UNIVERSIDAD DE ANTIOQUIA

Caso de estudio - Aprendizaje No Supervisado
Analítica para la toma de decisiones

Alanis Álvarez, Juan E. Soto, Paola A. Arabia, Santiago Restrepo
Departamento de Ingeniería Industrial
Universidad de Antioquia, Colombia

Introducción

Este informe presenta un proyecto de Machine Learning de utilidad en el ámbito de la salud, más específicamente en el tema nutricional, que tiene como objetivo realizar agrupaciones de personas a través de algoritmos de aprendizaje no supervisado basados en sus características. Para ello, se utiliza como insumo un conjunto de datos del repositorio de la Universidad de California (UC Irvine), el cual recopila datos sobre los hábitos alimenticios y la condición física de personas en Perú, Colombia y México.

Se dispone de un notebook de Jupyter dedicado a la exploración y preprocesamiento de los datos, seguido por la implementación de las técnicas de agrupación, comparación y selección de las mejores métricas arrojadas por los algoritmos, con el fin de dar solución a la problemática, evaluando, graficando y analizando los diferentes clústers identificados por el algoritmo, lo que permitirá comprender de manera más profunda las características de los patrones y su relación dentro de los datos.

Caso de estudio (Planteamiento del problema)

La obesidad está influenciada por una combinación de factores relacionados con los hábitos y estilos de vida de las personas. Se especula, que una baja actividad física, una alta ingesta calórica y hábitos alimenticios poco saludables contribuyen significativamente al desarrollo de esta condición.

Ahora bien, Carlos, un nutricionista que trabaja en el centro de salud de la EPS Sura está interesado en comprender cómo los diferentes hábitos y estilos de vida de sus pacientes influyen en que estos puedan llegar a tener obesidad. El nutricionista cuenta con una base de datos recopilada a partir de sus diferentes consultas donde se incluye información acerca de la edad, el género del paciente, altura, peso, si la persona suele contar la cantidad de calorías que consume, cuantas comidas principales realiza al día, cuánta actividad física realiza por semana, entre otras características del estilo de vida que lleva cada paciente. El objetivo del nutricionista es identificar los diferentes patrones o comportamientos en sus pacientes que puedan estar asociados con un mayor riesgo de obesidad y busca determinar si existen grupos de pacientes que comparten características comunes que influyen en esta situación.

Diseño de solución propuesto

De acuerdo a los clusters generados por los modelos, con el fin de aprovechar la visualización que brinda cada agrupación por características, se le propone a la empresa, el desarrollo de una plataforma digital o una extensión en la app actual que le permita tanto a los pacientes como al nutricionista encargado, ingresar información relevante sobre sus hábitos y estilos de vida, similar a la información recopilada durante las consultas, con el objetivo de que el sistema genere recomendaciones a través de notificaciones, que sean personalizadas para cada grupo de pacientes basadas en los patrones de riesgo identificados a través de algoritmos de aprendizaje automático. Estas recomendaciones incluirán pautas de alimentación, ejercicios físicos, seguimiento de calorías, entre otros hábitos saludables. Esto puede incluir recordatorios para registrar su ingesta calórica, realizar actividades físicas o programar consultas con el nutricionista. Esto aumentaría de manera efectiva las intervenciones para prevenir y controlar la obesidad, automatizando la intermediación entre el nutricionista y el paciente.

Limpieza y transformación de los datos

Inicialmente es importante mencionar que el dataset original contiene 2111 observaciones y 17 columnas, sin embargo, se elimina de entrada la variable objetivo dado que vamos a realizar un proyecto de aprendizaje no supervisado. Así mismo, ninguna de las variables posee datos faltantes y 8 de ellas son de tipo *float* y 8 son de tipo *object*, las cuales se proceden a convertir en numéricas. Finalmente, el nombre de las variables se modifica para facilitar el entendimiento de estas.

Para convertir las variables tipo *object* a tipo numérica, que en su mayoría son de tipo *bool* se reemplazan los valores de 'yes' por 1 y los valores de 'no' por 0. Adicionalmente, existen un par de variables de tipo ordinal (valores de 'no', 'Sometimes', 'Frequently' y 'Always') reemplazando su valor más bajo por 0 y el más alto por 4. Finalmente, una variable que contenía 5 clases relacionadas con el método de transporte utilizado por la persona, se ejecuta un procedimiento similar reemplazando las clases que no realizan esfuerzo físico para desplazarse por 0 y las clases que si realizan esfuerzo físico por 1.

Finalmente, se evidencia que el dataset posee 24 observaciones duplicadas, que se proceden a eliminar, y además, no se evidencia presencia de datos atípicos (por ende no se trata).

Análisis exploratorio de los datos

Para la exploración de los datos se ejecuta principalmente un análisis univariado a través de la descripción y conteo de valores por variables para identificar la distribución y comportamiento de las mismas. Así mismo, se realizan boxplots para las variables numéricas de las cuales se evidencia que no existen datos atípicos. Adicionalmente, se realiza la matriz de correlación de estas mismas variables, de la cual se evidencia que no existen correlaciones considerables entre ellas.

Selección de variables

A partir del análisis exploratorio se observa que el dataset no presenta datos atípicos o variables innecesarias que ameriten la eliminación de columnas. Sin embargo, se realizaron histogramas de las variables evidenciando que *'FUMADOR', 'MONITOREA_CALORIAS'* y *'MEDIO_TRANSPORTE'* presenta un desbalanceo notable en sus valores, por ende se deciden eliminar del dataset. Con esto, se obtiene un dataset con 2087 observaciones y 13 columnas después del análisis exploratorio y tratamiento de variables, el cual se procede a escalar con el método *StandardScaler()* para ejecutar los algoritmos de agrupación.

Selección y aplicación de algoritmos/ técnicas de modelado

Se decide hacer uso los siguientes algoritmos de agrupación y reducción de dimensionalidad:

- **K-means:** Se decide hacer uso de este algoritmo dado que suele ser el más usado en los problemas de agrupación. Además, una vez realizado el preprocesamiento de los datos se obtiene una base de datos mucho más limpia con la cual se pueden obtener agrupaciones más precisas utilizando K-means. Finalmente, para determinar el número de clústers a utilizar se emplea el método del codo (elbow method), lo que nos da como resultado un total de 5 clústers.
- **DBScan:** Por otro lado, se decide hacer uso del algoritmo DBScan dado que se cuenta con una base de datos con pocas muestras y además, es un algoritmo robusto en cuanto a la detección de datos atípicos, por ende, podría realizar agrupaciones mucho más precisas y diferenciables. Para determinar sus parámetros más importantes se hace uso del método K-vecino más cercano (punto de máxima curvatura) para determinar el valor de *epsilon* y se realiza el gráfico del score de silueta contra el mínimo número de muestras para obtener el parámetro *min_samples.*

- **PCA (Análisis de componentes principales):** Ahora, en los dos algoritmos mencionados se ejecuta reducción de la dimensionalidad a través del algoritmo de PCA con el objetivo de hallar componentes que puedan explicar de forma más precisa las agrupaciones realizadas.

Comparación y selección de técnicas

Para la ejecución de los algoritmos mencionados anteriormente se emplearon 2 modelos con cada uno de ellos (obteniendo un total de 4 modelos): uno de ellos utilizando los datos estandarizados resultantes del preprocesamiento y el otro de ellos utilizando el conjunto de datos con reducción de la dimensionalidad (PCA) y optimización de hiperparámetros que explicara al menos el 55% de la varianza. Al ejecutar el modelo definiendo una varianza del 55% se obtienen 5 componentes principales que explican alrededor del 60% de la varianza.

Ahora, para realizar la comparación entre los modelos empleados se decide utilizar las métricas de inercia, el score de silueta y Calinski Harabasz. El valor de estas métricas nos permiten evaluar el comportamiento de las agrupaciones generadas por los algoritmos, así mismo, el modelo que arroje un valor superior de las métricas mencionadas (mayormente silueta y Calinski) será el escogido como óptimo.

Afinamiento de hiperparámetros

Para la optimización de parámetros se hace uso del método de búsqueda aleatoria, en la cual se obtienen los parámetros que está utilizando por defecto el modelo (función '*get_params()'*), seguido a esto se realiza la consulta de los posibles valores que estos pueden tomar en la página de sklearn y finalmente, se emplea la búsqueda de los parámetros que maximicen el score. Sin embargo, cabe mencionar que una vez ejecutada la optimización de hiperparámetros en ambos algoritmos no se obtienen mejoras considerables en las métricas de desempeño.

Evaluación y análisis del mejor modelo

Una vez se ejecutan los algoritmos se obtienen las siguientes métricas de desempeño:

Modelo		Inertia	Silhouette Score	Calinski harabasz score
K-means	Base	19092.099	0.121	219.161
	PCA optimizado	8253.584	0.235	509.718
DBScan	Base	-	0.207	115.663
	PCA optimizado	-	0.222	209.495

Tabla 1. Métricas de desempeño de los modelos de aprendizaje no supervisado empleados

Al observar los resultados clasificados en la Tabla 1 se concluye que el mejor modelo obtenido es el K-means aplicando reducción de dimensionalidad con PCA y optimización de hiperparámetros a pesar de obtener un score de silueta más cercano a 0 que a 1, lo que nos permite afirmar que las agrupaciones realizadas por el algoritmo no son tan precisas y diferenciables. Sin embargo, se desea escoger este modelo dado que los demás obtienen métricas inferiores.

Ahora, para realizar el análisis de los clústers generados por el algoritmo se decide realizar un procedimiento gráfico utilizando los valores de las variables iniciales (antes de estandarizar),

recordando que este modelo arrojó un total de 5 grupos. En la Tabla 2 se presenta la descripción de las características encontradas por cada clúster:

Clúster	Características/Descripción de clúster
0	- Posee a las personas mayores (media: 28 años) - Posee más mujeres que hombres - Posee personas con la estatura más baja del estudio - Predominan las personas que no consumen alcohol y algunas personas que lo consumen a veces - Caracterizado por personas que a veces comen entre comidas - Bajo consumo de agua al día - Posee personas que menos actividad física practican - Personas que no suelen consumir 3 comidas principales al día - Personas que no suelen pasar mucho tiempo al frente de dispositivos tecnológicos
1	- Contiene a las personas con menor edad del estudio - Predomina la presencia de hombres - Personas con una altura promedio cercana a los 1.8 m. - Predominan las personas que no consumen alcohol y algunas que lo hacen a veces - Predominan las personas que a veces comen entre comidas - Personas que más agua consumen al día - Personas que más actividad física practican - Personas que menos consumen verduras - Contiene a las personas que más tiempo pasan al frente de dispositivos tecnológicos
2	- Contiene a las personas más livianas del estudio (menor peso) - Personas jóvenes - Predominan las mujeres - Personas con estatura baja - Predominan las personas que consumen alcohol a veces y algunas que no lo consumen - Este clúster contiene a la mayoría de personas que comen entre comidas frecuentemente y siempre, también algunas personas que lo hacen a veces (casi ninguna no lo hace) - Bajo consumo de agua - Consumo de verduras levemente superior - Poco tiempo al frente de dispositivos
3	- Personas levemente pesadas (media: 103 kg.) - Contiene personas con mayor edad del estudio - Prácticamente sólo posee hombres (alrededor del 98%) - Personas más altas del estudio - Prácticamente todas las personas consumen alcohol a veces - Prácticamente todas las personas a veces comen entre comidas - Alrededor del 95% de las personas consumen comidas altamente calóricas
4	- Contiene a las personas más pesadas del estudio (media: 114 kg.) - Prácticamente todas las personas son mujeres - Prácticamente todas consumen alcohol a veces - Todas comen entre comidas a veces - Alto consumo de agua - Es el clúster que contiene a las personas que menos actividad física practican - Personas con mayor consumo de verduras al día - Personas que normalmente consumen 3 comidas principales al día - Todas las personas consumen comidas altamente calóricas

Tabla 2. Principales características de los clústers generados por K-means con PCA optimizado

Entonces, gracias a la descripción de los grupos empleada en la Tabla 2 se decide nombrar a los grupos de la siguiente manera:

- **Clúster 0: Personas mayores propensas a sufrir obesidad**
- **Clúster 1: Jóvenes sanos propensos a ser pesados**
- **Clúster 2: Jóvenes sanos con peso insuficiente**
- **Clúster 3: Hombres con sobrepeso**
- **Clúster 4: Mujeres con sobrepeso**

De esta forma se podría concluir que las personas contenidas en el clúster 3 (hombres) y 4 (mujeres) son las que se encuentran en condición de obesidad más crítica, las personas contenidas en el clúster 0 son las que aún no poseen una condición de obesidad pero conservan hábitos que los pueden llevar a ello, las personas contenidas en el clúster 1 son principalmente hombres que no suelen cuidar su alimentación pero tienen hábitos saludables que los hacen mantener en peso normal, y finalmente, en el clúster 2 predominan las mujeres con menor peso o insuficiente.

Conclusiones finales y recomendaciones

Principalmente, para concluir sobre el modelo (K-means con PCA optimizado) se observó que obtuvo métricas de desempeño que pueden ser mejores, sin embargo, se obtuvieron agrupaciones valiosas y de utilidad para analizar y responder al problema. Así mismo, la elección del algoritmo a utilizar dependerá de diversos factores como el tamaño del dataset, la distribución de los datos y los datos atípicos que contenga.

En cuanto a la utilidad del modelo en la EPS se apreció que es considerable dado que se puede emplear en diferentes situaciones y planes que la entidad decida relacionada con el tratamiento de la obesidad, y de esa forma llegar a manejar tanto de forma preventiva como correctiva algunas acciones que favorezcan al estilo de vida y la salud de las personas. Adicionalmente, los clústers generados por el modelo permiten dar respuesta a la problemática planteada a través de las características de cada uno de ellos, y también, desplegar un correcto funcionamiento de la solución propuesta a través de la app.

Finalmente, se le recomienda a la empresa emplear mayores recursos en la optimización de hiperparámetros de modo que el modelo sea capaz de obtener mejores resultados, lo que generaría obtener agrupaciones mucho más precisas, y por ende, conclusiones mucho más contundentes. Por otro lado, en caso de que la EPS implemente la ejecución del modelo y ponga en marcha la solución propuesta, se le recomienda que mantenga en constante revisión los datos que recolecta dado que la presencia de datos atípicos puede generar que el modelo disminuya su rendimiento. Por último, lo más importante es que la EPS mantenga en constante observación y control principalmente a las personas agrupadas en los clústers más riesgosos.

Bibliografía

scikit-learn: machine learning in Python — scikit-learn 1.4.1 documentation. (s. f.). https://scikit-learn.org/stable/

UCI Machine Learning Repository. (s. f.). *https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition*