.

```
#---------------------------------------#
Cruce 'is_promoted' con 'awards_won?':
#---------------------------------------#
is_promoted          0          1
awards_won?
0            92.325078   7.674922
1            55.984252  44.015748
All          91.482995   8.517005
```

La variable awards won, al igual que el anterior, podemos observar una clara discriminación del target en sus categor´ıas, ya que cuando el colaborador no gana premios solo el 7 % es promovido, y en caso contrario el 40 % es promovido lo que hace, en proporción, una diferencia considerable.

```
#---------------------------------------#
Cruce 'is_promoted' con 'department':
#---------------------------------------#
is_promoted                0          1
department
Analytics          90.433483   9.566517
Finance            91.876972   8.123028
HR                 94.375517   5.624483
Legal              94.898941   5.101059
Operations         90.985196   9.014804
Procurement        90.361446   9.638554
R&D                93.093093   6.906907
Sales & Marketing  92.796912   7.203088
Technology         89.240684  10.759316
All                91.482995   8.517005
```

La variable departamento, en todas sus categorías, no discrimina al target, ya que no se observa una diferencia significativa entre sí el colaborador es promovido o no.

```
#---------------------------------------#
Cruce 'is_promoted' con 'gender':
#---------------------------------------#
is_promoted          0         1
gender
f            91.006621  8.993379
m            91.684850  8.315150
All          91.482995  8.517005
```

La variable género igual, en sus dos categorías no discrimina al target, no se observa una diferencia significativa entre sí el colaborador es promovido o no.

```
#---------------------------------------#
Cruce 'is_promoted' con 'previous_year_rating':
#---------------------------------------#
is_promoted                0         1
previous_year_rating
1.0                98.585891  1.414109
2.0                95.715976  4.284024
3.0                92.722097  7.277903
4.0                92.062367  7.937633
5.0                83.638532 16.361468
sin_rating         91.779825  8.220175
All                91.482995  8.517005
```

En la variable previous year rating igual, se puede notar una diferencia en la categoría 5, en comparación a las demás, lo que indicaría que en este caso esta variable si está discriminando al target en al menos una de sus categorías.

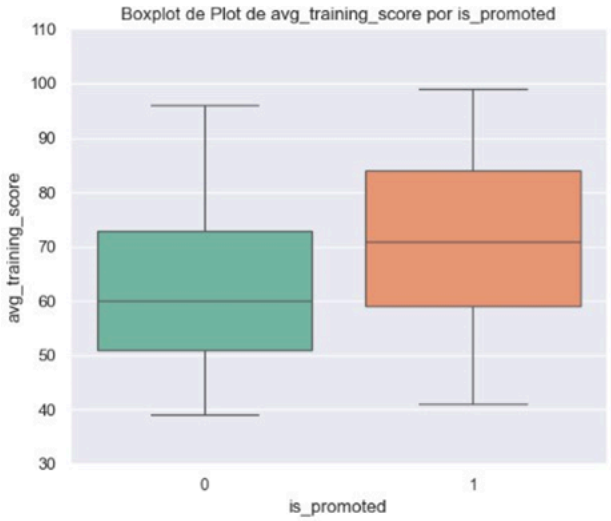
```
#---------------------------------------#
Cruce 'is_promoted' con 'education':
#---------------------------------------#
is_promoted              0         1
education
Bachelor's       91.796886  8.203114
Below Secondary  91.677019  8.322981
Master's & above 90.144054  9.855946
no_identificado  94.935658  5.064342
All              91.482995  8.517005
```

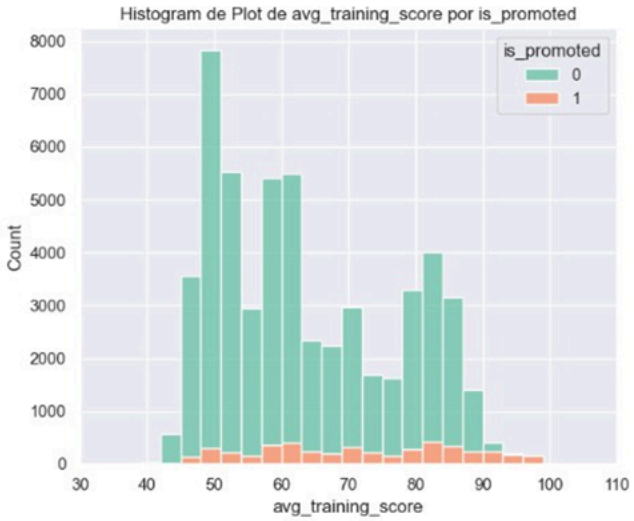
Por otro lado, variable educación, se observa que en sus categorías el target no es discriminado, no se observa diferencias significativas entre sí el colaborador es promovido o no.

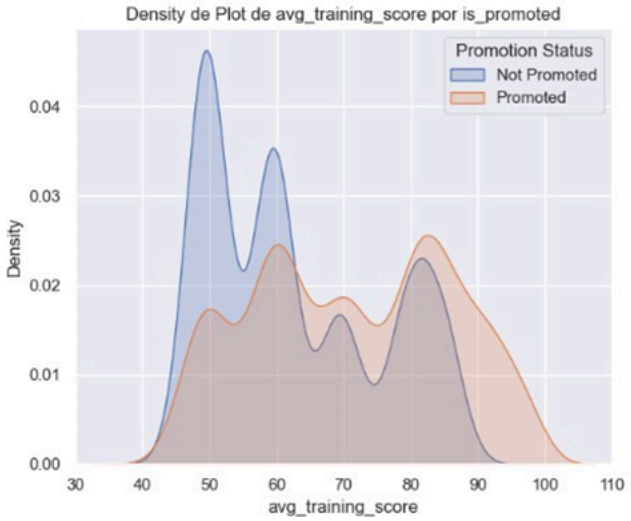
```
#---------------------------------------#
Cruce 'is_promoted' con 'recruitment_channel':
#---------------------------------------#
is_promoted                 0         1
recruitment_channel
other               91.604809  8.395191
referred            87.915937 12.084063
sourcing            91.498708  8.501292
All                 91.482995  8.517005
```

La variable recruitment channel, tiene una clara discriminaci´on en la categoría referidos, siendo un 12 % los que son referidos mientras que las demás categorías hay menos de 8.2

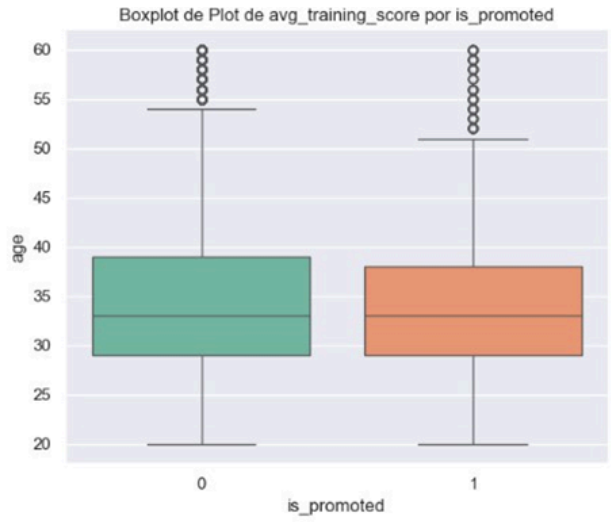
- Variables numericas

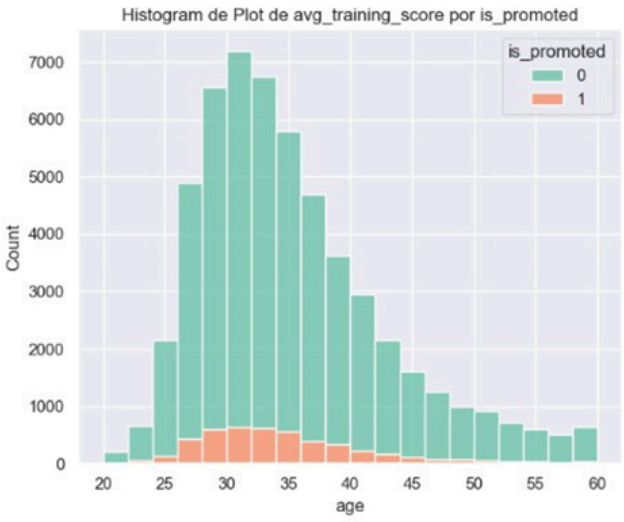


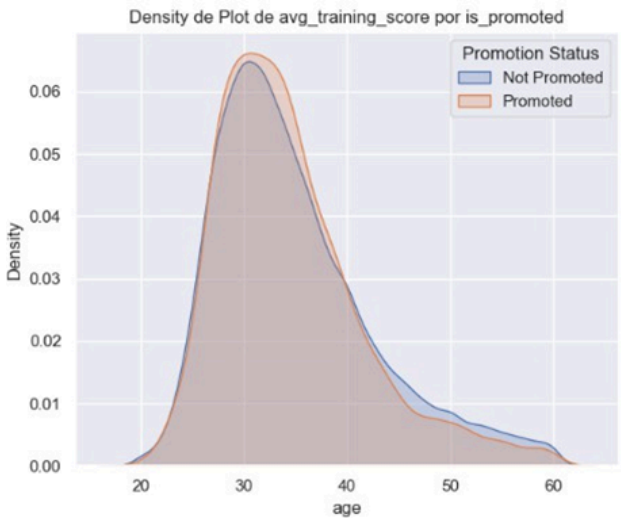




- **avg training score:**
  Se puede observar que hay una clara diferenciación en la distribución del puntaje promedio de las evaluaciones actuales respecto a si un colaborador es promovido o no. Para el caso de un colaborador promovido la distribuci´on se centra(hay m´as informaci´on) entre puntajes de 60 a 85 y por otro lado, donde los trabajadores no son promovidos la distribuci´on se centra entre valores de 50 a 70.







- **age:**
  Se observa que la variable edad no se diferencia entre el target, la distribución es casi similar. Podemos decir que esta variable no discrimina al target.

8. Principales drivers o factores que podrían ayudarnos a explicar el problema de estudio

De acuerdo al análisis en el punto anterior podemos identificar que drivers o factores podrían ayudarnos a explicar el problema de estudio.
Si se quiere identificar a los trabajadores que tienen alta probabilidad de ser promovidos debemos buscar que variables afectan positivamente al target, osea que lo discriminen bien.
Por esa razón las variables son las siguientes:

- KPIs met >80 %
- awards won
- previous year rating
- recruitment channel
- avg training score

9. Análisis del problema

9.1. Selección de muestra de entrenamiento (80%) y de evaluación (20%)

El conjunto de datos se divide en muestras de entrenamiento (80%) y prueba (20%) de manera estratificada para asegurar que las proporciones de la variable objetivo se mantengan en ambos conjuntos.

is_promoted	Frecuencia	Porcentaje (%)
NO_PROMOVIDO	50140	91.48%
PROMOVIDO	4668	8.52%

Distribución de Clases en el Dataset Original

is_promoted	Frecuencia	Porcentaje (%)
NO_PROMOVIDO	40112	91.48%
PROMOVIDO	3734	8.52%

Distribución de Clases en el Conjunto de Entrenamiento

is_promoted	Frecuencia	Porcentaje (%)
NO_PROMOVIDO	10028	91.48%
PROMOVIDO	934	8.52%

Distribución de Clases en el Conjunto de Prueba

9.2. Balanceo con Borderline-SMOTE

Se trabajó con dos tipos de balanceo, uno para cada modelo de interés. Se realizó de la siguiente manera:

is_promoted	Frecuencia	Porcentaje (%)
NO_PROMOVIDO	40112	50.00%
PROMOVIDO	40112	50.00%

Distribución de Clases después de aplicar Borderline-SMOTE para el modelo Árbol de Clasificación con el Algoritmo Bagging

is_promoted	Frecuencia	Porcentaje (%)
NO_PROMOVIDO	40112	68.97%
PROMOVIDO	18050	31.03%

Distribución de Clases después de aplicar Borderline-SMOTE de forma parcial para el modelo Catboost con smote y gritseach

9.3. Entrenamiento de los modelos

9.3.1. Modelo con Árbol de Clasificación con el Algoritmo Bagging

Resultados con Bagging:

Métrica	Valor
Sensibilidad	0.935082
Especificidad	0.961408
Accuracy	0.948245
Balanced Accuracy	0.948245

Matriz de Confusión:

	NO_PROMOVIDO (Pred)	PROMOVIDO (Pred)
NO_PROMOVIDO (Real)	38564	1548
PROMOVIDO (Real)	2604	37508

9.3.2. Modelo con Catboost gritseach

Resultados con Catboost y smote:

Métrica	Valor
Sensibilidad	0.886925
Especificidad	0.998604
Accuracy	0.963946
Balanced Accuracy	0.942765

Matriz de Confusión:

	NO_PROMOVIDO (Pred)	PROMOVIDO (Pred)
NO_PROMOVIDO (Real)	40056	56
PROMOVIDO (Real)	2041	16009

9.4. Predicción de los modelos en la data testing

9.4.1. Predicción del Modelo Árbol de Clasificación con el Algoritmo Bagging en la data testing

Resultados del Modelo con Bagging en la data testing:

Métrica	Valor
Sensibilidad	0.3801
Especificidad	0.9735
Accuracy	0.9229
Balanced Accuracy	0.6768
Log-Loss	0.8372
Average Precision	0.4971

Matriz de Confusión:

	NO_PROMOVIDO (Pred)	PROMOVIDO (Pred)
NO_PROMOVIDO (Real)	9762	266
PROMOVIDO (Real)	579	355

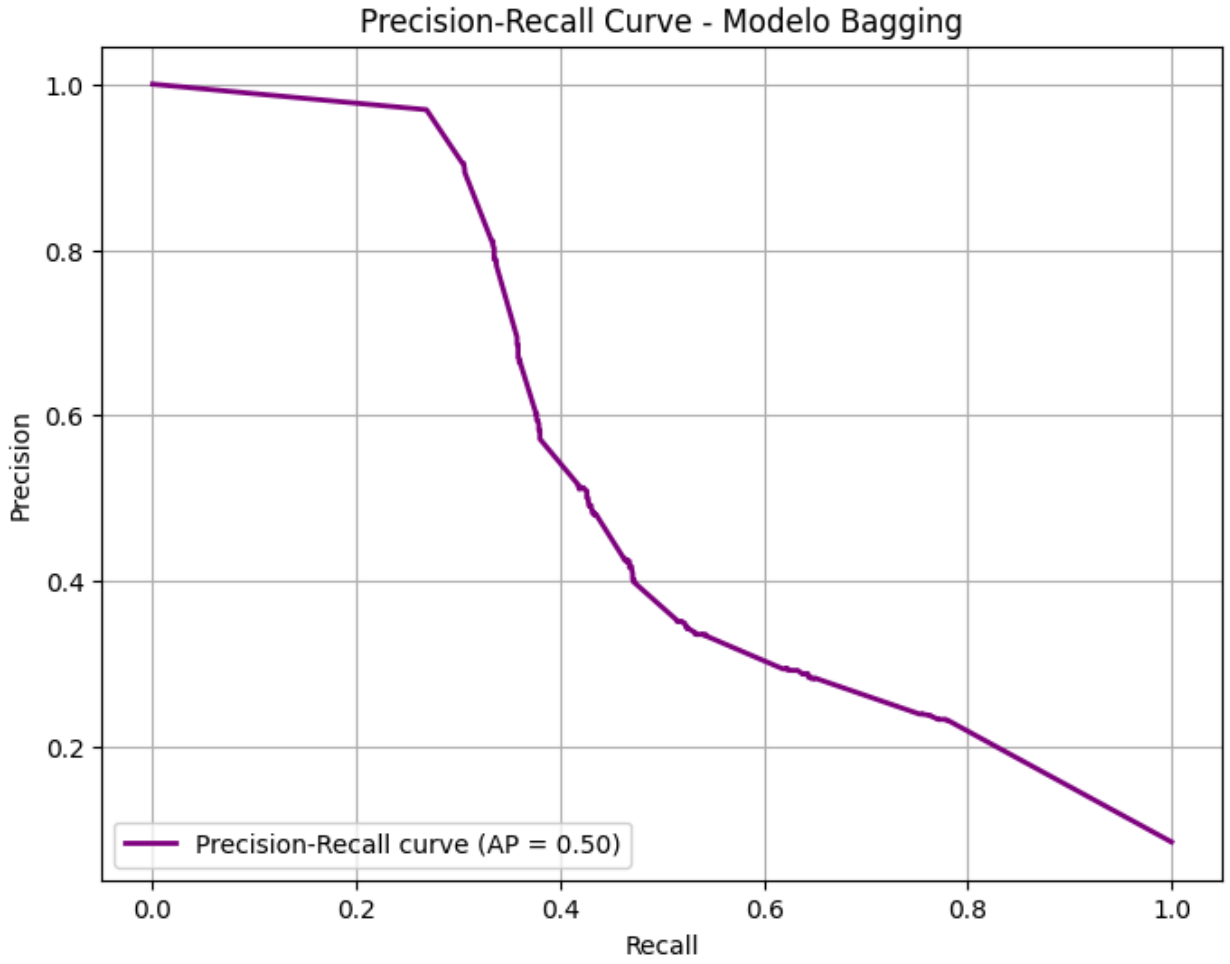


Gráfico de la curva Precision-Recall para el Modelo Bagging

Interpretación:

- Al inicio, el modelo mantiene una precisión alta cuando el recall es bajo, lo que significa que los pocos casos positivos que identifican son clasificados correctamente.
- A medida que el modelo intenta identificar más casos positivos (aumentando el recall), la precisión cae rápidamente, indicando que está incluyendo muchos falsos positivos en sus predicciones.

9.4.2. Predicción del Modelo Catboost con smote y grid en la data testing

Resultados del Modelo Catboost con smote y grid en la data testing:

Métrica	Valor
Sensibilidad	0.3704
Especificidad	0.9929
Accuracy	0.9399
Balanced Accuracy	0.6817
Log-Loss	0.1739
Average Precision	0.5863

Matriz de Confusión:

	NO_PROMOVIDO (Pred)	PROMOVIDO (Pred)
NO_PROMOVIDO (Real)	9957	71
PROMOVIDO (Real)	588	346

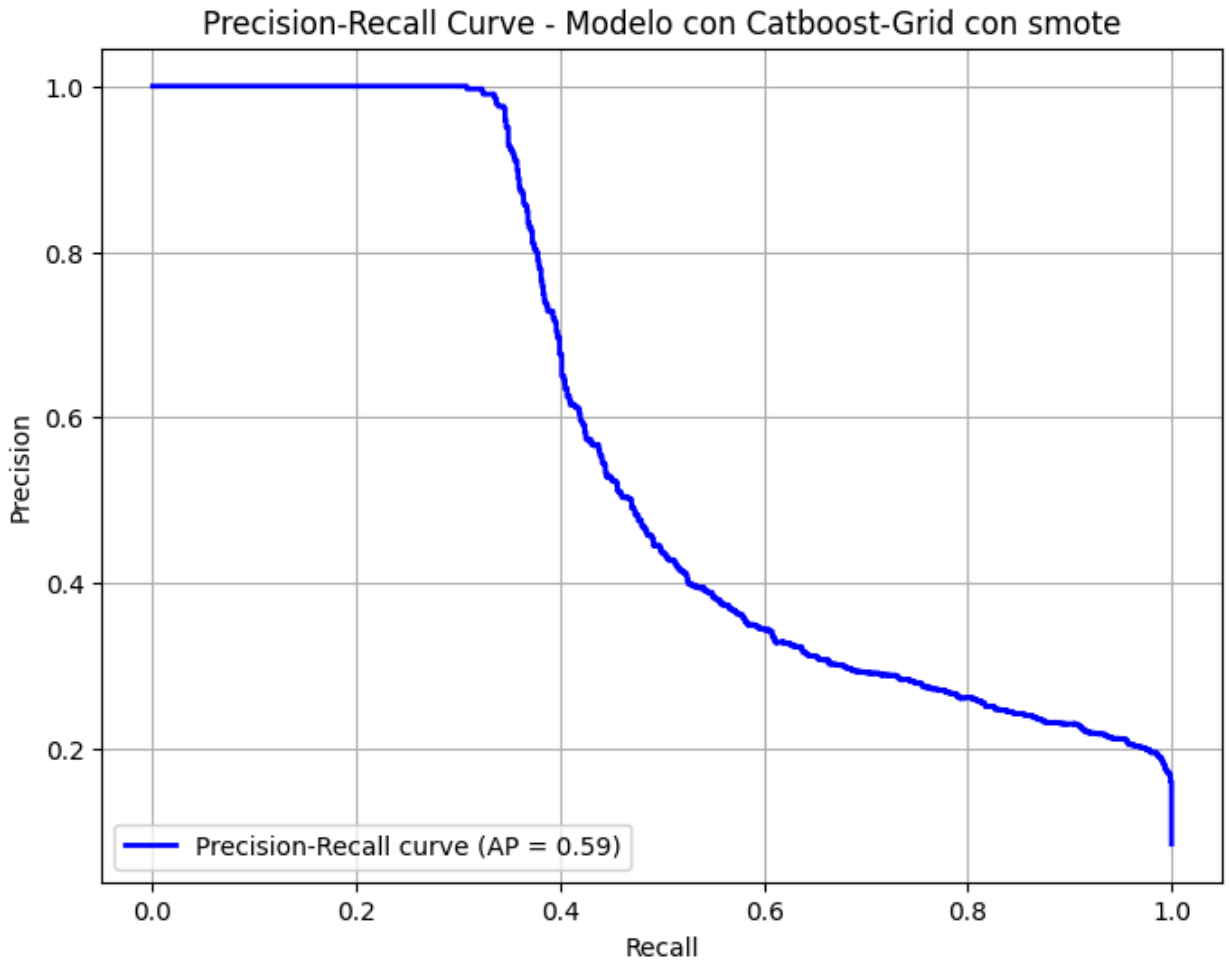


Gráfico de la curva Precision-Recall para el Modelo Catboost con smote y grid

Interpretación:

- A diferencia del modelo Bagging, este modelo mantiene la precisión durante un rango mayor de valores de recall antes de comenzar a caer. La curva más gradual sugiere que puede capturar más casos positivos sin sacrificar tanto la precisión.
- Este modelo es mejor para maximizar el recall, ya que permite alcanzar niveles más altos de recuperación sin una caída tan abrupta en precisión. Esto lo convierte en una opción más robusta si el objetivo es capturar la mayor cantidad de casos "PROMOVIDO" posible.

9.5. Interpretación de Resultados, Conclusiones y Recomendaciones

9.5.1. Interpretación de Resultados

Comparación de Accuracy entre Entrenamiento y Predicción:

Modelo	Accuracy (train)	Accuracy (test)
Bagging	0.948245	0.922916
Catboost-Grid	0.963946	0.939900

En ambos modelos tenemos una predicción cercana a lo esperado en el entrenamiento por lo tanto ambos modelos son buenos al momento de predecir la promoción del empleado.

Métricas de los modelos en la data testing:

Métrica	Bagging	Catboost-Grid
Sensibilidad	0.3801	0.3704
Especificidad	0.9735	0.9929
Accuracy	0.9229	0.9399
Balanced Accuracy	0.6768	0.6817
Log-Loss	0.8372	0.1739
Average Precision	0.4971	0.5863

- Average Precision: Este es el indicador clave para la evaluación de la precisión en un contexto de clases desbalanceadas, tanto Bagging como Catboost-Grid muestran un rendimiento destacable con valores de 0.4971 y 0.5863 respectivamente.

- Accuracy: Bagging como Catboost-Grid también se destacan con un Accuracy de 0.9293 y 0.9399 en la data de test, los cuales son muy alto y se acerca al máximo valor. Los modelos logran hacer predicciones correctas en una gran proporción de casos, lo que los colocan como los mejores modelos evaluados, y se acercan a sus train con unos Accuracy de 0.922916 y 0.939900, lo que indica una predicción estable.

- Log-Loss: Bagging y Catboost tienen un Log-Loss de 0.8372 y 0.1739, que son considerablemente bajo (Sobre todo el de Catboost), lo que indica que los modelos está realizando buenas predicciones probabilísticas, y las diferencias entre las predicciones y los valores reales son pequeñas. Este es un buen indicador de la capacidad de los modelos para clasificar correctamente mientras minimiza los errores.

9.5.2. Conclusiones y Recomendaciones

- Conclusiones:
  El modelo Bagging se destaca como el más robusto al combinar alto rendimiento en Accuracy , Average Precision y Log-Loss . Aunque el modelo Catboost muestra un rendimiento superior en Especificidad y logra el Log-Loss más bajo, su Sensibilidad es ligeramente menor en comparación con Bagging, lo que podría limitar su capacidad para identificar correctamente casos de promoción. La estabilidad de ambos modelos es evidente, ya que ambos presentan valores de precisión en el conjunto de prueba que son cercanos a los alcanzados durante el entrenamiento, indicando una buena generalización.

  El modelo Catboost con ajuste de hiperparámetros por grid search y balanceo de clases con SMOTE tiene un excelente Log-Loss de 0.1739, considerablemente más bajo que el de Bagging, lo que indica que Catboost realiza predicciones probabilísticas muy precisas, minimizando la diferencia entre los valores predichos y los valores reales. Asimismo, Catboost se destaca en Average Precision con un valor de 0.5863, superior al obtenido por Bagging. Esto hace de Catboost una opción fuerte y confiable para la predicción en situaciones de datos desequilibrados.

- Recomendaciones:
  Implementar Bagging : Dado su desempeño general, Bagging es recomendado para la implementación en predicciones de promoción, ya que ofrece una excelente precisión en la clasificación y un Log-Loss bajo que respalda la estabilidad de sus predicciones. Su alta Sensibilidad también asegura que identifica un buen número de casos positivos, siendo adecuado para el contexto de clases desequilibradas.

  Explorar Catboost como una alternativa avanzada : Aunque Bagging es más robusto en general, Catboost presenta una excelente Especificidad y un Log-Loss mucho menor, lo que lo convierte en una opción sólida si el enfoque principal es la precisión en la predicción de casos negativos y la

minimización de errores probabilísticos. Catboost puede ser especialmente útil en casos donde la especificidad es prioritaria o en contextos con un alto costo de falsos positivos.

Link de git: https://github.com/CieloLozada/Proyecto-final---Futura-

- Aquí pueden encontrar los notebooks de los modelos realizados.