Juan David Bahamon - A00375826
Samuel Hernandez - A00375392
David Peñaranda - A00375827

Introducción

La inteligencia artificial (IA) ha revolucionado la forma en que las empresas analizan y extraen conocimientos de los datos. Este documento presenta dos casos prácticos de aplicación de técnicas de aprendizaje automático en diferentes contextos.

El primer caso aborda el análisis de riesgo crediticio, donde se desarrolla un modelo predictivo para determinar si se aprobará o denegará un préstamo a un solicitante, basado en sus características personales, ingresos, historial crediticio y detalles del préstamo.

El segundo caso se enfoca en los supermercados, utilizando la IA para predecir el nivel de gasto esperado de los clientes. Al analizar características de los clientes se pueden desarrollar modelos que permitan a los supermercados anticipar las necesidades de sus clientes y optimizar sus estrategias.

En ambos casos, se siguió un proceso riguroso de limpieza de datos, validación cruzada y evaluación de diferentes algoritmos de aprendizaje automático, como árboles de decisión, bosques aleatorios y modelos de boosting. Estos ejemplos demuestran el potencial de la IA para abordar desafíos prácticos y resaltan la importancia de integrar estas técnicas en los procesos de toma de decisiones empresariales.

Consideraciones Éticas

Se consideraron los siguientes aspectos para ambas partes del laboratorio:

Privacidad de los Datos. En el caso del primer dataset los datos no contienen nombres, apellidos, documentos ni otra forma de identificador directo de las personas asociadas a los préstamos. En el caso del segundo, únicamente se compromete la identificación de los usuarios, pero como esta no representa una característica relevante a la hora de realizar las predicciones requeridas se puede prescindir de esta y los datos quedan aislados de las personas que los generaron. Así, se minimiza significativamente el riesgo de reidentificación. Además, los datos fueron usados con el único propósito de estudiarlos y no fueron compartidos con terceros no autorizados.

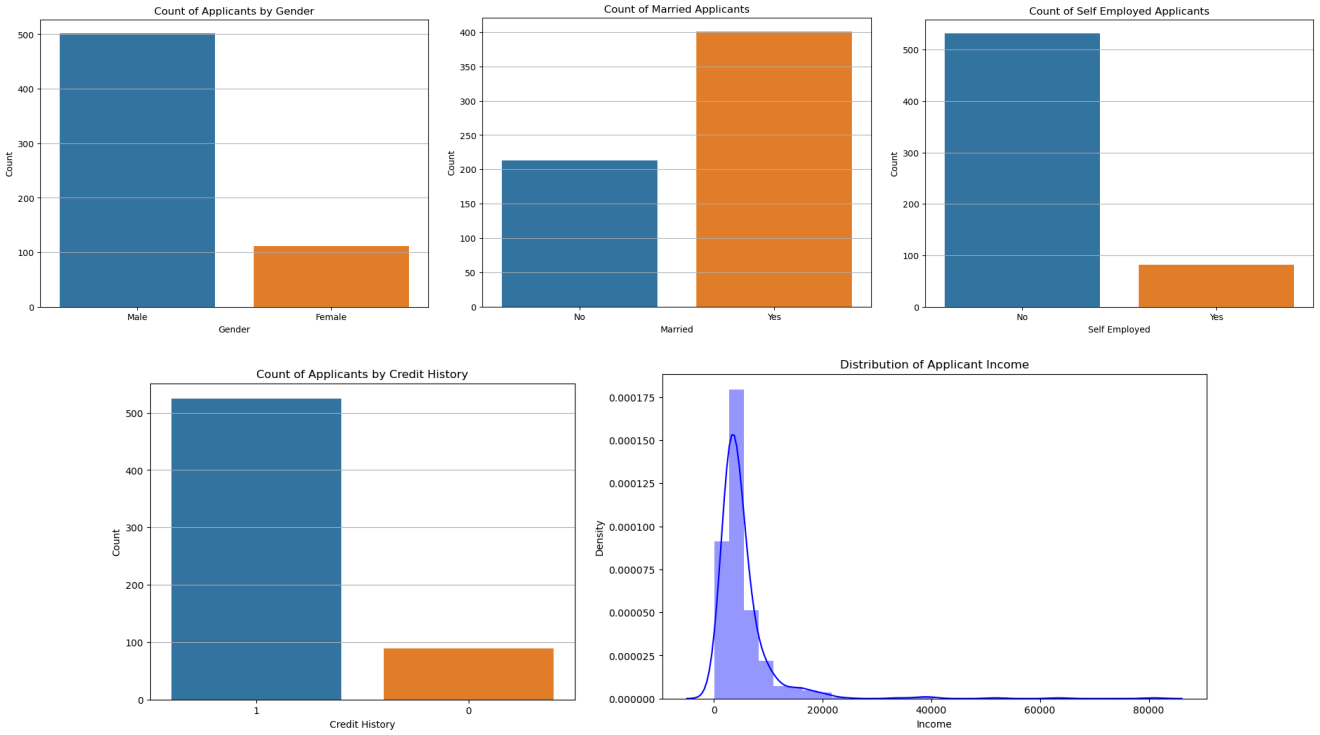
Equidad y No Discriminación. Se prestó especial atención a los posibles sesgos. Evitar los sesgos es crucial porque en este contexto el modelo no debería rechazar sistemáticamente a grupos determinados de personas que cumplen ciertas características por el mero hecho de tenerlas. En la evaluación de los modelos se puede observar que no existe discriminación de ninguna forma. Finalmente, se intentó garantizar la representación de los diferentes grupos de personas en los datos de entrenamiento.

Responsabilidad. En general, se documentó claramente los métodos y modelos utilizados en el estudio para que otros puedan entender cómo se tratan e interpretan los datos y cómo se obtuvieron los resultados expuestos. De este modo, se debe asumir que el modelo no es infalible y cualquier predicción errónea puede atribuirse únicamente a las limitaciones naturales en toda forma de inteligencia artificial y no a la manipulación maliciosa de los datos.

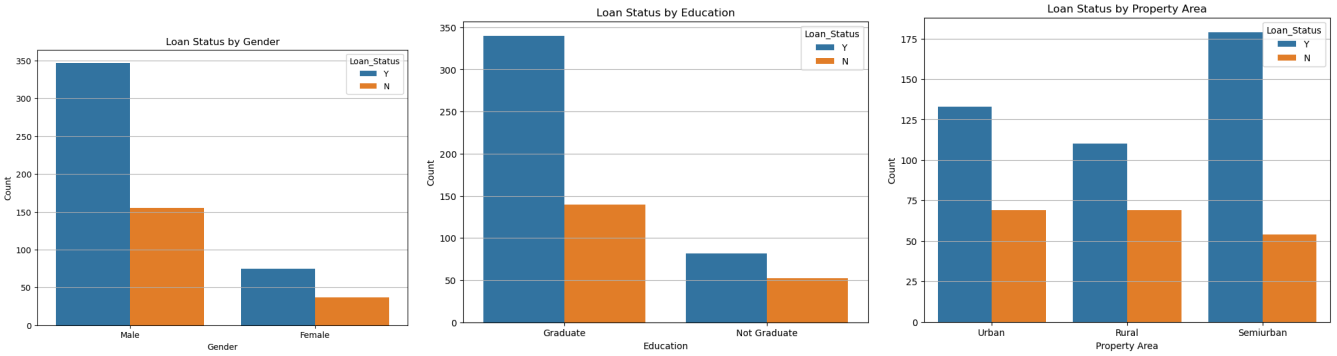
Laboratorio Parte 1

Identificación y Formulación del Problema

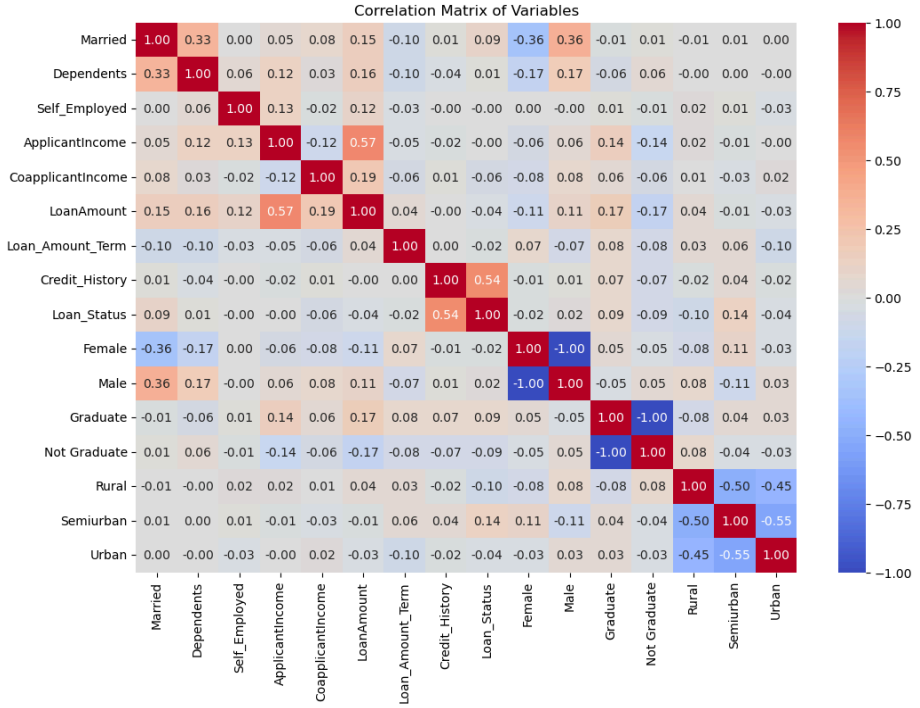
El problema se centra en la predicción de la aprobación de préstamos. El objetivo es construir un modelo que pueda predecir si un préstamo será aprobado o no, basándose en información previa de solicitudes que han sido aprobadas o rechazadas y datos del solicitante como el género, el estado civil, la educación, el número de dependientes, el ingreso, el monto del préstamo, el plazo del préstamo y el historial crediticio. La precisión de este modelo es crucial, ya que puede ayudar a tomar decisiones informadas y libres de sesgos.



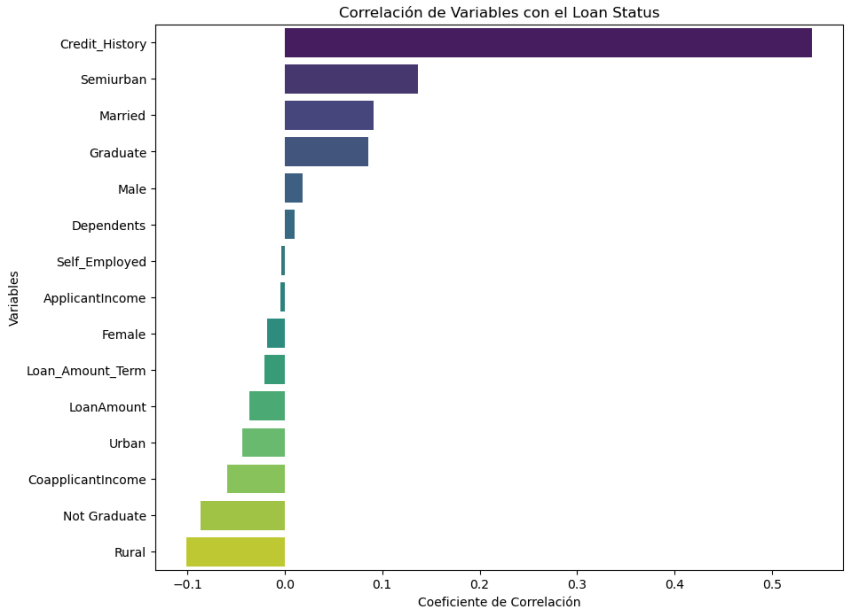
Como parte del análisis univariado se observa que las personas casadas, empleadas (no auto-empleadas) hombres y con historial crediticio conforman los grupos con mayor representación. Es importante revisar al final que estos desbalances no sean significativos en el entrenamiento del modelo. Esta misma observación aplica para el nivel de ingresos de la persona, pues como es natural existen niveles que no tienen tanta representación en el dataset.



Confrontando las variables se puede determinar que el porcentaje de hombres y mujeres cuyo préstamo no fue aceptado es similar dentro de su grupo de género. En cambio, el porcentaje de graduados que no recibieron el préstamo es inferior al de no graduados dentro de sus respectivos grupos según nivel de educación. Finalmente se observa cómo los grupos según área de la propiedad generan diferentes distribuciones: El área urbana y rural recibieron cantidades moderadas de rechazos respecto a las aprobaciones y el área semiurbana presentó más bien pocos rechazos en comparación con la cantidad de créditos aprobados.

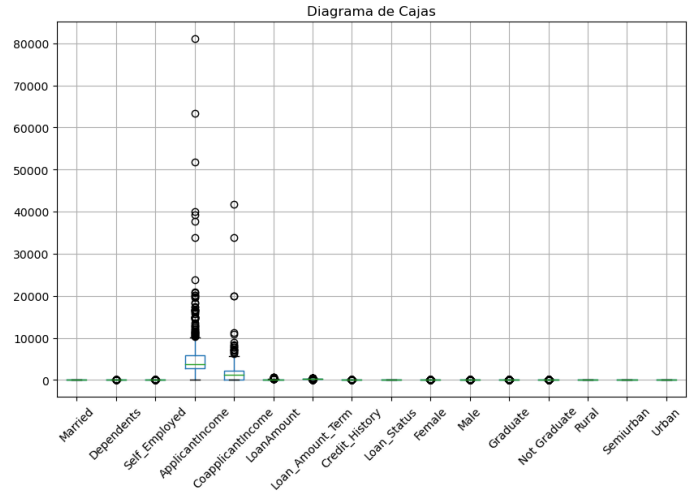


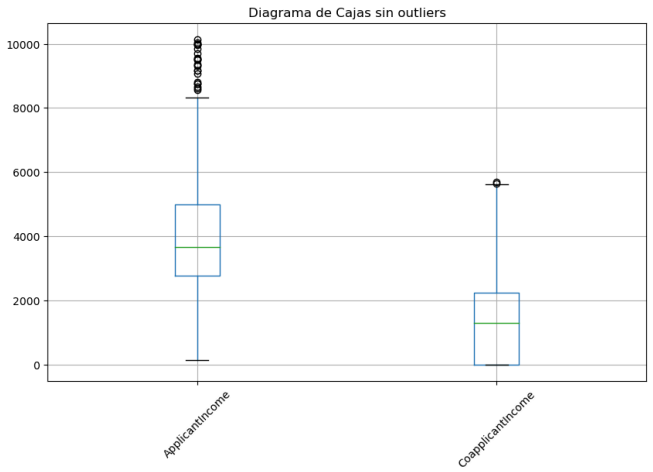
En primer lugar se ignora la diagonal principal y algunas correlaciones cuyo comportamiento obedece a clases mutuamente excluyentes (ej: no estar graduado y estar graduado, vivir en área rural, semiurbana o urbana). Se puede destacar dos relaciones que son evidentes en el contexto del problema: La aprobación del préstamo se ve influenciada por el historial crediticio y la cantidad solicitada está relacionada con el nivel de ingresos (ambas directamente proporcionales). Sin embargo, estas relaciones no son precisamente fuertes.



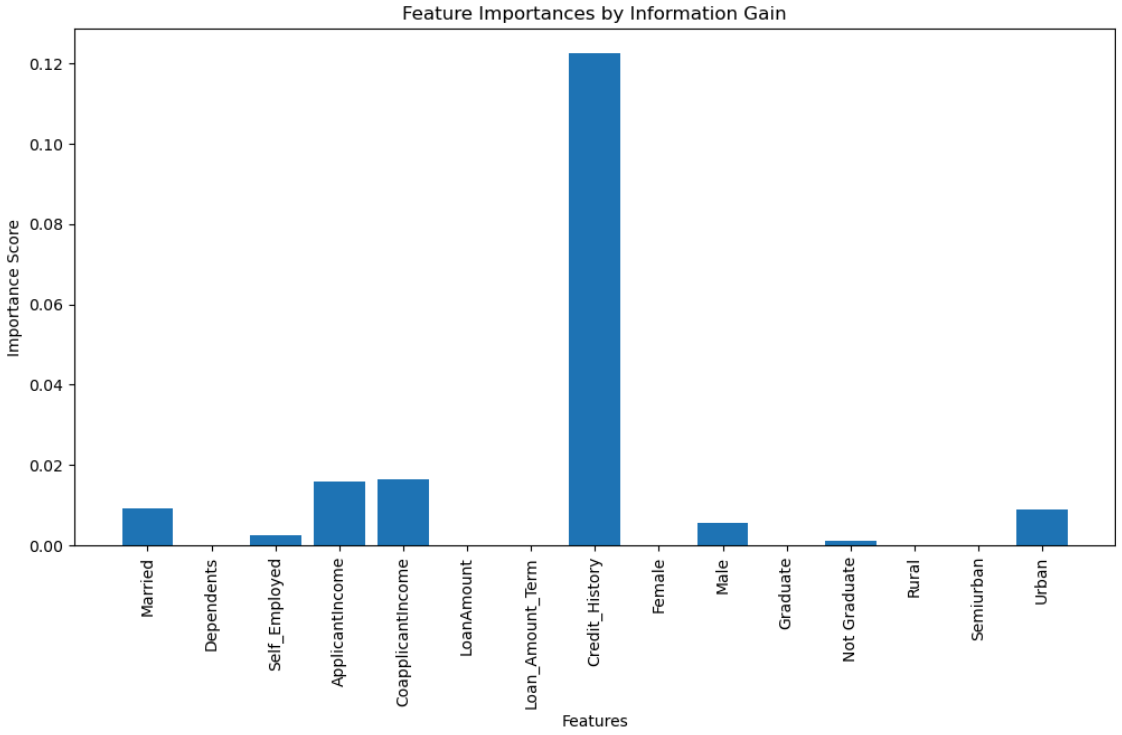
Mediante esta gráfica se puede aislar la relación de la variable objetivo (estado del préstamo) con las demás y se reitera la influencia del historial crediticio, el área semiurbana, el estado marital casado, y el nivel de educación graduado. Estos últimos se observaron inicialmente en los análisis de dos variables.

A continuación se observa el proceso de eliminación de datos atípicos, que es clave para garantizar una mejor toma de decisiones en el modelo.





Se observa que existen valores outliers especialmente en los ingresos del aplicante y del co-aplicante. A la derecha se observa los datos una vez eliminados aquellos outliers que se podían eliminar. Una vez que los datos pasaron por estos procesos se empieza a considerar las características más relevantes para elaborar modelos.



Finalmente, se probaron diferentes opciones de modelo y se realizó validación cruzada mediante el método de k-folds. A continuación, se resume los resultados de los mejores hiperparámetros de los modelos:

Best parameters for DecisionTree: {'max_depth': 3, 'min_samples_split': 10}
Best parameters for RandomForest: {'max_depth': 3, 'n_estimators': 10}
Best parameters for KNN: {'leaf_size': 20, 'n_neighbors': 7}
Best parameters for SVC: {'C': 0.1, 'kernel': 'linear'}
Best parameters for XGBoost: {'max_depth': 3, 'n_estimators': 50}

A continuación se resume el desempeño de cada modelo:

DecisionTree:

Tiene un buen balance entre precisión y recall para ambas clases.
La precisión para la clase False es alta (0.86), pero el recall es bajo (0.51), lo que indica que tiene dificultades para identificar correctamente todos los casos False.

Para la clase True, el recall es muy alto (0.96), lo que significa que identifica muy bien los casos True, pero la precisión es un poco más baja (0.81).

RandomForest:

Tiene un rendimiento ligeramente mejor que DecisionTree.
Logra una precisión excelente (0.94) para la clase False, pero el recall es bajo (0.49), similar al DecisionTree.
Para la clase True, el recall es casi perfecto (0.99) pero la precisión es un poco menor (0.80).

KNN:

Es el modelo con peor rendimiento general.
Tiene serios problemas para identificar la clase False, con precisión (0.08) y recall (0.03) extremadamente bajos.
Para la clase True, el recall es bueno (0.85) pero la precisión es baja (0.65).

SVC:

Logra una precisión perfecta (1.0) para la clase False, pero el recall es bajo (0.43).
Para la clase True, el recall es perfecto (1.0) pero la precisión es menor (0.79).
Es uno de los modelos con mayor desequilibrio entre precisión y recall.

XGBoost:

Tiene un rendimiento similar al DecisionTree.
Precisión y recall balanceados para ambas clases, sin valores extremos.

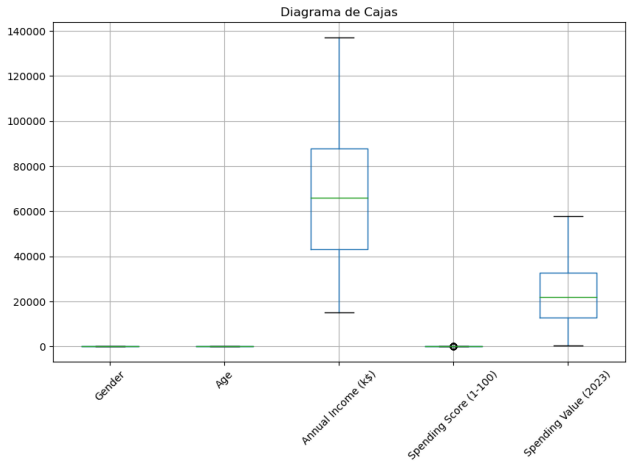
Elección del modelo

En caso de tener que elegir uno de los modelos se optaría por RandomForest y XGBoost. Esto porque parecen ser los modelos más equilibrados y con mejor rendimiento general. Respecto al DecisionTree y SVC, estos tienen problemas con al menos una de las clases. Finalmente KNN es el que peor se desempeña en esta tarea y puede ser descartado.

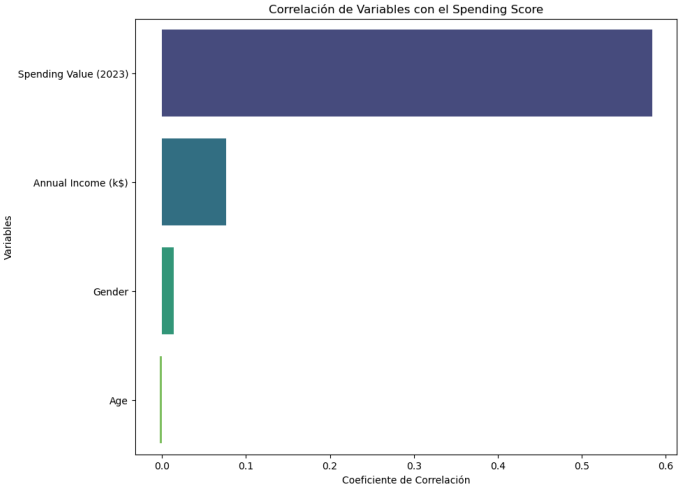
En un conjunto de datos con variables mixtas como este, que incluye características numéricas como ingresos y monto del préstamo, y características categóricas como género, estado civil, educación, etc., los modelos RandomForest y XGBoost tienen la flexibilidad de manejar bien estos datos heterogéneos. Además, como el conjunto de entrenamiento no es excesivamente grande, estos dos algoritmos relativamente complejos pueden ajustarse bien sin el riesgo de sobreajuste excesivo, aprovechando su capacidad para aprender patrones complicados.

Laboratorio Parte 2

Interpretación de las gráficas generadas en la fase de análisis

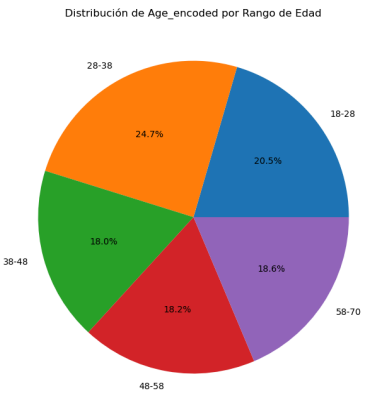


Se empleó un diagrama de caja para detectar outliers. A pesar de que no se observaron outliers, se implementó el código necesario para eliminarlos preventivamente en campos numéricos.
Los valores de género, edad, ingreso anual y puntuación de gasto no tenían valores atípicos y por tanto no se modificaron.

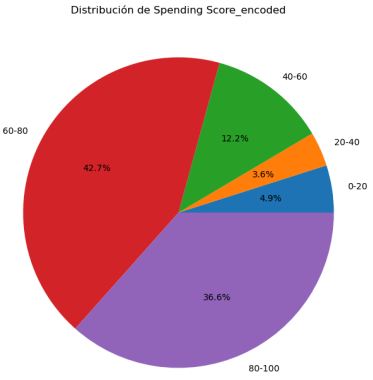


Además se realizó un análisis de correlación. Mediante este se determinó que annual income y spending score tienen una correlación insignificante (0.08). Annual income y spending value muestran una correlación alta y positiva de 0.82. Y La correlación entre spending score y spending value es moderada (0.58). Esto es, que los clientes con mayores ingresos y mayor puntuación de gasto tienden a gastar más en la tienda.

Después, se generó un gráfico de correlación entre cada una de las variables y el Spending Score (1-100). En la observación de este gráfico, se destaca que la variable Spending Value (2023) tiene la correlación más significativa con el Spending Score, sugiriendo así que el valor de gasto es un predictor más confiable del comportamiento de gasto de los clientes. Por otra parte, aunque la correlación con Annual Income (k$) es positiva, también indica que los ingresos anuales inciden en el Spending Score, aunque en menor grado que el valor de gasto.

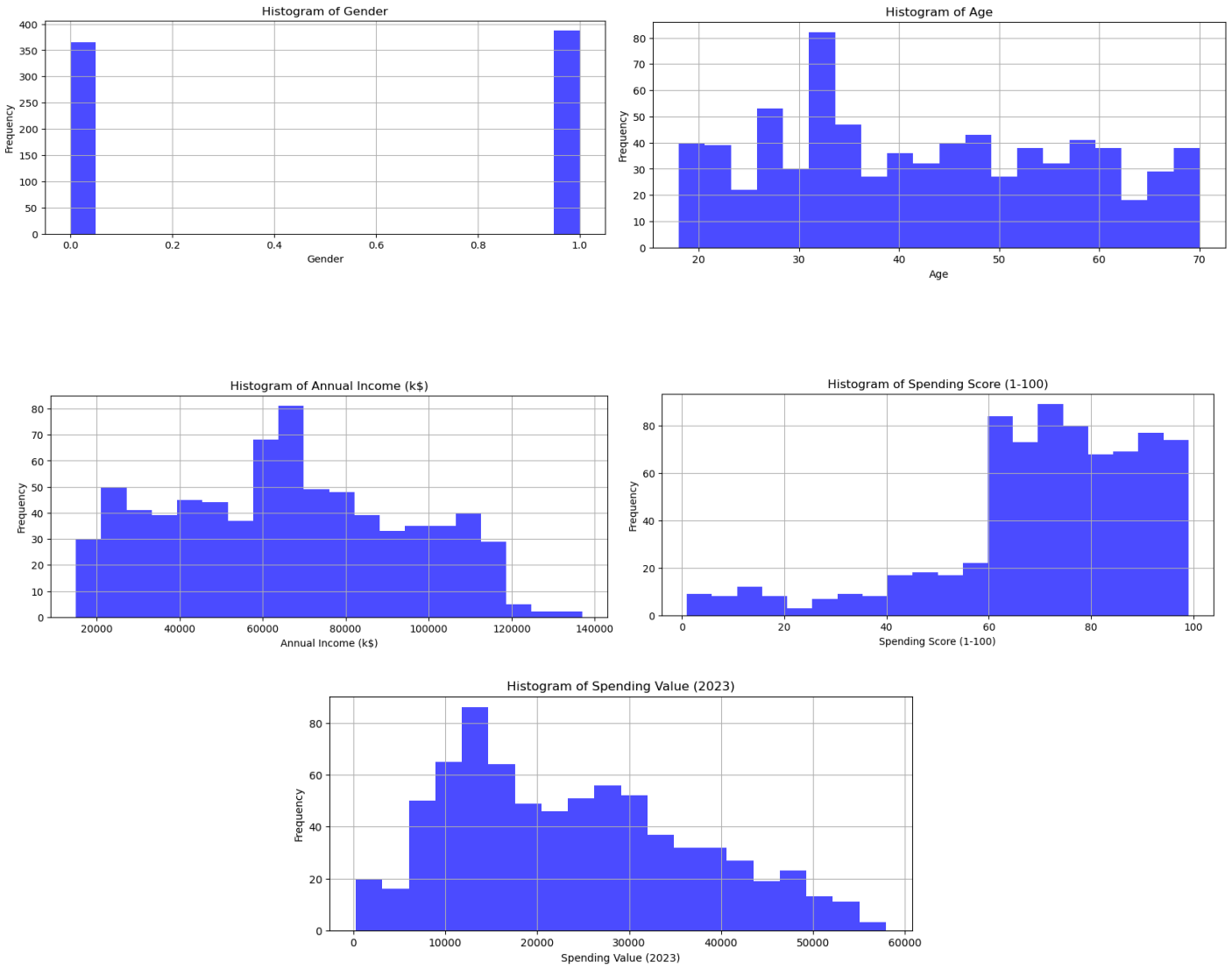


La gráfica de pastel proporciona una visión de la distribución de la variable "Age_encoded" por rangos de edad. El grupo de 28-38 años es el más numeroso, representando el 24.7% de la población. Le sigue el grupo de 58-70 años, con el 18.6% de la población. Los grupos de edad restantes se distribuyen de la siguiente manera: 18-28 años (20.5%), 38-48 años (18.0%), y 48-58 años (18.2%). Esta distribución nos permite identificar los grupos de edad predominantes.



La segunda gráfica circular presenta la distribución del "Spending Score_encoded" en diversos rangos. El rango de 60-80 destaca con la mayor proporción, representando el 42.7% y sugiriendo que una gran parte de los clientes posee una puntuación de gasto alta. Le sigue el grupo de 80-100, con el 36.6% de la distribución, indicando que muchos clientes muestran un comportamiento de gasto muy activo. Además, el segmento de 40-60 representa el 12.2% de la distribución, mostrando una cantidad moderada de

clientes con un gasto intermedio. Los segmentos de 0-20 y 20-40 son los más pequeños, con un 4.9% y 3.6% respectivamente, lo que sugiere que hay pocos clientes en estos rangos de puntuación. Esta información puede ser valiosa para identificar segmentos de clientes según su comportamiento de gasto.



El "Histograma de Género" muestra una distribución equitativa de frecuencias para las categorías "mujer" y "hombre". El histograma de edad tiene un pico en el rango de 30 a 35 años, indicando que este grupo de edad es el más común en el conjunto de datos. En cuanto al histograma del ingreso anual, se observa un pico significativo en el rango de 60,000 a 80,000 dólares, sugiriendo que la mayoría de las personas tienen ingresos dentro de este intervalo. El histograma del puntaje de gasto (1-100) muestra frecuencias más altas en el rango de 60 a 100, indicando que la mayoría de las personas tienen puntajes de gasto en esta escala. Finalmente, el histograma del valor del gasto (2023) muestra una concentración de frecuencias en el rango de 10,000 a 20,000, lo que sugiere que la mayoría de las personas tienen un valor de gasto dentro de este intervalo.

Justificación de elección de algoritmos para cada caso:

Caso #1:
Se optó por utilizar la regresión en este caso porque nuestro objetivo es predecir un valor numérico específico, en este caso, el valor de gasto de los clientes. La regresión es el enfoque más apropiado cuando se busca modelar la relación entre variables independientes y una variable dependiente continua. En este escenario, nuestras características predictoras incluyen datos como género, ingreso anual, puntuación de gasto y edad, que se espera

que influyan en el valor de gasto de un cliente. Por lo tanto, la regresión se considera la técnica más adecuada para modelar esta relación y proporcionar predicciones numéricas sobre el gasto de los clientes.

Caso #2:
Para este caso específico, donde nuestro objetivo es predecir el puntaje de gasto de nuevos clientes basados en características similares a los clientes existentes, se optó por utilizar un algoritmo de clasificación (KNN). Dado que la variable objetivo, el puntaje de gasto, ha sido codificada en grupos discretos, convirtiéndo así el problema en uno de clasificación. La tarea consiste en asignar a cada nuevo cliente a uno de los cinco grupos predefinidos de acuerdo con sus características. Dado que el objetivo es predecir la categoría a la que pertenece un cliente en función de sus características, la clasificación es la opción adecuada.

Caso #3:
Para este caso donde deseamos predecir el género de los nuevos clientes para tomar decisiones informadas sobre qué tipo de productos de cuidado personal ofrecer, se optó por utilizar el modelo de Máquinas de Vectores de Soporte (SVM). Este modelo es adecuado para encontrar un hiperplano que pueda separar efectivamente los datos en diferentes grupos basados en características específicas, como edad, ingresos y comportamiento de compra. Al usar SVM, queremos identificar patrones en el comportamiento de compra que nos ayuden a determinar si debemos enfocarnos en productos diseñados para un género específico o si deberíamos ofrecer una variedad equilibrada para ambos géneros. Además, se utiliza GridSearchCV para encontrar los mejores hiperparámetros del modelo, ayudando a aumentar la precisión de las predicciones. Al tener como objetivo predecir el género de los clientes, que es una variable categórica, este problema se clasifica como un problema de clasificación.

Caso #4
Para abordar la relación entre la edad de los clientes y su comportamiento de compra, se optó por utilizar el modelo de Árbol de Decisión (Decision Tree Classifier). Este modelo permite clasificar a los clientes en grupos basados en su edad y analizar cómo estas segmentaciones se correlacionan con diferentes patrones de compra. La capacidad de los árboles de decisión para estructurar los datos de manera jerárquica facilita la comprensión de las relaciones entre las variables, lo que los hace ideales para este análisis. Dado que el objetivo es clasificar a los clientes en grupos en función de su edad, este problema se clasifica como un problema de clasificación.

Explicación de las métricas de evaluación arrojadas para cada algoritmo y qué combinación de parámetros o pasos hiciste para obtener el mejor modelo posible.

En el caso 1, se evaluaron dos modelos de regresión lineal para predecir el valor de gasto de los clientes. El primer modelo, una regresión lineal simple que solo consideraba los ingresos anuales como variable predictora, mostró un rendimiento moderado tanto en el conjunto de entrenamiento como en el de prueba, con un MSE de aproximadamente 56,293,744 y 55,258,636, respectivamente. Por otro lado, el modelo de regresión lineal múltiple, que incluía género, ingresos anuales, puntuación de gasto y edad como características predictoras, mostró un rendimiento superior, con un MSE de aproximadamente 12,593,776 y 13,017,617 en los conjuntos de entrenamiento y prueba, respectivamente. La elección del modelo múltiple se basó en su capacidad para explicar una mayor varianza en los datos y su mejor rendimiento en el conjunto de prueba, lo que sugiere una mejor generalización a nuevos datos. Además, este modelo tiene un R2 de aproximadamente 0.926, lo que indica que el modelo explica alrededor del 92.6% de la varianza en los datos de entrenamiento.

En el caso 2, se evaluaron dos modelos: un clasificador Dummy y un clasificador KNeighbors (KNN). El clasificador Dummy, que predice siempre la clase más frecuente en el conjunto de entrenamiento, sirvió como línea base con una precisión del 37.1%. Por otro lado, el clasificador KNN ajustado con k=5 vecinos más cercanos obtuvo una precisión del 59% en el conjunto de entrenamiento, indicando un buen rendimiento en los datos de entrenamiento.

En el caso 3, se exploraron dos enfoques para clasificar el género de los clientes: un clasificador Dummy y un clasificador SVM (Support Vector Machine). El clasificador Dummy, proporcionó una precisión de referencia del 49.0%. Por otro lado, el clasificador SVM básico con kernel rbf y gamma scale logró una precisión del 58.6% en el conjunto de entrenamiento y del 49.7% en el conjunto de prueba. Luego, se realizó una búsqueda de hiperparámetros utilizando GridSearchCV, que seleccionó el mejor modelo SVM con C=10, gamma=1 y kernel rbf.

Este modelo mejorado mostró una precisión del 67.0% en el conjunto de entrenamiento y del 56.0% en el conjunto de prueba.

En el caso 4, se exploró la clasificación de clientes en grupos basados en su edad utilizando un clasificador de árbol de decisión. Primero, se entrenó un clasificador Dummy proporcionando una precisión de referencia del 17.2%. Luego, se utilizó un árbol de decisión básico sin ajuste de hiperparámetros, lo que resultó en una precisión del 72.5% en el conjunto de entrenamiento y del 21.9% en el conjunto de prueba. Posteriormente, se realizó una búsqueda de hiperparámetros utilizando GridSearchCV, que seleccionó el mejor modelo con criterion='entropy', max_depth=3 y min_samples_split=2. Este modelo mejorado mostró una precisión del 30.0% en el conjunto de entrenamiento y del 20.0% en el conjunto de prueba. A pesar de los esfuerzos de ajuste de hiperparámetros, la precisión del modelo en el conjunto de prueba no mejoró significativamente.

¿Mejoraron el/los modelos del caso #1 con la aplicación del k-fold cross validation o no hubo una diferencia significativa?

No hubo una diferencia significativa en la mejora del modelo con la aplicación del k-fold cross-validation. En el primer modelo de regresión lineal simple, la puntuación R2 para el conjunto de prueba fue de aproximadamente 0.685 antes de la validación cruzada y 0.664 después, lo que indica una leve disminución en la capacidad del modelo para explicar la varianza en los datos de prueba. Del mismo modo, para el modelo de regresión lineal múltiple, la puntuación R2 para el conjunto de prueba fue de aproximadamente 0.926 antes de la validación cruzada y 0.924 después, lo que indica una mejora mínima en la capacidad predictiva del modelo.

Valor de baseline

Calcular el valor de baseline es fundamental en el desarrollo de modelos de aprendizaje automático, ya que proporciona una referencia simple para evaluar el rendimiento de modelos más complejos. Es decir que, el valor de baseline va a representar el desempeño mínimo que se espera lograr sin la utilización de un modelo avanzado. Es necesario conocer este valor antes de implementar un modelo más sofisticado, porque ayuda a establecer expectativas realistas sobre el rendimiento del modelo y a evaluar si realmente está aportando valor. En cuanto a la precisión de los modelos KNN, SVM y decision tree, como se ha mencionado antes la precisión de los modelos es considerablemente mejor que la del baseline, lo que nos indica que el modelo está aprendiendo patrones útiles en los datos y está realizando predicciones significativas.

Impacto del GridSearchCV en los algoritmos KNN, SVM y decision tree

El GridSearchCV tuvo un impacto positivo en el rendimiento del algoritmo KNN y SVM al encontrar combinaciones óptimas de hiperparámetros, mejorando significativamente su precisión tanto en el conjunto de entrenamiento como en el de prueba. Sin embargo, para el Decision Tree, a pesar de los esfuerzos de ajuste de hiperparámetros, no se observó una mejora significativa en la precisión del conjunto de prueba