**Artículo sobre el Análisis y Segmentación de Datos sobre Obesidad mediante Técnicas de Clustering y Balanceo de Clases**

**1. Introducción**

El análisis de datos es una herramienta fundamental en la ciencia de datos, y su aplicación en la salud pública permite identificar patrones relevantes que pueden llevar a decisiones informadas. La obesidad es un problema de salud creciente en muchas poblaciones a nivel mundial, y su estudio puede verse beneficiado por la segmentación de individuos con características comunes, como la edad, el peso y la historia familiar.

El objetivo de este artículo es aplicar técnicas de análisis de datos para explorar un conjunto de datos sobre obesidad, donde se utiliza un enfoque de clustering para agrupar a los individuos según sus características y se emplean técnicas de balanceo de clases para mejorar los resultados de los modelos predictivos. A través de estas técnicas, se busca entender cómo las variables como la edad, el peso, la altura y otros factores influyen en la obesidad y la segmentación de la población.

**2. Revisión Bibliográfica**

**Preprocesamiento de Datos**

El preprocesamiento de datos es una fase crucial en cualquier análisis de datos, ya que garantiza que los modelos sean alimentados con datos de calidad. Esta fase incluye la limpieza de datos, la normalización de características, la imputación de valores faltantes y la codificación de variables categóricas.

En este estudio, se utilizó la normalización de características numéricas, como la edad, el peso y la altura, mediante el uso de **StandardScaler** de sklearn. Esta técnica garantiza que todas las características tengan un rango similar, lo que es importante para algoritmos como K-Means que son sensibles a la escala de los datos.

Además, las variables categóricas, como el género y la historia familiar de obesidad, fueron codificadas utilizando **LabelEncoder**, lo que permite convertir estas variables en valores numéricos y hacerlas utilizables para los modelos de machine learning.

**Balanceo de Clases: SMOTE**

Uno de los mayores problemas en el análisis de datos es el **desbalance de clases**, especialmente cuando una clase está sobrerrepresentada en comparación con otras. Este desbalance puede afectar negativamente el rendimiento de los modelos de clasificación. **SMOTE** (Synthetic Minority Over-sampling Technique) es una técnica que genera ejemplos sintéticos de la clase minoritaria, lo que ayuda a equilibrar la distribución de clases sin necesidad de eliminar datos de las clases mayoritarias.

En este artículo, se utilizó SMOTE para balancear el conjunto de datos, específicamente la variable NObeyesdad, que es la etiqueta de la obesidad (con clases como "obesidad severa", "sobrepeso", etc.). La aplicación de SMOTE permitió mejorar la capacidad del modelo para aprender las características de las clases minoritarias.

**K-Means y Técnicas de Clustering**

El **clustering** es una técnica de aprendizaje no supervisado que se utiliza para agrupar datos en segmentos o clusters basados en la similitud de sus características. En este estudio, se utilizó el algoritmo **K-Means**, que asigna los datos a **k** clusters según las características de los individuos. El valor de **k** fue determinado empíricamente, evaluando la dispersión de los datos y la homogeneidad de los clusters.

Una de las principales ventajas de K-Means es su simplicidad y rapidez, aunque una de las limitaciones es la necesidad de definir el número de clusters a priori. Por esta razón, en este artículo, se utilizó el índice de la silueta para evaluar la calidad del clustering y determinar un valor adecuado de **k**.

**Reducción de Dimensionalidad: PCA**

La reducción de dimensionalidad es otra técnica clave utilizada en este artículo. A menudo, los conjuntos de datos contienen un gran número de características, lo que puede llevar a problemas de **sobreajuste** y aumentar la complejidad computacional. **PCA** (Principal Component Analysis) es una técnica de reducción de dimensionalidad que proyecta los datos en un espacio de menor dimensión manteniendo la mayor variabilidad posible.

La aplicación de PCA en este estudio permitió reducir el número de características del dataset, mejorando la eficiencia del algoritmo de clustering y facilitando la interpretación de los resultados.

**3. Metodología**

**Descripción del Dataset**

El conjunto de datos utilizado en este estudio proviene de un estudio sobre obesidad, con información sobre varias características de los individuos, tales como:

* **Edad**: La edad de los individuos en años.
* **Altura**: La altura de los individuos en metros.
* **Peso**: El peso de los individuos en kilogramos.
* **Género**: El género de los individuos, representado por valores binarios.
* **Historia Familiar de Sobrepeso**: Variable binaria que indica si el individuo tiene antecedentes familiares de obesidad.

La variable dependiente, **NObeyesdad**, es una variable categórica que indica la condición de obesidad del individuo (por ejemplo, "obesidad severa", "sobrepeso", etc.).

**Preprocesamiento de Datos**

El proceso de preprocesamiento siguió los siguientes pasos:

1. **Normalización de Características Numéricas**: Se aplicó **StandardScaler** para transformar las características numéricas en un rango estándar de media 0 y desviación estándar 1.
2. **Codificación de Variables Categóricas**: Se utilizó **LabelEncoder** para convertir las variables categóricas como el género y la historia familiar en valores numéricos.

**Balanceo de Clases**

Se aplicó **SMOTE** para balancear la variable NObeyesdad y asegurar que las clases estuvieran equilibradas antes de aplicar cualquier modelo de machine learning. Este paso fue crucial para evitar que las clases minoritarias fueran ignoradas por el modelo.

**Clustering con K-Means**

El algoritmo **K-Means** fue utilizado para realizar la segmentación de los individuos. En este caso, se seleccionó un valor de **k = 3** para los clusters, ya que, basado en el análisis visual y la distribución de los datos, se observaron tres grupos principales que podrían representar diferentes perfiles de obesidad.

**Evaluación del Modelo**

La evaluación del modelo de clustering se realizó utilizando la **matriz de confusión**, así como la evaluación visual de los resultados de clustering y las características principales de cada cluster.

**4. Resultados**

**Distribución de Clusters**

Tras la ejecución de K-Means, se obtuvo la siguiente distribución de los individuos en los tres clusters:

* **Cluster 0**: 539 individuos
* **Cluster 1**: 946 individuos
* **Cluster 2**: 626 individuos

Cada cluster representa un grupo de individuos con características comunes, y a continuación se presenta un análisis detallado de las características promedio de cada grupo.

**Análisis de Clusters**

**Cluster 0**

Este grupo está compuesto principalmente por individuos de edad media, con un peso y altura equilibrados. La mayoría de estos individuos no presentan antecedentes familiares de obesidad. Este cluster puede representar a individuos con una condición de sobrepeso moderado o en riesgo de obesidad.

**Cluster 1**

Este grupo es el más grande y está compuesto por individuos jóvenes con una baja prevalencia de obesidad severa. Sin embargo, muchos de ellos tienen antecedentes familiares de obesidad, lo que puede aumentar su riesgo. Este cluster podría ser clave para intervenciones preventivas.

**Cluster 2**

El último grupo está compuesto por individuos con mayor peso y menos altura, que presentan una alta prevalencia de obesidad severa. Muchos de estos individuos tienen antecedentes familiares de obesidad, lo que sugiere que la genética juega un papel importante en su condición.

**5. Discusión**

El análisis de clustering permitió identificar tres grupos bien diferenciados con características comunes en cuanto a la edad, el peso, la altura y la historia familiar de obesidad. Estos resultados son coherentes con investigaciones previas sobre la obesidad, que sugieren que factores como la genética, la dieta y el estilo de vida influyen en la condición de obesidad.

El uso de SMOTE para balancear las clases permitió que el modelo aprendiera adecuadamente de las clases minoritarias, lo que se traduce en un mejor desempeño en términos de precisión y recall.

Aunque el valor de **k = 3** parece ser adecuado, es recomendable realizar una validación más rigurosa del número de clusters utilizando técnicas como el índice de silueta o el método del codo para obtener una segmentación aún más precisa.

**6. Conclusiones**

El uso de técnicas de **clustering**, **balanceo de clases** y **reducción de dimensionalidad** permitió segmentar con éxito a los individuos en grupos homogéneos con características comunes, lo que podría ser útil para el diseño de intervenciones de salud pública. Los resultados muestran cómo el clustering no supervisado puede proporcionar una comprensión más profunda de los factores que influyen en la obesidad.

Además, se demostró que el balanceo de clases mediante SMOTE es una técnica efectiva para manejar conjuntos de datos desbalanceados, y PCA puede ser útil para reducir la complejidad del modelo sin perder información crítica.

**7. Referencias**

* Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer. ISBN: 978-0-387-84857-0.