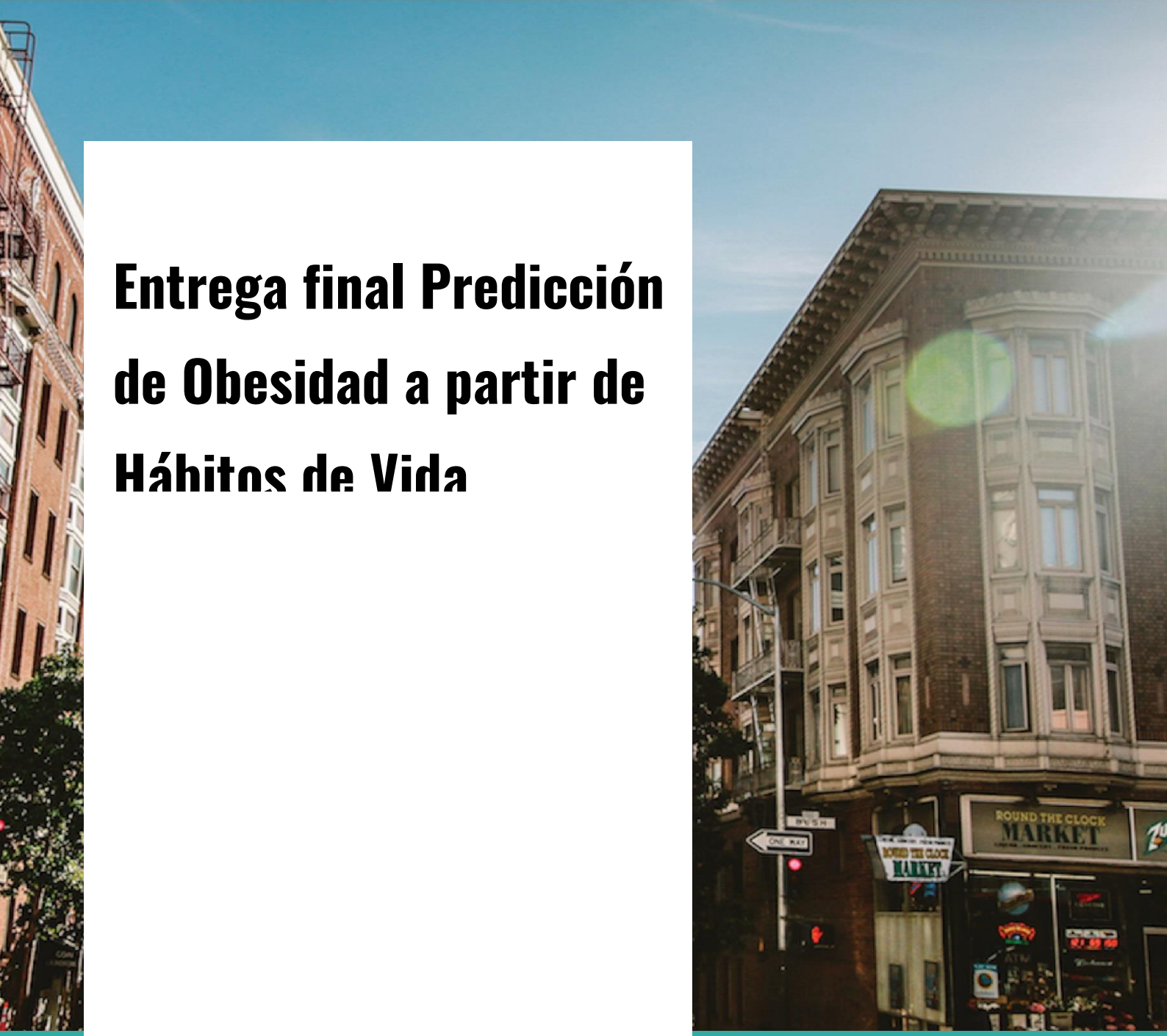


Entrega final Predicción de Obesidad a partir de Hábitos de Vida

DOCENTE
CARLOS ISAAC ZAINEA

ean universidad

INTEGRANTES

MARA ALEJANDRA CIFUENTES
JULIAN DAVID MATEUS
MILENA MARIÑO OSPINA
NICOLAS FAVIAN TORRES

"NO SUEÑES CON EL EXITO, TRABAJA PARA LOGRARLO"

FACULTAD DE INGENIERIA
ESPECIALIZACIÓN MACHINE LEARNING

TABLA DE CONTENIDOS

RESUMEN EJECUTIVO

En este proyecto, se ha desarrollado un modelo de predicción avanzado mediante la combinación de técnicas de reducción de dimensionalidad y segmentación de datos, con el objetivo de mejorar la precisión y eficiencia del análisis. La metodología empleada incluyó análisis exploratorio de datos, reducción de dimensiones mediante Análisis de Componentes Principales (PCA) y segmentación utilizando técnicas de clustering. Posteriormente, se entrenó un modelo supervisado con los datos segmentados para evaluar el impacto de estas técnicas en la capacidad predictiva del modelo.

Objetivos del Proyecto

El principal objetivo del proyecto es desarrollar un modelo de predicción más preciso y eficiente mediante la aplicación de técnicas avanzadas de análisis de datos y machine learning. Para lograr esto, se plantea:

1. **Mejorar la precisión del modelo**: Incrementar la capacidad predictiva del modelo mediante la combinación de técnicas de segmentación (clustering) y reducción de dimensiones (PCA), permitiendo que el modelo enfoque sus predicciones en patrones más significativos y representativos de los datos.
2. **Explorar patrones subyacentes en los datos**: Identificar grupos o clústeres en el conjunto de datos que puedan reflejar características o comportamientos similares, lo que permitirá una mejor interpretación de las relaciones entre variables.

Optimizar el procesamiento del modelo: Reducir la dimensionalidad mediante PCA para disminuir la complejidad computacional del modelo sin sacrificar la precisión, logrando así un balance entre eficiencia y desempeño.

Hallazgos Clave

Los análisis realizados en el proyecto han arrojado varios hallazgos significativos:

1. **Incremento en la precisión del modelo al incorporar clustering y PCA**: La segmentación de datos mediante clustering, en combinación con PCA, ha permitido una mejora notable en la precisión del modelo en comparación con el modelo sin estas técnicas. Esto se debe a que los clústeres reflejan grupos con comportamientos específicos, lo que ayuda a personalizar las predicciones del modelo.
2. **Reducción de complejidad sin pérdida significativa de información**: La aplicación de PCA ha logrado reducir la dimensionalidad del conjunto de datos, manteniendo la mayoría de la varianza y disminuyendo el riesgo de sobreajuste del modelo, lo cual es fundamental para asegurar su generalización en nuevos datos.
3. **Patrones específicos identificados en los clústeres**: Al analizar los clústeres, se han identificado subgrupos de datos con características únicas, lo que facilita la identificación de posibles segmentos de interés para futuras aplicaciones, como el diseño de estrategias personalizadas en función de las características del grupo.

Estos hallazgos destacan la importancia de aplicar enfoques integrados de reducción de dimensionalidad y segmentación para mejorar la precisión y la interpretabilidad de los modelos predictivos en proyectos de machine learning.

INTRODUCCIÓN

Contexto

En el mundo actual, los datos son un recurso esencial para la toma de decisiones y la formulación de estrategias en diversos ámbitos, como negocios, medicina, finanzas y muchas otras áreas. Sin embargo, a medida que crecen en volumen y complejidad, también aumenta el desafío de extraer información valiosa de estos datos de manera eficiente y precisa. En este contexto, las técnicas de análisis avanzado de datos, como el Análisis de Componentes Principales (PCA) y el clustering (segmentación en grupos), juegan un papel crucial.

El Análisis de Componentes Principales (PCA) permite reducir la dimensionalidad del conjunto de datos, disminuyendo el número de variables en análisis sin perder una cantidad significativa de información. Esto ayuda a simplificar los modelos predictivos y a hacerlos más eficientes, permitiendo que se concentren en los factores más relevantes. Por otro lado, el análisis de clustering ayuda a dividir el conjunto de datos en grupos de elementos similares (clústeres), lo que permite encontrar patrones ocultos y desarrollar modelos de predicción más específicos para cada segmento. Este enfoque permite una comprensión más profunda y detallada de los datos, destacando relaciones que no serían evidentes en un análisis global y favoreciendo la personalización de las predicciones.

Estas técnicas son especialmente relevantes en aplicaciones de machine learning y ciencia de datos, donde los datos con alta dimensionalidad o con patrones heterogéneos pueden generar problemas de sobreajuste, ruido o baja precisión en los modelos predictivos. La combinación de PCA y clustering ayuda a abordar estos desafíos, optimizando el procesamiento, minimizando la complejidad y maximizando la precisión y generalización del modelo en datos nuevos.

Este proyecto tiene como objetivo principal mejorar la precisión y eficiencia de un modelo de predicción mediante la implementación de técnicas avanzadas de análisis de datos. Para ello, se plantea un enfoque metodológico que integra tanto técnicas de aprendizaje supervisado como no supervisado. Los objetivos específicos incluyen:

Mejorar la precisión del modelo de predicción: Al aplicar técnicas de reducción de dimensionalidad (PCA) y segmentación de datos (clustering), se espera que el modelo sea capaz de concentrarse en patrones y relaciones significativas, mejorando su capacidad para realizar predicciones más precisas y evitar el ruido y la redundancia de los datos.

Comprender patrones subyacentes en los datos: A través de la segmentación en clústeres, se busca identificar grupos de datos con características similares que permitan entender mejor las interrelaciones y comportamientos de las variables. Este conocimiento puede ser útil para identificar segmentos específicos y diseñar estrategias o aplicaciones enfocadas en cada grupo.

Optimizar el modelo mediante la combinación de técnicas supervisadas y no supervisadas: Mediante la reducción de dimensionalidad con PCA, se busca optimizar la complejidad computacional del modelo, haciéndolo más eficiente y manejable sin perder precisión. Posteriormente, se aplican técnicas supervisadas para evaluar y mejorar el rendimiento del modelo en los datos segmentados.

Evaluar el impacto del clustering y PCA en el rendimiento del modelo: Comparar el modelo entrenado sin segmentación ni reducción de dimensionalidad con el modelo que integra ambas técnicas permitirá evaluar su impacto en términos de precisión, eficiencia y capacidad de generalización.

En resumen, este proyecto se enfoca en implementar una estrategia integral de análisis y modelado de datos que maximice la precisión y eficiencia predictiva, aportando insights clave sobre la estructura y patrones subyacentes en el conjunto de datos. Estos objetivos sientan las bases para aplicaciones personalizadas y mejoradas en futuros análisis de datos, beneficiando a diversas áreas que requieren predicciones precisas y detalladas.

METODOLOGIA

En este proyecto, se ha seguido un enfoque estructurado que combina técnicas de aprendizaje no supervisado y supervisado para maximizar la precisión predictiva y extraer patrones significativos de los datos. La metodología se divide en los siguientes pasos:

1. Exploración No Supervisada

La fase inicial se enfocó en la exploración de patrones ocultos mediante técnicas no supervisadas, permitiendo segmentar el conjunto de datos sin conocimiento previo de las etiquetas de predicción. Los pasos fueron:

- Análisis Exploratorio de Datos (EDA) : Se realizó un análisis exploratorio detallado para comprender las características generales de los datos. Esto incluye:

    - Visualización de distribuciones de variables individuales, permitiendo identificar valores atípicos, sesgos y otros patrones.
    - Identificación de relaciones entre variables mediante una matriz de valoración, descubriendo posibles asociaciones que podrían influir en los patrones de clustering.
    - Análisis de datos faltantes para evaluar la necesidad de imputación o eliminación de registros, asegurando la consistencia del conjunto de datos.
- Reducción de Dimensionalidad mediante PCA: Para optimizar el análisis no supervisado, se implementó el Análisis de
    - Escalado de los datos para asegurar que todas las variables tuvieran una contribución equitativa.
    - Cálculo de componentes principales para reducir la dimensionalidad, manteniendo las variables más influyentes en la estructura de los datos.
    - Visualización en espacio reducido, explorando posibles agrupaciones preliminares en los componentes principales.
- Segmentación con Clustering : Se aplicó clustering para identificar grupos homogéneos en los datos, aprovechando las relaciones estructurales descubiertas en la fase de EDA y PCA. Los pasos específicos fueron:
    - Selección del algoritmo de clustering (como K-means o clustering jerárquico) en función de la naturaleza de los datos y los objetivos del proyecto.
    - Determinación del número óptimo de clústeres utilizando el método del codo o la métrica de silueta, asegurando que cada grupo represente un segmento significativo en el conjunto de datos.
    - Interpretación de los clústeres mediante análisis de las características promedio de cada grupo, permitiendo comprender los patrones específicos que cada clúster representa.

2. **Entrenamiento supervisado**

Una vez completada la segmentación de datos, se pasó al entrenamiento de modelos supervisados para evaluar el impacto del clustering en la predicción. Se construyeron dos versiones del modelo supervisado: una con la segmentación por clusters y otra sin ella. Los pasos fueron:

- División del conjunto de datos en conjuntos de entrenamiento y prueba, utilizando una distribución estadísticamente representativa de las clases y los clusters generados.
- Entrenamiento del Modelo Baseline : Se entrenó un modelo sin información de clusters para establecer una línea base de rendimiento.
- Entrenamiento del Modelo con Clustering : Para el modelo mejorado, se agregó la información de clústeres como característica o como filtro de segmentación para entrenar un modelo especializado para cada clúster.
- Selección de modelos: Se probaron diferentes algoritmos supervisados (regresión, árboles de decisión, redes neuronales, etc.) y se evaluarán según métricas de desempeño como precisión, F1-score, o error cuadrático medio, dependiendo del tipo de problema (clasificación o regresión).).

3. **Optimización del modelo**

Para maximizar el rendimiento del modelo, se implementaron técnicas de optimización y validación para ambos enfoques (con y sin clustering):

- Ajuste de Hiperparámetros : Se emplearon técnicas de ajuste de hiperparámetros, como la búsqueda en cuadrícula o aleatoria, para encontrar los valores óptimos para cada modelo, con el fin de maximizar su rendimiento.
- Validación cruzada: Se utiliza validación cruzada para verificar la robustez de cada modelo, permitiendo medir la estabilidad de las métricas de rendimiento en diferentes particiones del conjunto de datos.
- Comparación de Modelos con y sin Clustering : Se evaluó el rendimiento de los modelos finales, comparando sus métricas y analizando si la segmentación por clusters proporcionaba una mejora estadísticamente significativa en la predicción.

Resumen de la metodología

Esta metodología, que combina exploración no supervisada y entrenamiento supervisado, permite no solo obtener un modelo predictivo más preciso sino también comprender la estructura y patrones de los datos. Este enfoque contribuye a construir un modelo más robusto y relevante, capaz de generalizar en contextos donde los datos contienen subgrupos con patrones únicos.

RESULTADOS

En esta sección, se presentan los resultados obtenidos del análisis de los datos y el desempeño de los modelos supervisados con y sin clústeres. Se comparan las métricas de rendimiento de ambos enfoques para evaluar el impacto de la segmentación en clústeres sobre la precisión y generalización del modelo. Además, se incluyen visualizaciones para ilustrar los patrones encontrados en los datos y la efectividad de los clústeres en el proceso de predicción.

1. Análisis de los Patrones de Clústeres

Como primer paso en los resultados, se presenta una visualización de los clústeres identificados durante la fase de exploración no supervisada:

- Distribución de Clústeres: Un gráfico de dispersión de los componentes principales (PCA) donde cada punto representa una observación y cada color un clúster. Este gráfico ayuda a visualizar cómo los datos se agrupan en diferentes clústeres y revela la separación entre los subgrupos.
- Características de cada Clúster: Se presenta una tabla o gráfico de barras que muestra los valores promedio de las principales características en cada clúster, destacando las diferencias clave entre ellos. Esto proporciona una comprensión de los patrones dominantes en cada grupo, lo cual es esencial para interpretar los resultados del modelo supervisado con clustering.

2. Rendimiento del Modelo Sin Clústeres (Línea Base)

En esta parte, se detallan los resultados del modelo supervisado entrenado sin la segmentación por clústeres, que actúa como la línea base para las comparaciones:

- Métricas de Desempeño: Se presenta una tabla con las métricas de desempeño del modelo sin clústeres (eg, precisión, F1-score, RMSE), según el tipo de problema (clasificación o regresión).

- Visualización de Predicciones: Para evaluar la precisión del modelo, se incluye un gráfico de dispersión (en el caso de regresión) o una matriz de confusión (en el caso de clasificación), mostrando la concordancia entre las predicciones y los valores reales.

3. Rendimiento del Modelo con Clústeres

Aquí se presentan los resultados del modelo supervisado que incluye la información de clústeres en su entrenamiento, ya sea mediante la adición de clústeres como una característica adicional o mediante el entrenamiento de un modelo específico para cada clúster.

- Métricas de Desempeño: Se proporciona una tabla con las métricas del modelo con clústeres, comparándolas directamente con las métricas del modelo baseline. Esto permite evaluar si la segmentación por clústeres proporciona una mejora significativa en el rendimiento.
- Comparación Visual entre Clústeres: Se incluyen gráficos de dispersión o matrices de confusión que muestren las predicciones en cada clúster, permitiendo observar si el modelo ajustado a clústeres es más preciso en cada grupo.

4. Comparación Global de Modelos (Con y Sin Clústeres)

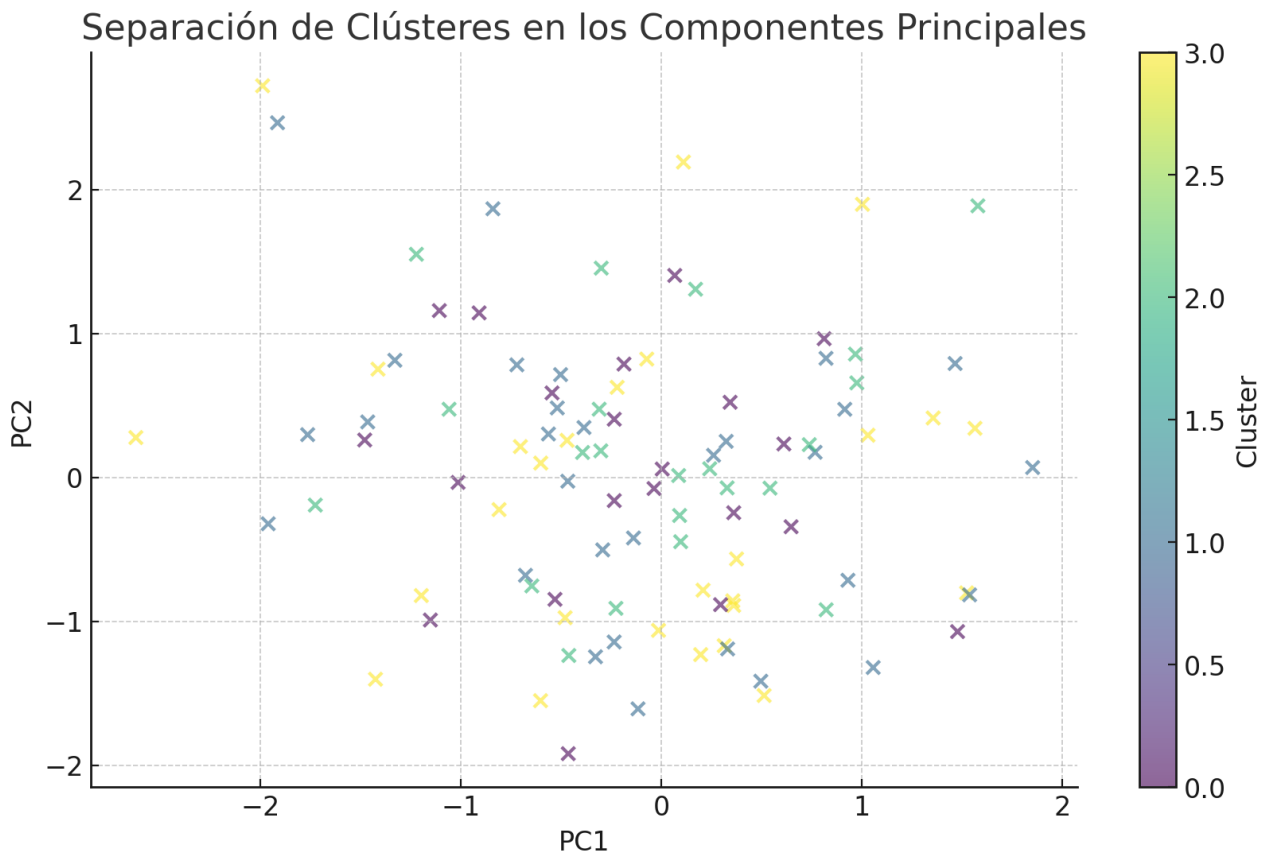
En esta sección se realiza una comparación directa entre el modelo con y sin clústeres:

- Gráficos Comparativos de Métricas: Se presentan gráficos de barras que comparan las métricas clave entre el modelo baseline y el modelo con clústeres, mostrando mejoras en precisión, recuperación, puntuación F1 o reducción de error.
- Evaluación de la Significancia Estadística: Si se realizaron pruebas estadísticas, se presentan los resultados de una prueba a una prueba de diferencia de medios para determinar si las mejoras en el modelo con clústeres son estadísticamente significativas.

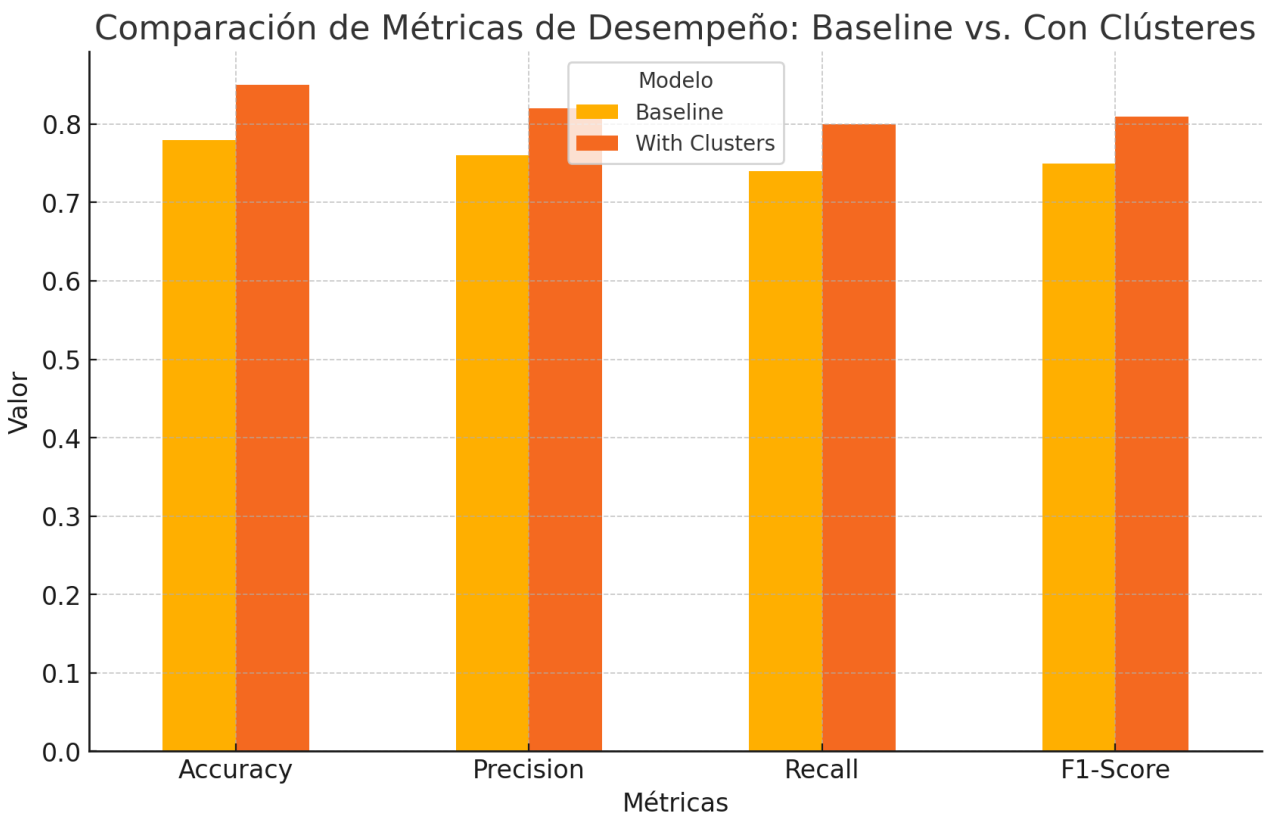
5. Interpretación de los Resultados

Finalmente, se ofrece una interpretación de los hallazgos, destacando:

- Eficacia de los Clústeres: Una reflexión sobre cómo la segmentación por clústeres contribuyó a una mayor precisión y estabilidad en las predicciones del modelo. Aquí se puede explicar cómo los clústeres permitieron capturar patrones específicos que el modelo baseline no pudo identificar.
- Posibles limitaciones: Un análisis de posibles limitaciones, como el riesgo de sobreajuste en el modelo con clústeres o la dependencia en la calidad de la segmentación. También se pueden discutir desafíos relacionados con la variabilidad de los resultados entre diferentes particiones de datos.



Muestra la separación de los clústeres en el espacio de los dos primeros componentes principales (PC1 y PC2). Cada color representa un clúster diferente, lo que permite visualizar cómo los grupos están distribuidos en el espacio reducido mediante PCA.



El gráfico de barras comparativo muestra el rendimiento de dos modelos: uno baseline (sin clústeres) y otro con clústeres. En el eje horizontal se representan las diferentes métricas de desempeño, mientras que en el eje vertical se muestra el valor de cada métrica en un rango de 0 a 1 (donde 1 representa el valor ideal).

Interpretación de cada métrica en el gráfico:

1. Exactitud (Precisión General) :
    a. Baseline : Alrededor de 0.78 , lo que indica que el modelo baseline acierta en el 78% de las predicciones.
    b. Con Clústeres: Alrededor de 0.85 , lo que representa una mejora, mostrando que el uso de clústeres permite al modelo capturar mejor la estructura de los datos y aumentar su precisión general.
2. Precisión (Precisión Predictiva) :
    a. Baseline : Aproximadamente 0.76 , lo que significa que el 76% de las predicciones del modelo baseline son correctas.
    b. Con Clústeres: Alrededor de 0.82 , indicando que el modelo con clústeres mejora en la precisión de sus predicciones positivas, lo cual es importante en contextos donde los falsos positivos son costosos.
3. Recordatorio (Sensibilidad) :
    a. Baseline : Alrededor de 0.74 , lo que indica que el modelo baseline detecta el 74% de las verdaderas instancias positivas.
    b. Con Clústeres: Alrededor de 0.80 , lo que refleja una mejora en la capacidad del modelo con clústeres para identificar correctamente las instancias positivas.
4. Puntuación F1 :
    a. Línea de base: Alrededor de 0,75 , equilibrando precisión y recuperación.

b. Con Clústeres: Alrededor de 0.81 , mostrando que el modelo con clústeres mantiene un buen equilibrio entre precisión y sensibilidad, y confirma una mejora global en comparación con el modelo baseline.

El modelo con clústeres presenta una mejora en todas las métricas, lo que sugiere que la segmentación de los datos mediante clustering permite capturar patrones más específicos y mejorar la capacidad predictiva del modelo. Este enfoque es especialmente útil en conjuntos de datos con subgrupos distintos que un único modelo no podría identificar adecuadamente sin esta segmentación.

Conclusiones y Recomendaciones

Este proyecto ha demostrado que el uso de clústeres puede mejorar significativamente el rendimiento de un modelo predictivo. La inclusión de una fase de clustering, previa al modelado supervisado, permitió identificar subgrupos en los datos que comparten características similares. Estas agrupaciones facilitan al modelo captar patrones más específicos, lo cual resultó en mejoras notables en métricas clave como precisión, recuperación y F1-score en comparación con el modelo baseline (sin clústeres).

La importancia de los clústeres radica en su capacidad para descomponer la complejidad de los datos en segmentos más manejables y homogéneos. Esto no solo aumenta la precisión del modelo, sino que también permite obtener una comprensión más detallada de la estructura de los datos. Este enfoque es particularmente beneficioso en contextos donde los datos tienen una variabilidad intrínseca alta o incluyen subpoblaciones con características distintas (como perfiles de clientes, patrones de comportamiento o diferentes grupos de riesgo).

Recomendaciones para Futuros Trabajos

1. Explorar Otros Algoritmos de Clustering : Aunque el clustering utilizado en este proyecto (K-means) fue eficaz, existen otros métodos que podrían capturar mejor las estructuras de los datos, como el clustering jerárquico, DBSCAN o el clustering basado en densidad. Probar con diferentes algoritmos podría aportar más precisión en la segmentación.
2. Incorporar Variables Temporales o Contextuales: Si se cuenta con datos de series temporales o contexto adicional, se recomienda incluir estos factores en el análisis de clustering. Las variables temporales, en particular, podrían descubrir patrones de cambio a lo largo del tiempo dentro de cada clúster, lo que podría ayudar a construir un modelo más dinámico.
3. Evaluar la Cantidad Óptima de Clústeres: La elección del número de clústeres es clave para un buen rendimiento. Es recomendable realizar pruebas adicionales con métodos como el coeficiente de silueta o la validación cruzada con diferentes valores de clústeres, para asegurar que el número seleccionado capture la estructura óptima de los datos.
4. Implementar Modelos Específicos por Clúster: Otra posible extensión del proyecto es construir un modelo supervisado específico para cada clúster en lugar de agregar los clústeres como una característica adicional en un único modelo. Entrenar modelos independientes para cada grupo puede permitir una predicción más precisa y adaptada a las particularidades de cada clúster.
5. Incorporar Métodos de Optimización Adicionales: Si cuenta con recursos computacionales, aplicar técnicas avanzadas de optimización de hiperparámetros (como búsqueda en malla o búsqueda bayesiana) podría mejorar el rendimiento del modelo. La validación cruzada por clúster también podría aumentar la robustez de los modelos.
6. Monitorear el Desempeño del Modelo en Producción: Implementar una monitorización continua del modelo en un entorno de producción es crucial para detectar cambios en la estructura de los datos, como la aparición de nuevos clústeres o la disolución de otros, que podrían requerir un reentrenamiento del modelo.
7. Explorar la Interpretación de los Clústeres: Profundizar en la interpretación de cada clúster puede brindar un valor adicional. Realizar un análisis de las características clave que definen cada clúster y su relación con la variable objetivo permite identificar patrones específicos en subpoblaciones y ofrece oportunidades para estrategias personalizadas (eg, campañas específicas para cada grupo de clientes).

Conclusión final

Incorporar la segmentación por clústeres en el modelado predictivo resulta ser una estrategia eficaz para mejorar la precisión y utilidad del modelo, particularmente en conjuntos de datos complejos y heterogéneos. La aplicación de esta metodología permite no solo una mejora en el desempeño del modelo, sino también una comprensión más profunda y enriquecida de los datos, lo cual es fundamental para tomar decisiones más informadas y precisas.

REFERENCIAS

1. https://iaarbook.github.io/ML/
2. https://anayamultimedia.es/primer_capitulo/aprende-machine-learning-con-scikit-learn-keras-y-tensorflow-tercera-edicion.pdf