**Informe Técnico del Proyecto de Clasificación de Niveles de Obesidad**

**Proyecto:** Desarrollo de un Modelo de Machine Learning para la Clasificación Multiclase de Niveles de Obesidad.  
**Fecha:** 23 de mayo de 2024  
**Versión:** 0.1  
**Autores:** Anca Bacria, Fernando García, Omar Lengua y Abigaíl Masapanta.

1. **Introducción y Objetivos del proyecto.**

Este documento presenta un resumen técnico inicial basado en el Análisis Exploratorio de Datos (EDA) realizado sobre un conjunto de datos destinado al desarrollo de un modelo de Machine Learning. El objetivo principal del proyecto es construir un sistema capaz de realizar una clasificación multiclase para predecir el nivel de obesidad de un individuo (NObeyesdad) a partir de un conjunto de características antropométricas, demográficas y de hábitos de vida. El modelo deberá ser capaz de distinguir entre un mínimo de tres clases, y el dataset actual provee siete clases distintas.

1. **Descripción del Conjunto de datos.**

* **Fuente:** Dataset train.csv (muestra de 50 filas en prueba.csv).
* **Dimensiones:** El dataset de entrenamiento (train.csv) consta de 20,758 registros (individuos) y 18 columnas (1 ID, 16 características predictoras y 1 variable objetivo).
* **Variable Objetivo:** La columna NObeyesdad es la variable categórica a predecir. Presenta 7 clases únicas:
  1. Insufficient\_Weight
  2. Normal\_Weight
  3. Overweight\_Level\_I
  4. Overweight\_Level\_II
  5. Obesity\_Type\_I
  6. Obesity\_Type\_II
  7. Obesity\_Type\_III
* **Características (Features):**
  1. **Numéricas:** Age, Height, Weight, FCVC, NCP, CH2O, FAF, TUE.
  2. **Categóricas:** Gender, family\_history\_with\_overweight, FAVC, CAEC, SMOKE, SCC, CALC, MTRANS.
  3. **Identificador:** id (eliminado durante el EDA por no ser predictivo).

1. **Resumen del Análisis Exploratorio de Datos (EDA)**
   1. **Calidad de los Datos**

* **Valores Nulos:** No se encontraron valores nulos en ninguna columna del dataset.
* **Duplicados:** No se encontraron filas completamente duplicadas.
* **Tipos de Datos:** Los tipos de datos parecen ser consistentes con la naturaleza de las variables (numéricas
* como float64, categóricas como object).
  1. **Variable Objetivo (NObeyesdad)**
* **Distribución:** Las clases de la variable objetivo presentan un desbalance moderado. La clase más frecuente es Obesity\_Type\_III (4046 instancias) y la menos frecuente es Overweight\_Level\_I (2427 instancias). Este desbalance deberá ser considerado durante el modelado.

|  |  |
| --- | --- |
| **NObeyesdad** | **Count** |
| Obesity\_Type\_III | 4046 |
| Obesity\_Type\_II | 3248 |
| Normal\_Weight | 3082 |
| Obesity\_Type\_I | 2910 |
| Insufficient\_Weight | 2523 |
| Overweight\_Level\_II | 2522 |
| Overweight\_Level\_I | 2427 |

* 1. **Análisis de Variables Numéricas**
* **Distribuciones:**
  + Age: Sesgada a la derecha, concentración en edades jóvenes.
  + Height, Weight: Distribuciones aproximadamente normales.
  + FCVC, NCP, CH2O, FAF, TUE: Muestran concentraciones en valores discretos, sugiriendo escalas o categorías codificadas. **Se requiere un diccionario de datos para su correcta interpretación.**
* **Correlaciones:**
  + Weight y Height muestran una correlación positiva moderada (0.46).
  + La correlación de Weight (0.82) y Age (0.30) con la variable objetivo NObeyesdad (codificada numéricamente) es la más destacada entre las variables numéricas. FAF (-0.20) y Height (-0.12) muestran correlaciones negativas.
* **Multicolinealidad (VIF):** Todas las variables numéricas presentan valores de VIF muy bajos (< 2), indicando ausencia de multicolinealidad problemática.
* **ANOVA:** Los tests ANOVA indican que las medias de Age, Height, Weight, FAF, y TUE difieren significativamente entre las distintas clases de NObeyesdad (p-values <<< 0.05), sugiriendo su alta relevancia predictiva.
  1. **Análisis de Variables Categóricas**
* **Asociación (V de Cramér):**
  + Las variables CAEC, MTRANS, CALC y family\_history\_with\_overweight muestran los niveles más altos de asociación con NObeyesdad.
  + Gender, FAVC, SMOKE, SCC muestran asociaciones más débiles, pero potencialmente útiles.
* **Visualizaciones:** Los countplot muestran patrones distintivos en la distribución de NObeyesdad según las categorías de las otras variables (ej. una mayor proporción de ciertos tipos de obesidad en individuos con historial familiar).
  1. **Outliers**

Se identificaron y aplicó un método de eliminación de outliers (IQR) para las columnas Weight, Height, Age, FAF, TUE. Se recomienda verificar el impacto (cantidad de datos eliminados) de este paso.

1. **Conclusiones Preliminares del EDA**

* El dataset es de buena calidad inicial (sin nulos, sin duplicados).
* La variable objetivo NObeyesdad está bien definida con 7 clases, cumpliendo el requisito del proyecto. Existe un desbalance moderado que debe ser manejado.
* Variables numéricas como Weight, Age, y Height son fuertes candidatas para la predicción.
* Variables categóricas como CAEC, MTRANS, y family\_history\_with\_overweight también muestran un potencial predictivo importante.
* No hay problemas graves de multicolinealidad entre las *features* numéricas.
* **La principal limitación actual es la falta de un diccionario de datos detallado, especialmente para las columnas con abreviaturas (FAVC, FCVC, NCP, CH2O, SCC, FAF, TUE) y aquellas numéricas con distribuciones discretas.**

1. **Recomendaciones y Próximos Pasos (Técnicos)**

* **OBTENER DICCIONARIO DE DATOS:** Es prioritario definir el significado exacto y las posibles unidades/escalas de todas las variables, especialmente FAVC, FCVC, NCP, CAEC, CH2O, SCC, FAF, TUE, CALC.
* **Feature Engineering:**
  + Considerar la creación del Índice de Masa Corporal (IMC) = Weight / (Height^2).
  + Explorar la creación de categorías a partir de variables continuas (ej. grupos de edad).
* **Preprocesamiento Detallado:**
  + Codificación adecuada de variables categóricas (One-Hot Encoding, Label Encoding según el caso y el modelo a usar).
  + Escalado/Normalización de variables numéricas (StandardScaler, MinMaxScaler).
* **Selección/Entrenamiento de Modelos:**
  + Comenzar con modelos base (Regresión Logística, k-NN, Naive Bayes).
  + Probar modelos más avanzados (Máquinas de Vectores de Soporte, Random Forest, Gradient Boosting como XGBoost, LightGBM, CatBoost).
  + Considerar el uso de redes neuronales si la complejidad lo justifica.
* **Manejo de Desbalance de Clases:** Implementar técnicas como SMOTE, ADASYN, o el uso de pesos de clase en los algoritmos si el rendimiento en clases minoritarias es un problema.
* **Evaluación:** Utilizar métricas apropiadas para clasificación multiclase (Accuracy, Precision, Recall, F1-score por clase y promediados, Matriz de Confusión, Curva ROC AUC (OvR/OvO)).
* **Interpretabilidad:** Planificar el uso de técnicas (SHAP, LIME) para entender las predicciones del modelo.

1. **Potencial del Modelo**

Un modelo preciso y robusto para la clasificación de niveles de obesidad tiene un alto potencial de aplicación en áreas como:

* Sistemas de recomendación de salud personalizados.
* Herramientas de apoyo para profesionales de la salud.
* Diseño de programas de bienestar corporativo y de salud pública.