Proyecto Final

"Análisis Integral de Salud y Condición Física: De Datos a Decisiones"

Profesor: Gustavo Benítez

Tutores: Hugo Mon, Ignacio Manuel Fernández

Alumno: Mauricio Horacio Zenere

Comisión: 61220

Dataset: https://www.kaggle.com/datasets/kukuroo3/body-performance-data

Abstract

Este proyecto de Ciencia de Datos se enfoca en un conjunto de datos que abarca aspectos relacionados con la salud, la condición física y las características de individuos. Con variables que van desde la edad hasta la capacidad de salto en largo, junto con una clasificación en clases estratificadas (A, B, C y D), el objetivo es realizar un análisis exploratorio de datos (EDA) exhaustivo, identificar patrones, generar hipótesis y proponer visualizaciones que proporcionen insights valiosos. Además, se explorarán posibles objetivos comerciales, y se considerará el contexto analítico y el problema comercial subyacente.

Este nombre del Proyecto refleja la amplitud del mismo, el que abarca aspectos relacionados con la salud, la condición física y la toma de decisiones basadas en datos. También indica la orientación hacia la comprensión completa de la información disponible y su aplicación en la toma de decisiones informadas. Por lo que es posible que sea ampliado o que surjan nuevos planteos durante el desarrollo del mismo.

Preguntas de Investigación:

1. ¿Cuál es la distribución de la población en las clases estratificadas (A, B, C y D)?
2. ¿Existe una correlación entre la edad y la clasificación?
3. ¿Influye el género en la capacidad de hacer abdominales?
4. ¿Qué factores están relacionados con la presión sanguínea sistólica y diastólica?
5. ¿Cuál es la relación entre la fuerza de agarre y la capacidad de salto en largo?

Hipótesis:

1. La edad está inversamente relacionada con la clasificación; los individuos más jóvenes tienden a estar en clases superiores.
2. El género influye en la capacidad de hacer abdominales, con una tendencia hacia un mejor rendimiento en hombres.
3. Se espera que haya una correlación positiva entre la fuerza de agarre y la capacidad de salto en largo.
4. La presión sanguínea sistólica podría estar relacionada con el índice de grasa corporal.

Análisis Exploratorio de Datos (EDA)

El EDA se centrará en:

- Resumen estadístico de las variables clave.
- Visualización de distribuciones de edad, altura, peso, etc.
- Gráficos de dispersión para explorar relaciones entre variables.
- Boxplots para identificar valores atípicos.
- Análisis de correlación y mapas de calor.

Propuesta de Visualizaciones para la primera etapa del trabajo

1. Histogramas de edad, altura y peso.
2. Diagramas de dispersión de presión sanguínea sistólica vs. diastólica.
3. Boxplot de edad vs. clasificación.
4. Mapa de calor de correlación entre variables numéricas.
5. Gráfico 3D

Insights Potenciales

1. Los individuos más jóvenes tienden a estar en clases superiores (hipótesis 1).
2. Los hombres tienden a tener un mejor rendimiento en la capacidad de hacer abdominales (hipótesis 2).
3. Existe una relación positiva entre la fuerza de agarre y la capacidad de salto en largo (hipótesis 3).
4. Se identificaron correlaciones significativas entre la presión sanguínea y el índice de grasa corporal (hipótesis 4).

Objetivos Comerciales

1. Los gimnasios pueden utilizar los resultados para personalizar programas de entrenamiento.
2. Las compañías de seguros pueden ajustar las primas según la salud y condición física de los asegurados.
3. Las clínicas de salud pueden identificar factores de riesgo en función de la presión sanguínea y el índice de grasa corporal.

Contexto Analítico y Problema Comercial

El contexto analítico involucra la recopilación de datos sobre la salud y la condición física de las personas, con el objetivo de comprender las relaciones y los factores que influyen en la clasificación. El problema comercial subyacente radica en cómo utilizar esta información para tomar decisiones informadas en campos como el fitness, la atención médica y los seguros. Identificar patrones y relaciones puede conducir a estrategias comerciales más efectivas y al mejoramiento de la calidad de vida de las personas.

Storytelling: Análisis de Clases en Función de Características Físicas

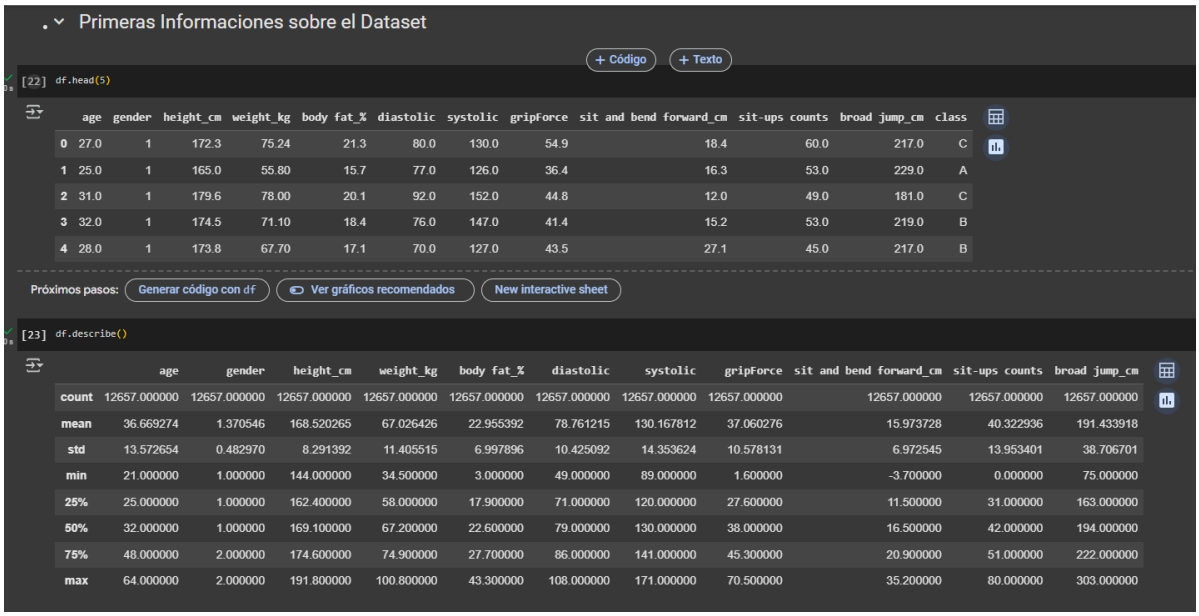
Inicio del Proyecto: Contexto y Objetivos

En mi proyecto, el desafío inicial fue construir un modelo que pudiera predecir la "clase" a la que pertenece una persona basándose en características físicas como su edad, género, altura y peso. Este análisis se origina en la necesidad de ayudar a servicios de salud, gimnasios y lugares de entrenamiento a personalizar sus programas de salud y ejercicio, ofreciendo recomendaciones basadas en la predicción de una clase que podría corresponder a distintas categorías de salud o condición física.

Paso 1: Preparación y Entendimiento de los Datos

El primer paso fue entender el dataset. El conjunto de datos contiene información de diferentes personas, en donde cada entrada incluye detalles sobre su edad, género, altura, peso y una categoría llamada "class", que en principio representaba su clase o grupo relacionado con su salud o condición física.

Con el objetivo de obtener los mejores resultados en el análisis, fue necesario revisar cuidadosamente la calidad de los datos. Comenzamos con el **data wrangling**, un proceso crucial que aseguraría que los datos estuvieran en el mejor estado posible para los modelos predictivos.



Primeras Informaciones sobre el Dataset

+ Código + Texto

```
[22] df.head(5)
```

	age	gender	height_cm	weight_kg	body fat_%	diastolic	systolic	gripForce	sit and bend forward_cm	sit-ups counts	broad jump_cm	class
0	27.0	1	172.3	75.24	21.3	80.0	130.0	54.9	18.4	60.0	217.0	C
1	25.0	1	165.0	55.80	15.7	77.0	126.0	36.4	16.3	53.0	229.0	A
2	31.0	1	179.6	78.00	20.1	92.0	152.0	44.8	12.0	49.0	181.0	C
3	32.0	1	174.5	71.10	18.4	76.0	147.0	41.4	15.2	53.0	219.0	B
4	28.0	1	173.8	67.70	17.1	70.0	127.0	43.5	27.1	45.0	217.0	B

Próximos pasos: Generar código con df | Ver gráficos recomendados | New interactive sheet

```
[23] df.describe()
```

	age	gender	height_cm	weight_kg	body fat_%	diastolic	systolic	gripForce	sit and bend forward_cm	sit-ups counts	broad jump_cm
count	12657.000000	12657.000000	12657.000000	12657.000000	12657.000000	12657.000000	12657.000000	12657.000000	12657.000000	12657.000000	12657.000000
mean	36.669274	1.370546	168.520265	67.026426	22.955392	78.761215	130.167812	37.060276	15.973728	40.322936	191.433918
std	13.572654	0.482970	8.291392	11.405515	6.997896	10.425092	14.353624	10.578131	6.972545	13.953401	38.706701
min	21.000000	1.000000	144.000000	34.500000	3.000000	49.000000	89.000000	1.600000	-3.700000	0.000000	75.000000
25%	25.000000	1.000000	162.400000	58.000000	17.900000	71.000000	120.000000	27.600000	11.500000	31.000000	163.000000
50%	32.000000	1.000000	169.100000	67.200000	22.600000	79.000000	130.000000	38.000000	16.500000	42.000000	194.000000
75%	48.000000	2.000000	174.600000	74.900000	27.700000	86.000000	141.000000	45.300000	20.900000	51.000000	222.000000
max	64.000000	2.000000	191.800000	100.800000	43.300000	108.000000	171.000000	70.500000	35.200000	80.000000	303.000000

Paso 2: Data Wrangling - Limpieza y Preparación de los Datos

Tratamiento de Valores Faltantes: El primer desafío encontrado en el dataset fueron los valores faltantes, algo común en muchos conjuntos de datos reales, por suerte en este como se puede observar en código, no hubo faltantes, por lo que no fue necesario por ejemplo, proceder a eliminar filas con valores faltantes.

Revisión de Outliers: Otro paso importante en la fase de wrangling fue la detección de outliers (valores atípicos) en las variables numéricas como 'age' (edad), 'height_cm' (altura) y 'weight_kg' (peso). Los outliers pueden distorsionar el análisis y la construcción de modelos predictivos, por lo que era fundamental identificarlos. Utilizamos gráficos como los box plots para visualizar estos valores atípicos. Luego se procedió a eliminarlos, a fin de lograr una mejor calidad de datos.

Paso 3: Análisis Univariado, Bivariado y Multivariado

Una vez que el conjunto de datos estaba limpio, era importante realizar un análisis exploratorio para entender mejor las relaciones entre las diferentes características.

Análisis Univariado: El análisis univariado nos permitió estudiar cada variable por separado. Utilizamos gráficos como histogramas y boxplots para visualizar la distribución de las variables individuales, como la edad, altura y peso. Esto nos ayudó a detectar la dispersión de los datos, identificar posibles sesgos y detectar si las variables seguían una distribución normal.

Análisis Bivariado: En el análisis bivariado, tomamos dos casos: Primero la relación entre la presión sistólica y la presión diastólica: este análisis permite ver si existe alguna correlación o patrón entre la presión sistólica y diastólica. Las observaciones agrupan de cierta manera y siguen una tendencia, lo que podría indicar que estas dos variables están relacionadas. En segundo lugar, se utilizó boxplot, uno de los ejes representa la columna Class (Clase) mientras que en el otro, la variable Edad (Age), lo que nos permite ver cómo se distribuye cada edad en cada clase.

Ambos gráficos son ejemplos de análisis bivariado, pero con enfoques ligeramente diferentes:

- El diagrama de dispersión se usa para visualizar la relación entre dos variables numéricas.
- El boxplot se utiliza para comparar la distribución de una variable numérica frente a una variable categórica.

Análisis Multivariado: Finalmente, realizamos un análisis multivariado donde exploramos cómo se relacionaban múltiples variables entre sí. Usamos gráficos de dispersión en 3D para visualizar las relaciones entre edad, peso y class. Así como también un Mapa de Calor de Correlación entres las variables.

Conclusión del Data Wrangling

El proceso de data wrangling fue necesario para poder luego enfocarnos de lleno en la parte de modelado. Al tratar los valores faltantes, detectar y manejar los outliers, logramos un conjunto de datos limpio y confiable. Además, los análisis univariado, bivariado y multivariado nos proporcionaron una visión más profunda de las relaciones entre las variables y las posibles correlaciones, lo cual es fundamental para la interpretación y la posterior fase de modelado.

Con los datos listos, el siguiente paso será aplicar los modelos de predicción y evaluar su rendimiento. Pero todo este trabajo de preparación nos ha dejado en una posición fuerte para obtener resultados más precisos y útiles en la etapa de análisis predictivo.

Modelado Predictivo

Estandarización de Datos

Para garantizar que los modelos predictivos trabajen de manera óptima, las variables predictoras (age, gender, height_cm, weight_kg) fueron estandarizadas utilizando StandardScaler. Este proceso convierte los datos a una distribución con media cero y desviación estándar unitaria, eliminando problemas asociados con diferentes escalas en las variables.

Implementación de Modelos

En una primera instancia se utilizó la Regresión Lineal y el RanfomForest, y se obtuvieron resultados buenos, pero mejorables como se puede observar en el Notebook del proyecto, pero igualmente se intentó ir un poco más allá y probar con modelos como el Rigde, Lasso y Gradient Booter Regresor, atento que son modelos, con una robustez frente al overfitting, ofrecían una mejor intepretabilidad, consistencia y enfoque en el objetivo del proyecto.

Validación y Evaluación - Puesta a Prueba

Regresión Ridge

Este modelo utiliza penalización L2 para reducir la complejidad del modelo y prevenir el overfitting. Se entrenó con los datos escalados y se observó su rendimiento en términos de consistencia y estabilidad.

Regresión Lasso

Este modelo emplea penalización L1, lo que permite seleccionar automáticamente las características más relevantes al establecer a cero los coeficientes de las variables menos importantes. Es especialmente útil para interpretar el impacto de las variables predictoras.

Asimismo, respecto del **Gradient Booster Regresor,** no se lo consideró para la puesta a prueba del modelo, luego de ver los resultados que se obtuvieron del entrenamiento del modelo; haciendo suponer con su RMSE muy elevado, que posiblemente podría estar sobreajustado.

```
Gradient Boosting Regressor - RMSE: 1898.6587720643151
Gradient Boosting Regressor - R²: 0.9998769916412803
Regresión Ridge - RMSE: 15.977091119542873
Regresión Ridge - R²: 0.9999999912896379
Regresión Lasso - RMSE: 7.722893526390291
Regresión Lasso - R²: 0.9999999979648291
```

Por su parte, también se aplicaron herramientas como GridSearchCV con la finalidad de lograr una validación cruzada y un ajuste de hiperparámetros.

```
Mejores parámetros para Ridge: {'alpha': 10}
Mejor RMSE para Ridge: 1.06
Mejores parámetros para Lasso: {'alpha': 0.01}
Mejor RMSE para Lasso: 1.06
```

Se evaluaron los modelos mediante las métricas de error cuadrático medio (RMSE) y coeficiente de determinación (R^2). Los resultados reflejan un buen ajuste de los modelos a los datos disponibles.

3.4 Funcionalidad de Predicción

Se creó una función interactiva que permite al usuario ingresar sus datos y recibir predicciones basadas en los modelos entrenados. La función convierte las entradas a un formato estandarizado, realiza las predicciones y las convierte en clases legibles para el usuario.

Además, se realizaron pruebas con casos específicos, como la entrada de un usuario con las siguientes características:

- Edad: 44 años.
- Género: Masculino.
- Altura: 190 cm.
- Peso: 100 kg.

Ambos modelos predijeron consistentemente la **Clase C**, lo que indica un alineamiento en los resultados.

CONCLUSIONES FINALES

Ambos modelos (Ridge y Lasso) predicen la Clase C para los mismos datos de entrada proporcionados.

Impacto de la Validación Cruzada y el Ajuste de Hiperparámetros:

La validación cruzada y la optimización de los parámetros alpha permitieron seleccionar configuraciones más robustas para los modelos, mejorando su capacidad de generalización.

Los resultados sugieren que el ajuste de hiperparámetros ha influido en cómo los modelos ponderan las variables predictoras (age, gender, height_cm, weight_kg), lo que podría haber llevado al cambio en la clase predicha.

Ambos modelos ahora coinciden en la predicción de la clase, lo que sugiere que las configuraciones optimizadas logran un enfoque similar en la evaluación de las características del usuario.

Evaluación Global:

Con un RMSE de 1.06, los modelos son precisos, aunque queda un pequeño margen de error (aproximadamente ±1 clase en promedio).

Los modelos optimizados tienen mayor probabilidad de ser fiables en un rango más amplio de datos.

Pasos a seguir - Futuras Aplicaciones

Uno de los primeros pasos sería optimizar aún más los modelos predictivos. Ya que si bien los datos obtenidos con la puesta a prueba, son los que se esperaban, se podría trabajar en la mejora de la calidad de los datos, realizando un análisis más profundo de las características que influyen en la predicción.

Un segundo paso crucial es poner en marcha el sistema en un entorno real. Esto podría implicar el desarrollo de una interfaz de usuario amigable donde los datos puedan ser ingresados fácilmente, y las predicciones se muestren de manera comprensible.

Esta funcionalidad puede ser expandida para generar reportes personalizados sobre el estado físico de una persona, basados en su clase predicha. Además, podrías crear recomendaciones automáticas que indiquen posibles mejoras en su salud, ya sea a través de un plan nutricional, ejercicios personalizados o pautas de bienestar.

En el ámbito de servicios de salud, gimnasios y centros de entrenamiento, este sistema podría ser integrado para ayudar a entrenadores y especialistas a evaluar rápidamente a los usuarios. La predicción de la clase proporcionaría una base sobre la cual se pueden sugerir entrenamientos específicos, ajustar dietas y realizar un seguimiento de los progresos en función de sus características físicas. Por ejemplo, alguien que esté en la "Clase D" podría recibir un plan especializado para reducir peso o mejorar su capacidad cardiovascular.

Finalmente, el proyecto puede escalarse aún más al integrar una funcionalidad de retroalimentación en tiempo real. A medida que los usuarios mejoran su condición física o cambian sus hábitos, los datos se pueden actualizar en el sistema, permitiendo ajustes dinámicos en sus recomendaciones. Esto no solo personaliza las sugerencias, sino que también asegura que el sistema sea adaptable a las necesidades de cada persona a lo largo del tiempo.