**Articulo Resumen del Análisis de Datos sobre Obesidad utilizando Técnicas de Clustering, Preprocesamiento y Balanceo de Clases**

**Introducción**

El análisis de datos sobre obesidad es fundamental para entender los factores que contribuyen a la salud pública. En este artículo, se aplicaron varias técnicas de procesamiento y modelado de datos, que incluyen **preprocesamiento de datos**, **balanceo de clases**, **clustering no supervisado** y **reducción de dimensionalidad**, con el fin de obtener un análisis más preciso de los patrones asociados con la obesidad. El enfoque que se sigue es integral, comenzando con la limpieza y transformación de los datos, seguido de la aplicación de modelos para segmentación y predicción.

**1. Preprocesamiento y Balanceo de Datos**

El primer paso en el análisis de datos es la limpieza y preparación de los datos. Para ello, se realizaron los siguientes pasos:

* **Normalización de Características**: Utilizamos **StandardScaler** para normalizar las características numéricas del dataset (edad, peso, altura). Esto permitió que los algoritmos de machine learning trabajaran con datos escalados y homogéneos.
* **Codificación de Variables Categóricas**: Las variables categóricas como **género** e **historia familiar de obesidad** fueron codificadas utilizando **LabelEncoder**, convirtiendo las categorías en valores numéricos para ser procesadas adecuadamente por los modelos.
* **Balanceo de Clases con SMOTE**: Como el dataset original presentaba un desbalance de clases en la variable objetivo **NObeyesdad**, se aplicó **SMOTE** (Synthetic Minority Over-sampling Technique), que genera ejemplos sintéticos de la clase minoritaria para equilibrar las clases. Este paso mejoró la capacidad predictiva del modelo al evitar que se inclinara hacia la clase mayoritaria.

**2. Selección del Clasificador: Random Forest**

Para clasificar a los individuos según su obesidad, se seleccionó el algoritmo **Random Forest**, un modelo robusto y adecuado para manejar datos desbalanceados. Utilizamos **validación cruzada** para evaluar la confiabilidad del modelo y obtener métricas clave de desempeño, como la **matriz de confusión**. El modelo fue entrenado y evaluado, proporcionando una base sólida para la clasificación de datos de obesidad.

**Justificación del clasificador**: El **Random Forest** es adecuado debido a su capacidad para manejar relaciones no lineales, interacciones entre variables y su resistencia al sobreajuste, lo que lo convierte en una opción ideal para este tipo de análisis de datos.

**3. Evaluación del Modelo: Confiabilidad y Matriz de Confusión**

En la primera ejecución del modelo, se evaluó su confiabilidad utilizando la **matriz de confusión**, que permitió observar la precisión y el recall del clasificador. La matriz de confusión mostró qué tan bien el modelo predice las clases correctas (obesidad) en comparación con las incorrectas.

**4. Split de Datos: Entrenamiento y Prueba**

Para realizar una evaluación más robusta, se dividieron los datos en dos sets: uno para **entrenamiento (80%)** y otro para **prueba (20%)**. Además, se realizaron al menos **100 asignaciones** con validación cruzada, calculando la **mediana de confiabilidad** para evaluar la estabilidad del modelo. También se realizó una segunda ejecución con un split **50/50** para verificar la generalización del modelo en otro tipo de división de datos.

**5. Reducción de Dimensionalidad: PCA**

Para optimizar los modelos y reducir la complejidad computacional, se aplicó **PCA** (Análisis de Componentes Principales). Este paso ayudó a reducir la cantidad de variables a un número más manejable, sin perder la variabilidad importante del conjunto de datos. Se probaron diferentes configuraciones de componentes principales (12, 10, 9, 5, 3) y se seleccionó el número óptimo que ofreció el mejor rendimiento para el modelo de clasificación.

**6. Clustering: Segmentación No Supervisada con K-Means**

El análisis de **clustering** fue el siguiente paso. Utilizamos el algoritmo **K-Means**, que segmentó a los individuos en tres grupos (clusters) según sus características. Cada cluster representa un grupo de individuos con características similares, lo que permite entender cómo se agrupan los datos en relación con factores como la edad, el peso, la altura y los antecedentes familiares.

La distribución de los clusters mostró que existían tres grupos predominantes, y la visualización de estos clusters en gráficos 2D y 3D facilitó la comprensión de cómo se agrupan los datos en función de las características seleccionadas.

**7. Resultados del Clustering y Análisis Final**

Los resultados del análisis de clustering revelaron tres segmentos clave:

* **Cluster 0**: Personas jóvenes con un peso moderado y antecedentes familiares limitados.
* **Cluster 1**: Individuos jóvenes con antecedentes familiares de obesidad, un grupo más grande y en riesgo potencial.
* **Cluster 2**: Individuos con mayor peso y antecedentes familiares significativos de obesidad, lo que indica una mayor predisposición genética.

Estos resultados demuestran cómo el uso de técnicas no supervisadas como **K-Means** puede revelar patrones ocultos en los datos y ayudar a segmentar a los individuos según sus características.

**Conclusiones**

Este artículo ha demostrado cómo una combinación de técnicas de preprocesamiento, balanceo de clases, clasificación y clustering puede proporcionar una visión más profunda de los patrones relacionados con la obesidad. Las técnicas aplicadas en este análisis, como **SMOTE**, **Random Forest**, **PCA** y **K-Means**, han permitido segmentar y comprender mejor los datos sobre obesidad, lo que puede tener implicaciones significativas para la salud pública.

El uso de estas técnicas en conjunto permite no solo realizar predicciones precisas, sino también identificar grupos de riesgo, lo que puede ser útil para diseñar estrategias de intervención más eficaces.