Equipo Data

Proyecto de modelo predictivo de ingresos

Alfredo André Durán Treviño - A01286222
Eliani González Laguna - A00836712
Fedra Fernanda Mandujano López - A00835797
Juan Marco Castro Trinidad - A01742821
Miranda Isabel Rada Chau - A01285243

Objetivo

Desarrollar un modelo predictivo para determinar si una persona tiene ingresos por encima o por debajo de 25,000 pesos mexicanos.

Con el propósito de identificar factores influyentes en los ingresos para comprender mejor las condiciones socioeconómicas en México y tener los datos necesarios para ser aplicados en la toma de decisiones informadas en políticas sociales con el objetivo de reducir la brecha económica de la población mexicana.

EXPLORACIÓN Y COMPRENSIÓN DE DATOS

Para reducir la dimensionalidad de la base de datos original con 192 varaibles y 309638 registros y preparar los datos para el modelado predictivo de ingresos se realizó:

- Eliminación de variables no relevantes para la predicción de ingresos como:
  - Identificadores
  - Causas específicas de discapacidad
  - Detalles precisos de actividades diarias.

- Variables Mantenidas:
  - Seleccionadas 25 variables más relevantes que incluyen características socioeconómicas, educativas y de salud.

Preparación de la Variable Objetivo:

- Creación de la variable objetivo 'ingreso_prom', que indica si el ingreso es mayor o menor/igual a 25,000 pesos mexicanos.

Limpieza y preparación de datos

Para fines del proyecto se eliminaron todas las filas con edades menores a 18 años para enfocar el análisis en adultos con ingresos relevantes.

Manejo de Valores Nulos:

- Eliminación de columnas con valores nulos
- Rellenado de valores nulos:
  - Columnas categóricas: Rellenadas con la moda.
  - Columnas numéricas: Rellenadas con la media.

Agrupación de Columnas

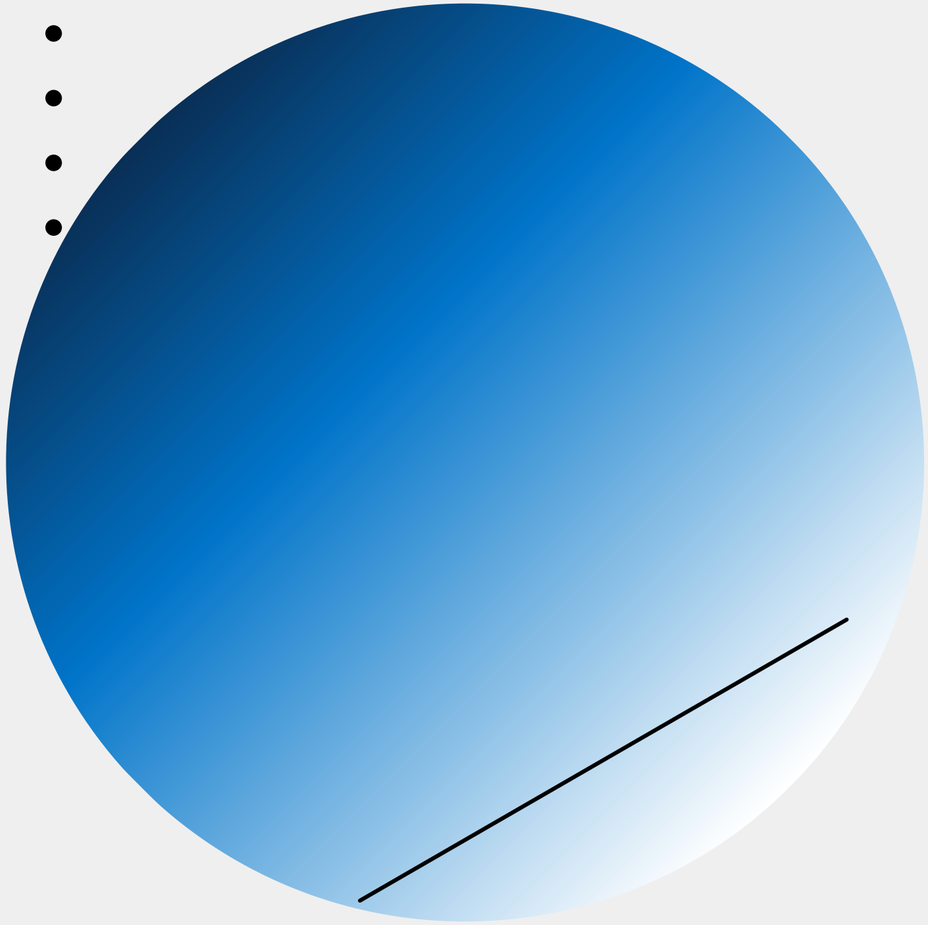
- Combinación de múltiples columnas de discapacidad en una sola ('disc') para obtener al número de personas con alguna discapacidad

Transformación de Datos

- Conversión de columnas de tipo 'object' a tipo numérico para análisis y modelado.

Data Engineering

- Decisión de mantener solo las columnas con pocos valores nulos para evitar sesgos en el análisis final.
- Enfoque en la calidad y coherencia de los datos para garantizar resultados precisos en el modelado predictivo.



Análisis Univariado

Variables Categóricas

Variable	Mode	Min	Max	Descripción
sexo	2	1	2	Mayoría femenino
disc	1	0	1	Mayoría sin discapacidad
hablaind	2	1	2	Mayoría no habla lengua indígena
etnia	2	1	2	Mayoría no se considera indígena
alfabetism	1	1	2	Mayoría sabe leer
asis_esc	2	1	2	Mayoría no asiste a la escuela
nivelaprob	4	0	9	Mayoría terminó preparatoria
edo_conyug	2	1	6	Mayoría casados

Variable	Mode	Min	Max	Descripción
segsoc	1	1	2	Mayoría contribuye a seguro social
entidad	8	1	32	Mayoría en Chihuahua
hablaesp	1	1	2	Mayoría habla español
atemed	1	1	2	Mayoría afiliados a atención médica
pagoaten_1	0	0	1	Mayoría no pagó por consulta
diabetes	2	1	2	Mayoría no tiene diabetes
pres_alta	1	1	2	Mayoría no tiene presión alta
trabajo_mp	1	1	2	Mayoría trabajó el mes pasado
num_trabaj	1	1	2	Mayoría tiene un trabajo
usotiempo1	9	8	9	Mayoría no trabajó la semana pasada
ingreso_prom	0	0	1	Relacionado con ingreso >25k

Variables Númericas

Variable	Mean	Std	Min	Max	Descripción
edad	43.55	17.48	18	109	Edad promedio, eliminados menores
hor_1	23.39	25.61	0	168	Horas trabajadas promedio

Variables no consideradas

De estas dos columnas se calculó la columna objetivo. Están altamente relacionadas

Variable	Mean	Std	Min	Max	Descripción
ing_tri_max	22440.66	45432.95	1.46	6854754.09	Ingreso máximo
ing_tri_total	21913.14	49854.46	0	6854754.09	Ingreso total

Valores atípicos encontrados

- **Edad**

Se identificó un número reducido de personas con 100 años o más de edad (21 datos). Aunque no fueron eliminadas para los fines del proyecto, como buena práctica deberían considerarse para eliminación debido a su escasa representación en el conjunto de datos.

- **Horas trabajadas**

Se detectaron varios casos en los que personas reportaron trabajar más de 125 horas a la semana, lo cual equivale a trabajar sin descanso. Este hallazgo indica una situación laboral extrema que requiere una revisión detallada y un análisis más profundo de la situación de cada quien. En este caso, se vio que hubo 35 personas que trabajaban más de 125 horas a la semana.

Análisis Bivariado

En esta sección se hicieron varios estudios para identificar el comportamiento de ciertas variables en relación con el ingreso promedio. Después de hacer este análisis pudimos ver que hay dos categorías en donde se puede ver una relación muy clara con los ingresos.

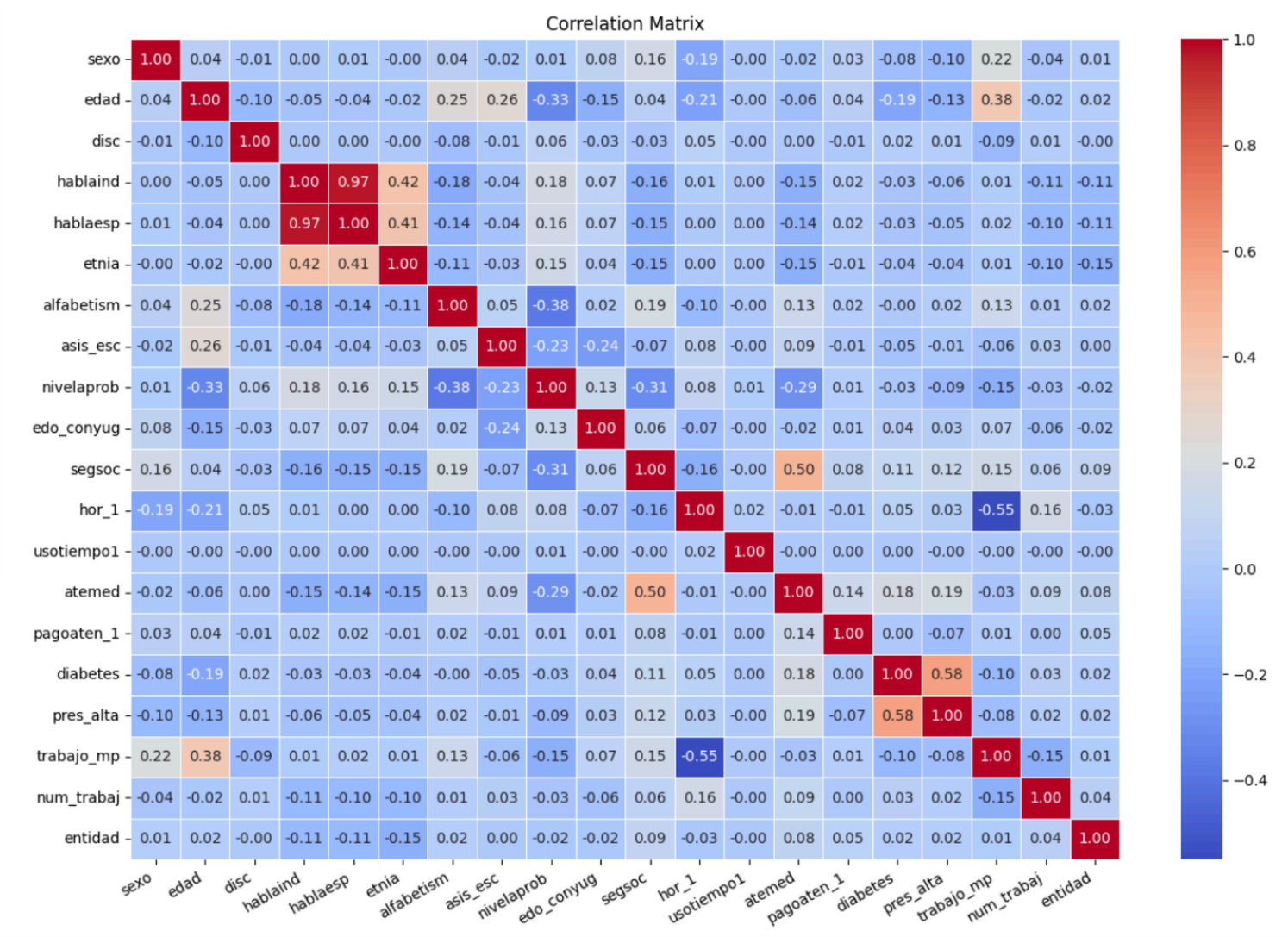
- **Edad**

En el gráfico de la edad, se puede ver una clara relación entre la cantidad de ingresos y la edad. Se puede ver que la mayor parte de las personas que caen arriba del umbral están en un rango de edad entre 20-65.

- **Nivel Aprobado**

En el caso del nivel aprobado, se puede ver una relación muy clara entre los niveles más altos de educación y la cantidad de personas con ingresos por encima de 25,000.

Matriz de Correlación



Ya habiendo seleccionado las variables que vamos a utilizar para el modelado, decidimos hacer un análisis para ver que relaciones existen entre estas variables. En general, se ven muy pocas variables altamente correlacionadas. Algunas de estas son: hablaind, hablaesp y segsoc y atemed.

Métodos

- **Modelos utilizados**

Para esta situación consideramos apropiado utilizar tanto árboles de decisión como modelos de Random Forest. Cada uno de estos modelos tenía sus ventajas y desventajas. El desempeño de cada uno de los modelos y métodos utilizados se muestra más adelante.

- **División de datos**

Antes de comenzar con la creación de los modelos, fue necesario hacer el split de los datos en la muestra de entrenamiento y en la muestra de prueba. En este caso, nosotros decidimos usar el 80% de los datos para entrenamiento y un 20% para las pruebas y la validación.

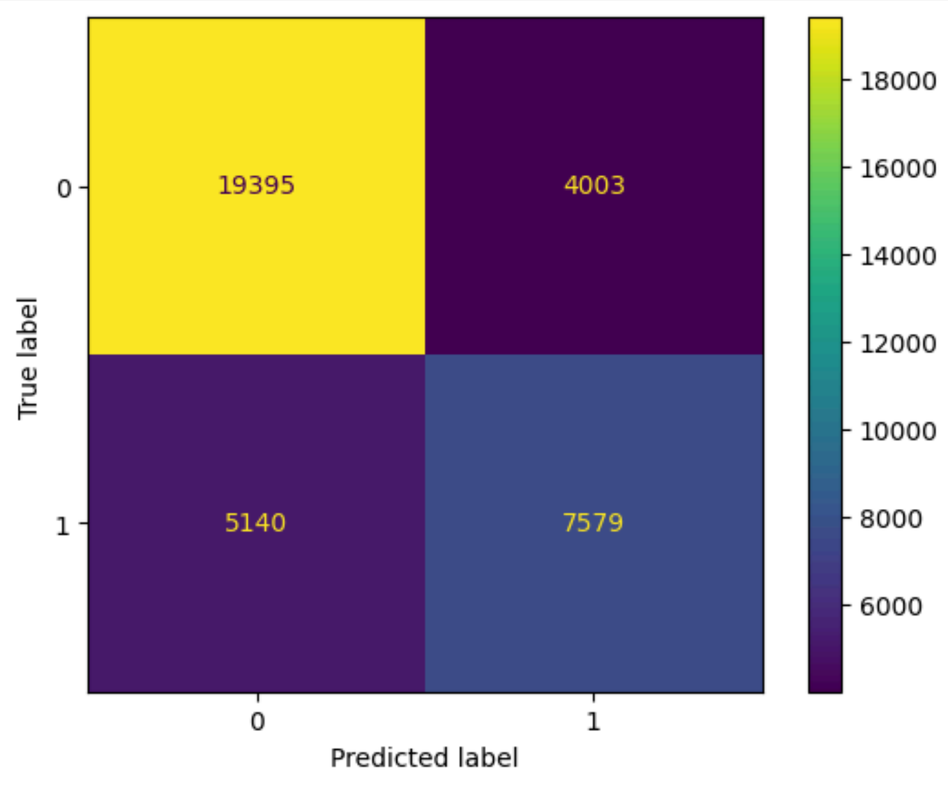
Árbol de Decisión

Para nuestro modelo simple utilizamos un arbol de decision de clasificación y obtuvimos los siguientes resultados:

Métricas:

- La exactitud (0.7468).
- La precisión (0.6543).
- La sensibilidad (0.5958).
- El puntaje F1 (0.6237).
- El AUC (71.239).

Matriz de Confusión:



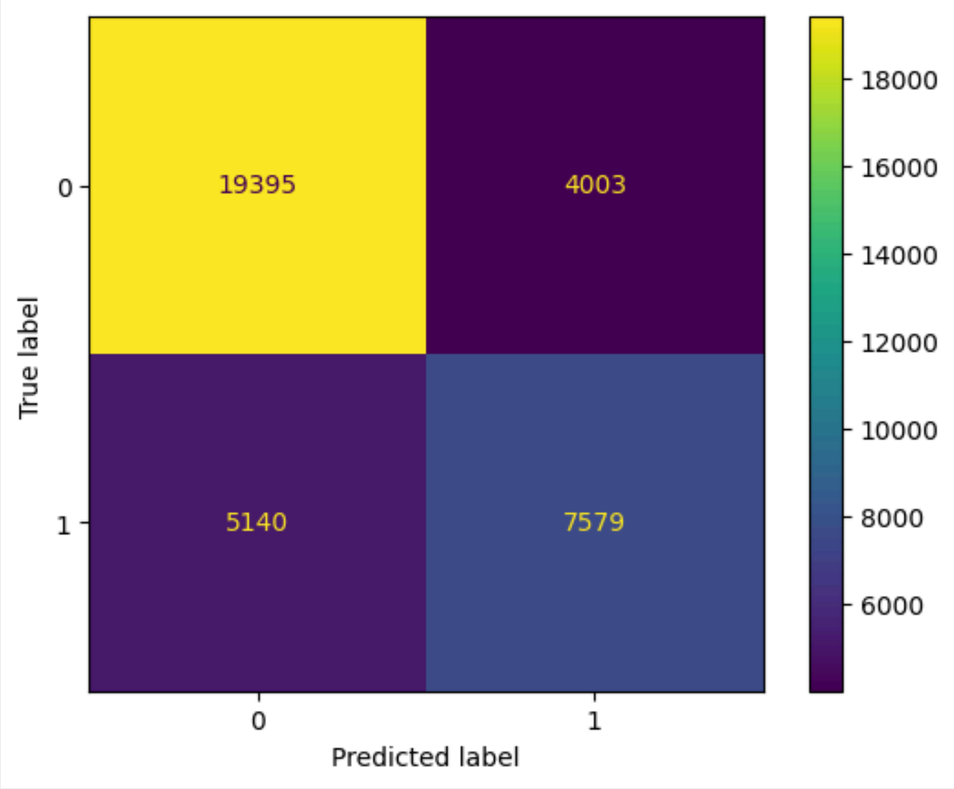
Árbol de Decisión GridSearch

Para nuestro modelo optimizado utilizamos un árbol de decisión de clasificación, pero con hiperparámetros de Gridsearch y obtuvimos los mismos resultados:

Métricas:

- La exactitud (0.7468).
- La precisión (0.6543).
- La sensibilidad (0.5958).
- El puntaje F1 (0.6237).
- El AUC (71.239).

Matriz de Confusión:



Árbol de Decisión - Método Over-sampling

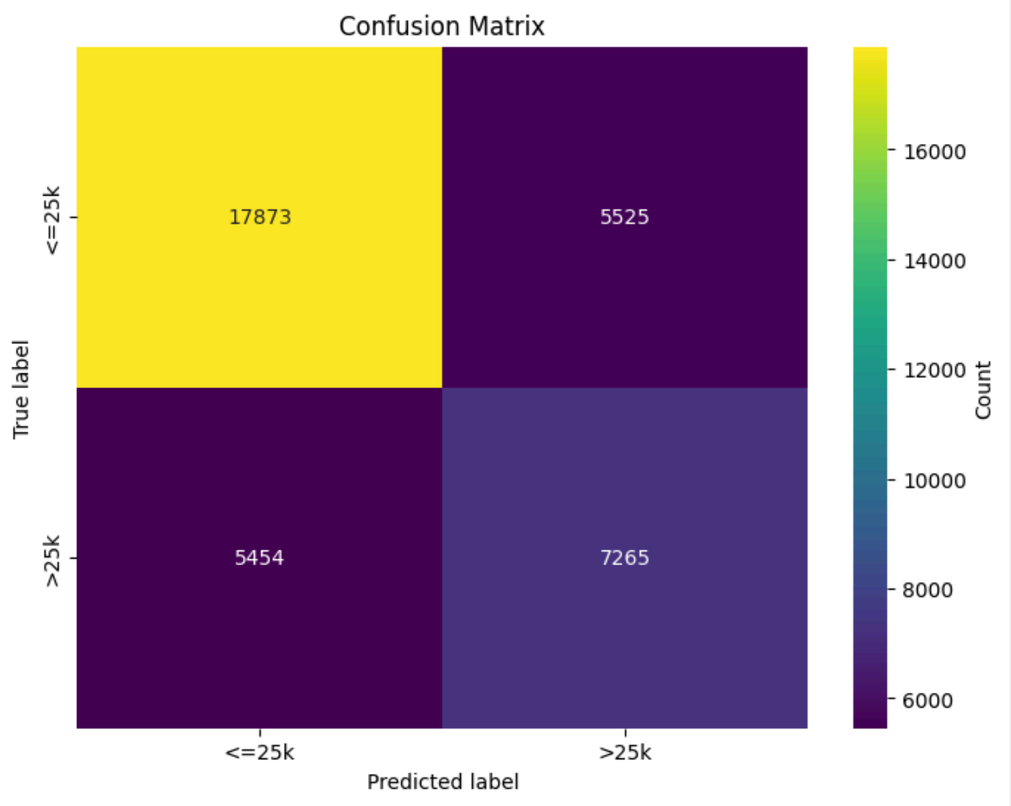
Grid Search

Para este modelo se utilizó un árbol de decisión clasificatorio y se obtuvo los siguientes resultados con la técnica de over-sampling buscando las mejores métricas con grid search:

Métricas:

- La exactitud (0.696).
- La precisión (0.568).
- La sensibilidad (0.5712).
- El puntaje F1 (0.569).
- El AUC (66.75).

Matriz de Confusión:



Árbol de Decisión - Método Under-sampling

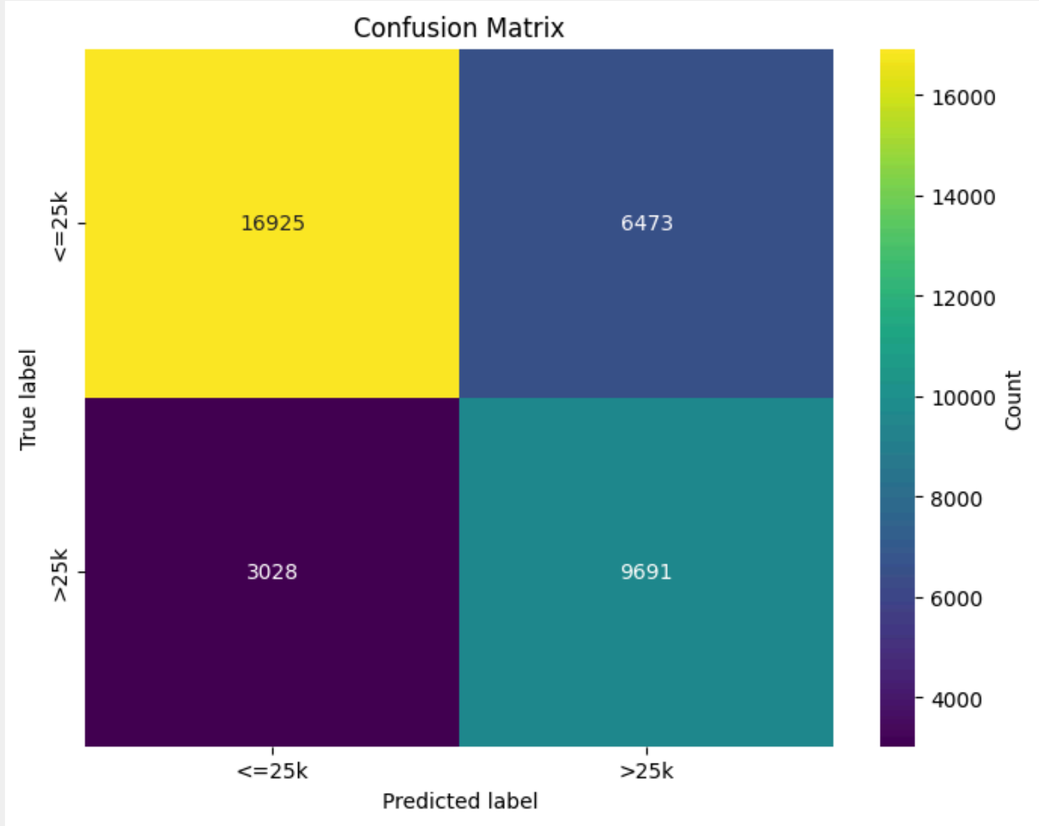
Grid Search

Para este modelo se utilizó un árbol de decisión clasificatorio y se obtuvo los siguientes resultados con la técnica de under-sampling buscando las mejores métricas con grid search:

Métricas:

- La exactitud (0.7369).
- La precisión (0.6).
- La sensibilidad (0.7619).
- El puntaje F1 (0.6711).
- El AUC (74.26).

Matriz de Confusión:



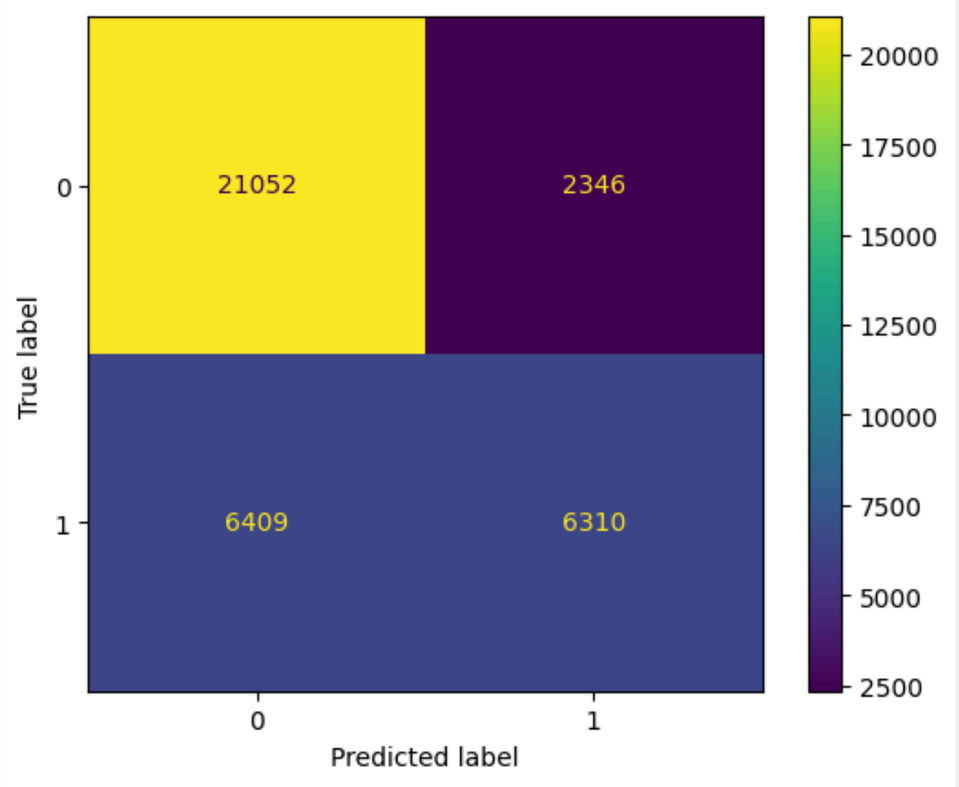
Random Forest

Este modelo combina varios clasificadores de árboles de decisión en diferentes submuestras y se ajusta con estos para mayor precisión y controlar el sobre ajuste.

Métricas:

- La exactitud (0.757).
- La precisión (0.728).
- La sensibilidad (0.496).
- El puntaje F1 (0.590).
- El AUC (69.7).

Matriz de Confusión:



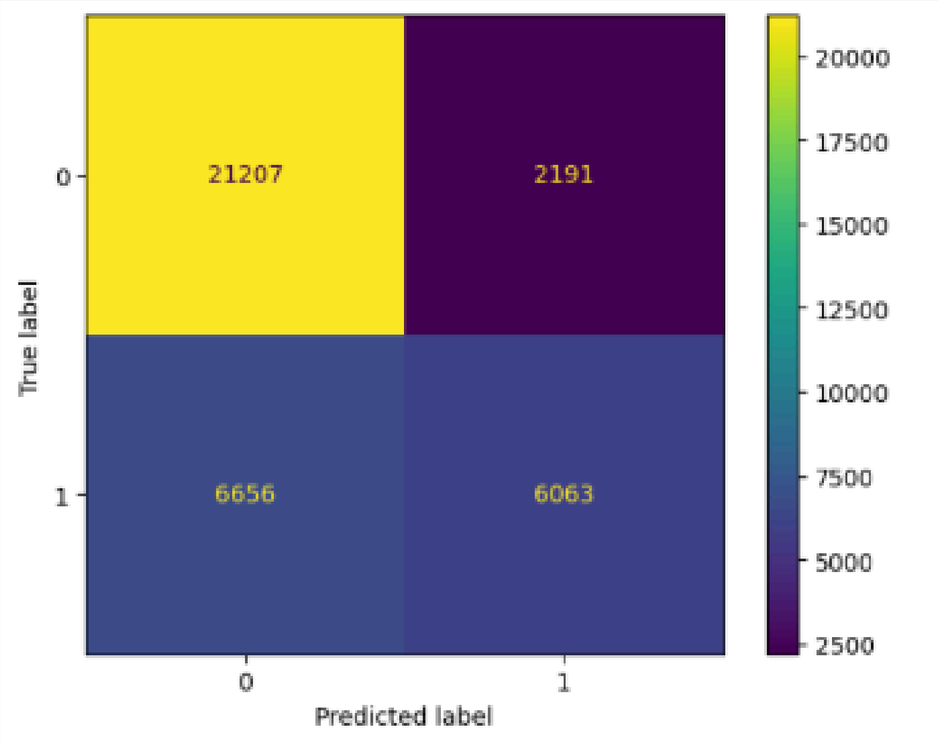
Random Forest Grid Search

Se buscan los mejores parametros del modelo para nuestro datos.

Métricas:

- La exactitud (0.755).
- La precisión (0.734).
- La sensibilidad (0.476).
- El puntaje F1 (0.578).
- El AUC (69.79).

Matriz de Confusión:



RF Undersampling

Nuestro mejor modelo resulto ser un Random Forest con GridSearch y Undersampling, el cual nos dio los siguientes resultados.

Métricas:

- La exactitud (0.7391).
- La precisión (0.6020).
- La sensibilidad (0.7646).
- El puntaje F1 (0.6736).
- El AUC (74.49).

Matriz de Confusión:

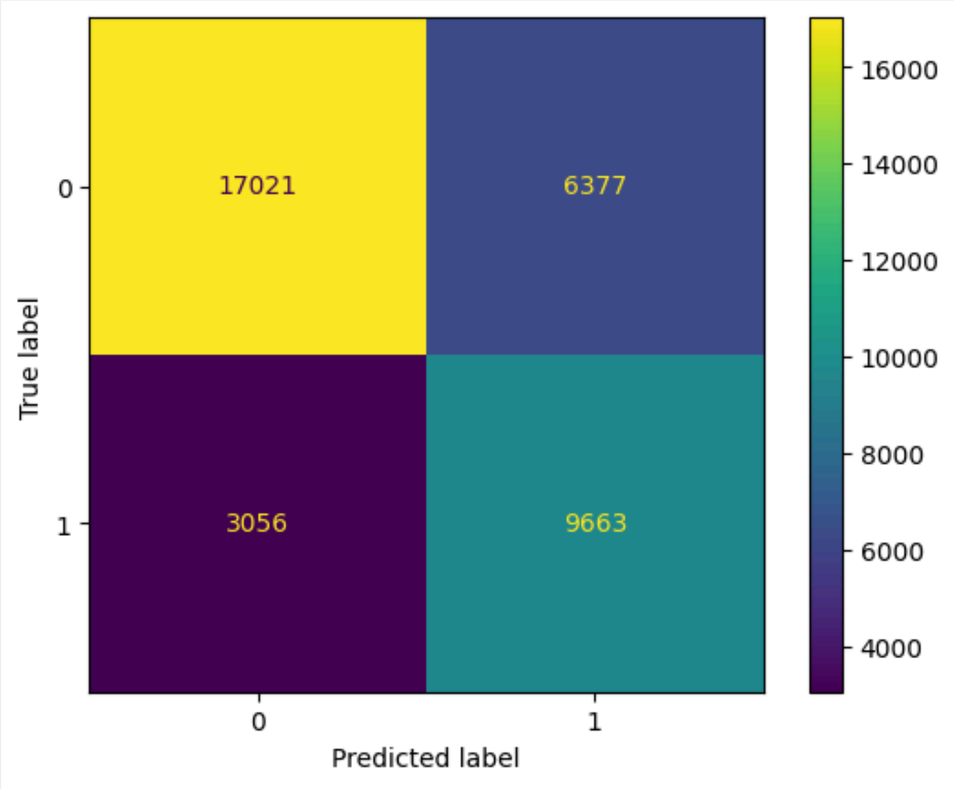
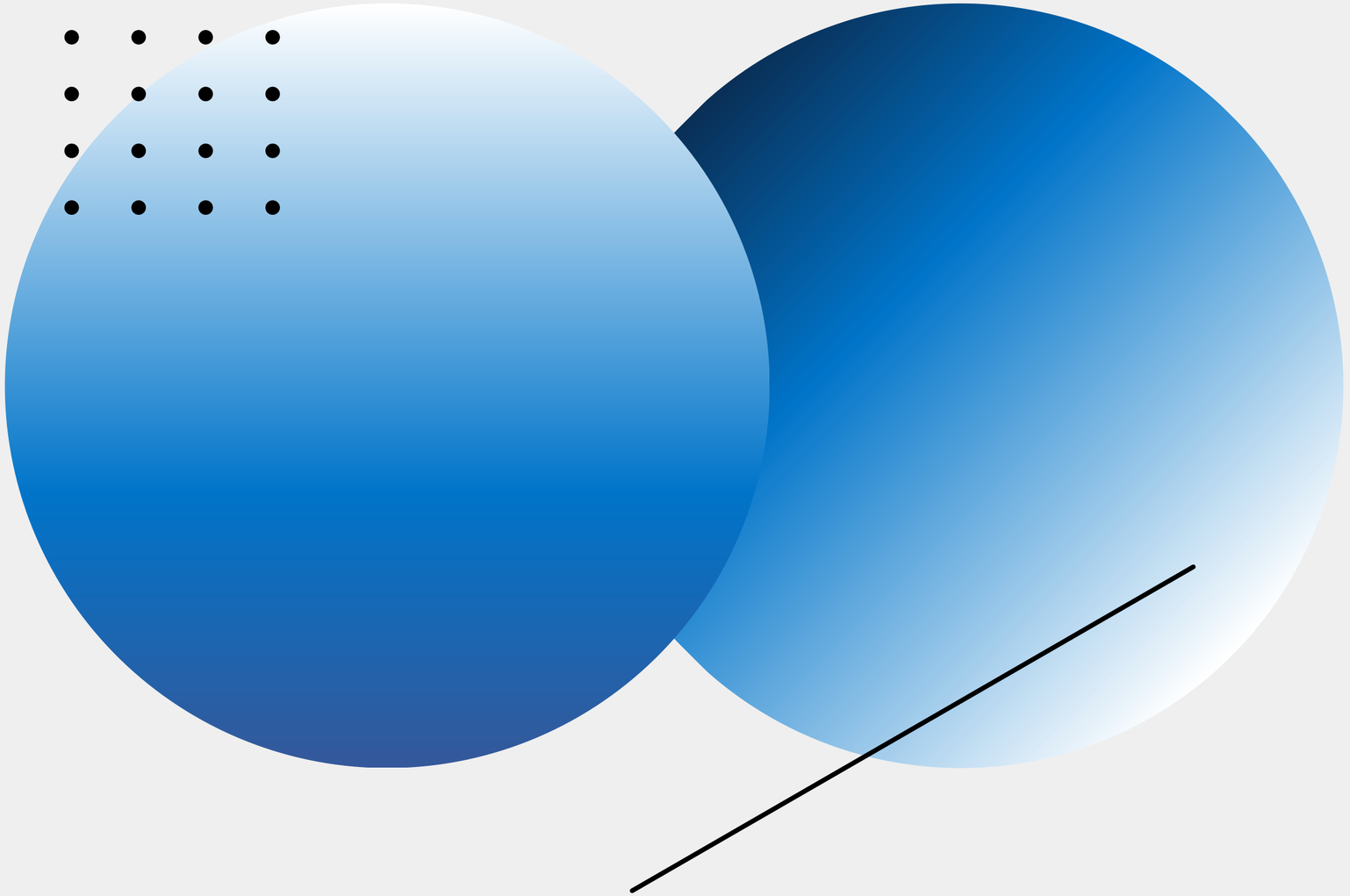


Tabla comparativa de las metricas de desempeño

Modelo	Variables	VP	FP	VN	FN	Total	Precison	Recall	Accurracy	F1 score	AUC
Árbol de desición	23	7579	4003	19395	5140	36117	0.6544	0.5959	0.7469	0.6238	0.7123
Árbol de desición con grid search	23	7579	4003	19395	5140	36117	0.6544	0.5959	0.7469	0.6238	0.7123
Árbol de desición con grid search y undersampling	23	9691	6473	16925	3028	36117	0.5995	0.7619	0.7369	0.6711	0.7426
Árbol de desición con grid search y oversampling	23	7265	5525	17873	5454	36117	0.5680	0.5712	0.6960	0.5996	0.6675
Random forest	23	6063	2191	21207	6656	36117	0.7346	0.4767	0.7550	0.5782	0.6915
Random forest con grid search	23	6063	2191	21207	6656	36117	0.7346	0.4767	0.7550	0.5782	0.6915
Random forest con grid search y undersampling	23	9725	6428	16970	2994	36117	0.6021	0.7646	0.7391	0.6737	0.7449



Mejor Modelo

Análisis del mejor modelo

En este caso, se pudo identificar que el mejor modelo fue el Random Forest con undersampling. Acorde a este modelo podemos identificar que en la matriz de confusión se determinaron 16,970 datos como negativos verdaderos y 9725 datos como positivos verdaderos, contando con solo 9422 errores de categorización de los datos. La dispersión de los conjuntos por mediciones de una magnitud representa una precisión considerable, por ende, los valores tienden a contar con cierta dispersión.

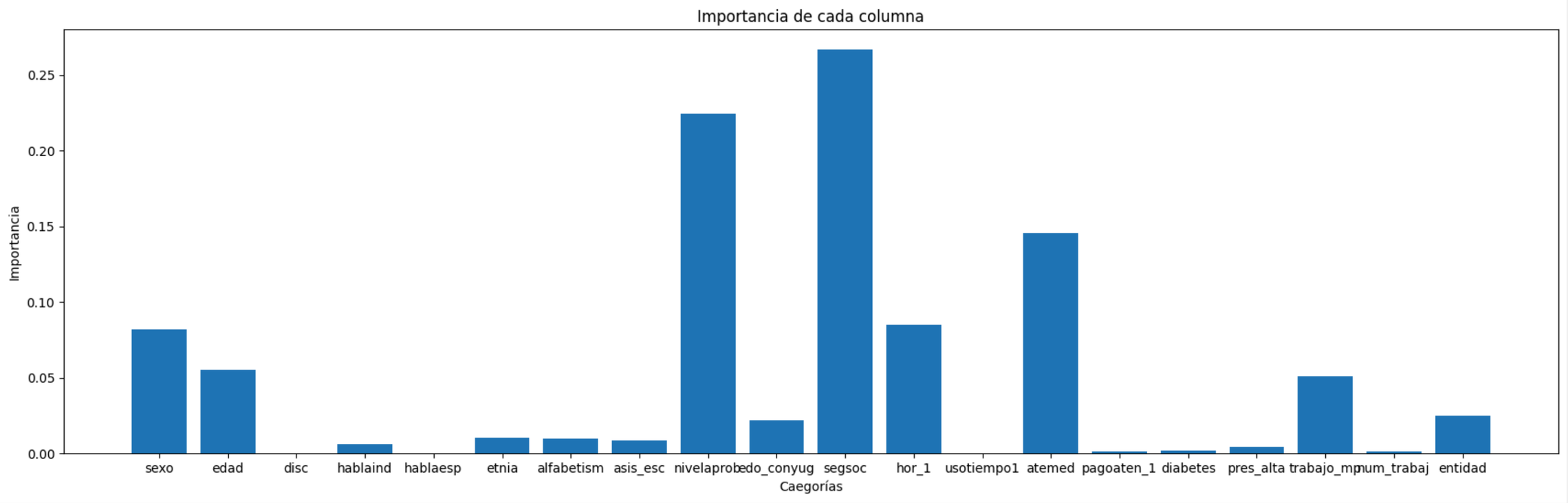
El ratio de verdaderos positivos indica una buena clasificación, al igual que, la fracción de predicciones en la precisión del modelo, pues, se cuenta con un balance adecuado. La capacidad de clasificación es la mejor siendo el modelo más preciso, como se ve en el grafico de la curva AUC.

Validación cruzada

Promedio de la precisión del conjunto de entrenamiento: 0.7671
Promedio de la precisión del conjunto de prueba: 0.7649

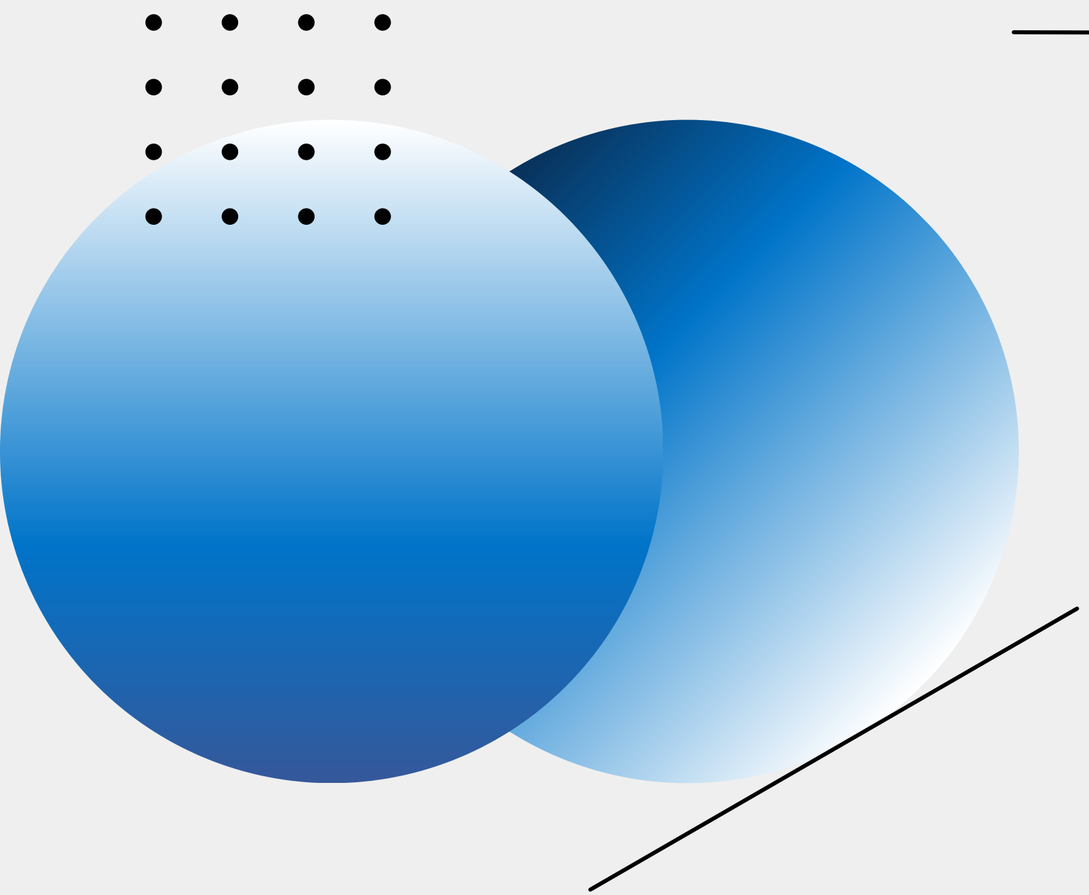
Lo cual representa la precisión con la que cuenta el modelo en base a la técnica de evaluación por medio del conjunto de entrenamiento y de validación. A su vez, estos buenos resultados indican que se evita el sobre ajuste y se proporciona una estimación más eficiente en el rendimiento del modelo.

Características importantes en el mejor modelo



Evaluación

Análisis de Desempeño



1. El modelo Random Forest con GridSearch y Under-sampling identifica bien los casos negativos y tiene un buen F1 score, indicando equilibrio entre precisión y sensibilidad.

2. La baja precisión implica falsos positivos, lo que es problemático para predecir ciudadanos por encima del umbral de $25,000.

3. Sobresale en el score AUC, mostrando una buena capacidad para diferenciar entre casos y ofreciendo una clasificación efectiva.

Implicaciones y limitaciones del mejor modelo

Manejo de Datos Desbalanceados:

- Ventajas:
  - Mejora el rendimiento del modelo al equilibrar las clases
  - Evita el sesgo hacia la clase mayoritaria

- Limitaciones:
  - Pérdida de patrones importantes al eliminar registros de la clase mayoritaria.
  - Puede afectar la capacidad del modelo para generalizar.

Aplicación de Grid Search

- Ventajas:
  - Encuentra los mejores hiperparámetros, personalizando el modelo.
  - Mejora la precisión y robustez del modelo.

- Limitaciones:
  - Riesgo de sobreajuste al dataset de entrenamiento.
  - Alto costo computacional.

Uso de Random Forest:

- Ventajas:
  - Reduce la varianza y evita el sobre ajuste (método bagging).
  - Robusto a valores atípicos y datos ruidosos.

- Limitaciones:
  - Incremento en el costo de memoria y tiempo de cómputo al aumentar el número de árboles.

Conclusiones

Al final de nuestro análisis, hemos comprendido la importancia de elegir el tipo de modelo adecuado, considerando tanto su precisión como el tiempo de procesamiento requerido.

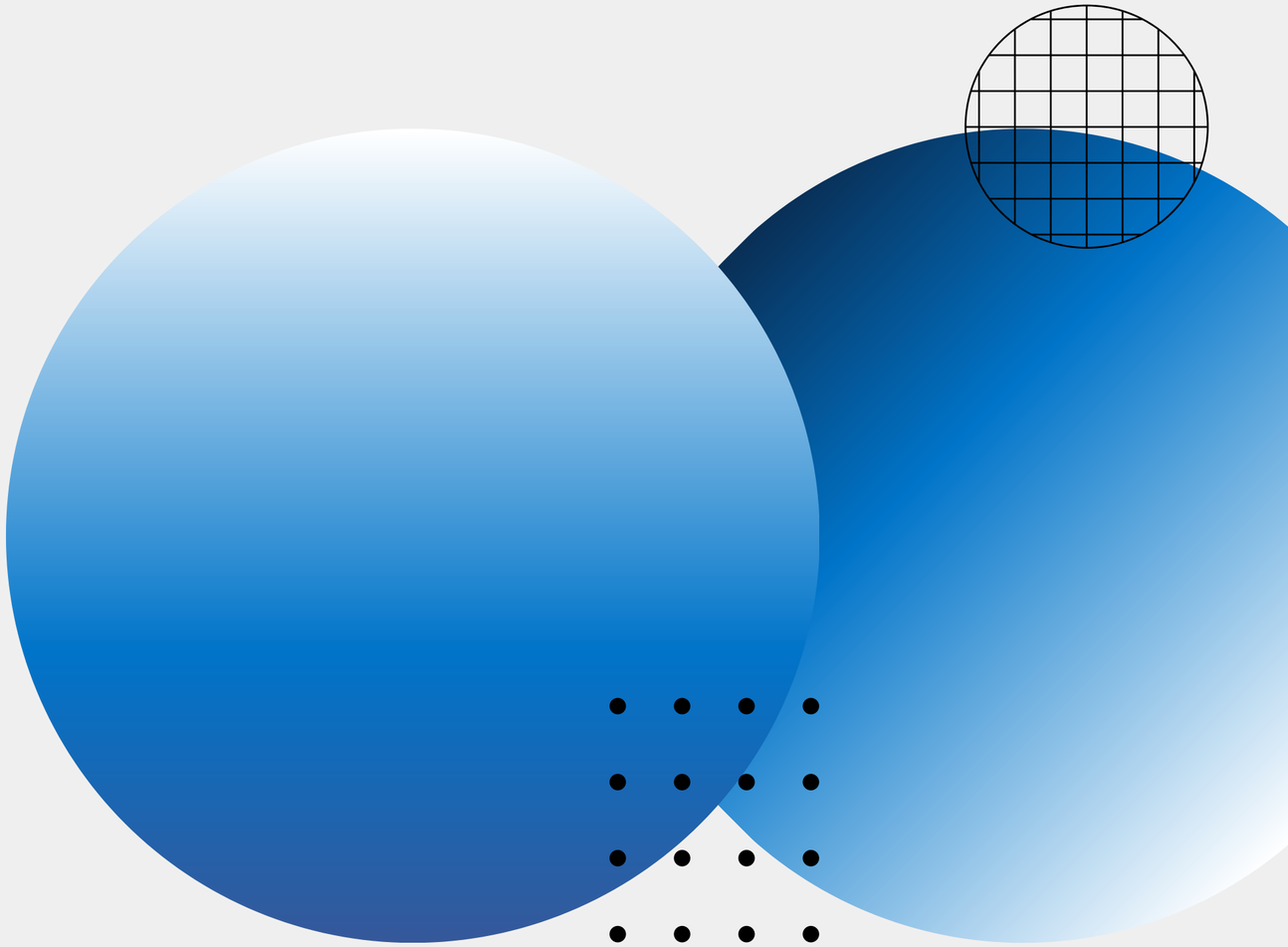
Además, destacamos la trascendencia del preprocesamiento de datos. Mediante la limpieza y análisis detallado, pudimos identificar características clave y manejar outliers notables. Este proceso no solo mejoró la calidad de los datos, sino que también aumentó la precisión de nuestros modelos.

Durante el análisis, también aprendimos sobre la necesidad de codificar correctamente las variables. Descubrimos que algunas columnas, aparentemente numéricas, en realidad eran de tipo objeto, lo que requirió un análisis más profundo para asegurar la coherencia y precisión de los datos.

El análisis gráfico y numérico fue esencial para entender las relaciones entre las variables. Las visualizaciones y estadísticas nos proporcionaron insights claros y profundos, facilitando la interpretación de los datos y el desarrollo de estrategias basadas en resultados concretos.

En resumen, este proyecto fue una oportunidad invaluable para aplicar nuestros conocimientos teóricos en un entorno práctico. Nos enseñó que el preprocesamiento meticuloso y la correcta codificación de variables son fundamentales para obtener resultados precisos y significativos en cualquier estudio de datos.

¡Muchas Gracias!



Referencias:

INEGI. (2024) Quiénes somos By Nacional de Container: Inegi.org.mx URL: https://www.inegi.org.mx/inegi/quienes_somos.html

INEGI. (2024). Instituto Nacional de Estadística y Geografía (INEGI). Inegi.org.mx. https://www.inegi.org.mx/