Empleando las librerías *Lazypredict* y *Optuna* para analizar la regresión con el dataset *Alumnos beneficiarios del comedor universitario de la Universidad Nacional Hermilio Valdizán – 2023 (UNHEVAL)*

Lizbeth Estefany Cáceres Tacora

Mayo 2025

1 Introducción

En la situación actual, en la que la toma de decisiones fundamentada en datos se ha transformado en una práctica indispensable en varias áreas, la implementación de modelos predictivos se vuelve indispensable. La habilidad para anticipar comportamientos, tendencias o valores venideros basándose en datos pasados posibilita la optimización de recursos, la mejora de servicios y la anticipación de necesidades. En este contexto, los métodos de regresión son un instrumento potente para modelar vínculos entre variables y producir proyecciones numéricas.

El propósito de este informe es examinar un grupo de datos obtenidos de un portal de acceso público, utilizando técnicas de regresión supervisada. Este análisis tiene como objetivo valorar el rendimiento de dos tipos de regresión: simple y ridge.

Para facilitar la exploración inicial de los modelos, se emplea la biblioteca *lazypredict*, que permite ejecutar rápidamente múltiples algoritmos de regresión sin necesidad de configuraciones manuales extensas. Esta herramienta proporciona una visión comparativa del rendimiento de distintos modelos utilizando métricas estándar como el coeficiente de determinación (R^2) y el error cuadrático medio (RMSE).

Sin embargo, la simple comparación entre modelos no siempre es suficiente. La calidad de un modelo predictivo también depende de la adecuada selección de sus hiperparámetros. Por esta razón, se recurre a **Optuna**, una biblioteca de optimización bayesiana que permite ajustar automáticamente los hiperparámetros de los modelos con el fin de mejorar su rendimiento. En particular, se hace uso de Optuna para optimizar el hiperparámetro *alpha* en la regresión Ridge, un tipo de regresión lineal con regularización L2 que es especialmente útil cuando existen problemas de multicolinealidad entre las variables predictoras.

A través de este enfoque combinado —exploración automática de modelos y optimización de hiperparámetros— se busca no solo encontrar el modelo más eficaz, sino también entender cómo influye la regularización en la estabilidad y precisión de las predicciones. Este proceso es esencial cuando se trabaja con conjuntos de datos que presentan relaciones complejas o variables altamente correlacionadas.

2 Elección del Dataset

Para este análisis se seleccionó la base de datos de acceso libre titulada *"Alumnos beneficiarios del comedor universitario de la Universidad Nacional Hermilio Valdizán – 2023 (UNHEVAL)"*, debido a que contiene información relevante y estructurada sobre variables relacionadas con el consumo alimentario de los estudiantes universitarios. Esta base de datos ofrece un contexto real que permite evaluar el comportamiento de distintos modelos de regresión, facilitando así la comprensión del funcionamiento y la aplicabilidad de herramientas de aprendizaje automático.

Es importante destacar que no todas las variables contenidas en el dataset inicial resultaban relevantes para la finalidad de este estudio. Algunas de estas no tenían una conexión directa con las actividades de predicción, por lo que se llevó a cabo un proceso de depuración y transformación de datos con el objetivo de elaborar un subconjunto más apropiado para el estudio. Las siguientes fueron las acciones clave llevadas a cabo durante esta etapa de limpieza y preprocesamiento:

- **Conversión de la fecha de nacimiento a formato `datetime`:** Se transformó la columna correspondiente con el fin de facilitar su uso posterior, especialmente para el cálculo de la edad.
- **Cálculo de la edad al 31 de diciembre de 2023:** Como se detectaron discrepancias entre la edad registrada y el año de nacimiento, se decidió determinar una edad homogénea a partir de la fecha de nacimiento, tomando como punto de referencia el final del año 2023. Esta transformación posibilita disminuir eventuales prejuicios y garantizar una mayor consistencia en los datos.
- **Eliminación de columnas irrelevantes o redundantes:** Se eliminaron las variables que no proporcionaban datos relevantes para el análisis predictivo, lo que facilitó la disminución de la dimensionalidad del dataset y enfocó la investigación en variables realmente valiosas para el modelo de regresión.
- **Revisión de valores nulos o atípicos en las variables de raciones servidas:** Se examino y depuró los registros que contenían valores ausentes o incongruentes en las columnas relacionadas con las comidas (desayuno, almuerzo, cena), con el objetivo de asegurar la confiabilidad del análisis y la adecuada utilización de los modelos.

Este proceso de preparación resultó en un conjunto de datos depurado, consistente y centrado en variables pertinentes, idóneo para emplearse en modelos de regresión lineal y para valorar el efecto de la regularización y optimización de hiperparámetros en el rendimiento predictivo.

3 Uso de la librería *Lazypredict*

Una vez realizado el proceso de limpieza y depuración del dataset, se procedió a su preparación para su análisis mediante la librería *Lazypredict*. Para ello, se seleccionaron únicamente las variables numéricas que aportan valor al modelo de regresión, considerando tanto la variable dependiente como las variables independientes con potencial explicativo.

Las variables utilizadas fueron definidas de la siguiente manera:

- **Variable dependiente:** Cantidad de raciones de almuerzo (*N RA ALMUERZO*)
- **Variables independientes:**
  - Edad
  - Año de la primera matrícula (*ANIO MATRICULA*)
  - Cantidad de raciones de desayuno (*N RAC DESAYUNO*)
  - Cantidad de raciones de cena (*N RAC CENA*)

Una vez establecidas las variables que representarían tanto la variable objetivo como las predictoras, se implementó el código correspondiente para aplicar la librería *Lazypredict*. Esta herramienta permite comparar rápidamente el desempeño de distintos modelos de regresión sobre el mismo conjunto de datos, proporcionando métricas como el R^2 y el error cuadrático medio (RMSE), lo que facilita la identificación de los modelos con mejor ajuste.

El código empleado para esta etapa se presenta a continuación:

```
import pandas as pd; from sklearn.model_selection import train_test_split; from
    lazypredict.Supervised import LazyRegressor
df = pd.read_csv("beneficiarios_comedor_2023_unheval_limpio_final.csv")[lambda x: x['
    N_RAC_ALMUERZO'] > 0]
X, y = df[['EDAD', 'ANIO_MAT1', 'N_RAC_DESAYUNO', 'N_RAC_CENA']], df['N_RAC_ALMUERZO']
models, _ = LazyRegressor(verbose=0).fit(*train_test_split(X, y, test_size=0.2,
    random_state=42))
print(f"Modelos:\n{models}")
```

Listing 1: Uso de la libreria Lazypredict

Donde los resultados optenidos fueron los siguientes:

Resultados de Modelos

Modelo	R^2	RMSE	Tiempo
LinearRegression	0.71	3.03	0.01 s

Cada uno de estos valores se interpretan de la sigueinte forma :

- R-Squared ($R^2 = 0.71$): Esto significa que el 71% de la variabilidad en las raciones de almuerzo ($N\_RAC\_ALMUERZO$) se explica por el modelo lineal usando las variables EDAD,$ANIO\_MAT1$, $N\_RAC\_DESAYUNO$ y N_RAC_CENA.. Es un valor aceptable, indica una buena relación lineal entre las variables predictoras y la variable objetivo. Sin embargo, el 29% restante de la variabilidad no es explicada por el modelo, lo que puede deberse a otros factores no incluidos, errores de medición, o relaciones no lineales.
- RMSE (Raíz del Error Cuadrático Medio = 3.03): En promedio, el modelo se equivoca por unas 3 raciones al predecir $N\_RAC\_ALMUERZO$. El RMSE debe evaluarse en el contexto del rango de $N\_RAC\_ALMUERZO$. Por ejemplo: Si los valores típicos están entre 0 y 10, entonces un error de 3 puede ser alto. Si están entre 0 y 50, un RMSE de 3 puede ser bajo.
- Tiempo de ejecución (0.01 s): Extremadamente rápido. Ideal para producción o para ejecutar muchas predicciones al instante.

Por otro lado tenemos que los modelos que mejor se ajsuatan son

Resultados de Modelos

Modelo	R^2	RMSE	Tiempo
KNeighborsRegressor	0.87	2.03	0.01 s
HistGradientBoosting	0.80	2.53	0.22 s
LGBMRegressor	0.80	2.53	0.10 s
ExtraTreesRegressor	0.80	2.55	0.23 s

4 Uso de la libreria Optuna

La regresión lineal simple (LinearRegression de sklearn) no tiene hiperparámetros que valga la pena optimizar con Optuna. Es decir: No tiene un parámetro como alpha, *max_depth*, n_estimators, etc., que pueda ajustarse. Solo ajusta una línea recta a los datos. No puedes "afinarla" más allá de los datos que le das. Por eso, no tiene sentido usar Optuna para buscar los "mejores hiperparámetros" de un modelo que no tiene ninguno relevante que optimizar.

```python
import optuna, numpy as np, pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import cross_val_score, KFold

df = pd.read_csv("beneficiarios_comedor_2023_unheval_limpio_final.csv")[lambda x: x['N_RAC_ALMUERZO'] > 0]
X, y = df[['EDAD', 'ANIO_MAT1', 'N_RAC_DESAYUNO', 'N_RAC_CENA']], df['N_RAC_ALMUERZO']

def objective(trial): return -cross_val_score(LinearRegression(), X, y, cv=KFold(5, shuffle=True, random_state=42), scoring='neg_root_mean_squared_error').mean()

study = optuna.create_study(direction="minimize").optimize(objective, n_trials=30)
print(f"Mejor RMSE: {study.best_value:.4f}\nMejores hiperparametros: {study.best_params}")
```

Listing 2: Ejemplo de código en Python

Indica que Optuna llevó a cabo un ensayo (Trial) buscando identificar los hiperparámetros más adecuados para el modelo, obteniendo un resultado de RMSE = 2.485... (error cuadrático medio raíz). No obstante, todos los parameters: se encuentran en blanco.

5 Uso de ridge con Optuna

En este escenario se aplicó la regresión Ridge, una modalidad de regresión lineal optimizada a través de regularización, también denominada regresión con penalización L2. Este procedimiento resulta especialmente beneficioso cuando se busca potenciar la estabilidad del modelo y su habilidad para generalizarse.

Uno de los mayores beneficios de Ridge radica en su habilidad para gestionar la multicolinealidad. Cuando existen altas correlaciones entre las variables predictivas, los coeficientes logrados mediante una regresión lineal convencional pueden tornarse inestables, lo que disminuye la confiabilidad de las predicciones. Ridge resuelve este inconveniente al disminuir la cantidad de coeficientes, estabilizando de esta manera el modelo.

Un motivo más para emplear Ridge es que contribuye a disminuir el sobreajuste (overfitting). En grupos de datos reducidos o ruidosos, es habitual que los modelos se adecuen excesivamente a los datos de entrenamiento, lo que reduce su capacidad para generalizar. La penalización L2 de Ridge funciona restringiendo los coeficientes de gran magnitud, previniendo un ajuste excesivo y optimizando el desempeño en datos recientes.

Adicionalmente, Ridge dispone de un hiperparámetro ajustable denominado alpha, que regula la intensidad de la regularización. Esto posibilita ajustar el modelo de acuerdo a las particularidades del conjunto de datos. Herramientas automáticas de optimización como Optuna pueden emplearse para determinar el valor ideal de alpha y de esta manera optimizar el desempeño del modelo.

5.1 ¿Cuándo es recomendable usar Ridge?

- Cuando se tiene un número elevado de variables predictoras.
- Cuando algunas de esas variables están correlacionadas entre sí.
- Cuando se busca mejorar la capacidad de generalización del modelo, evitando el sobreajuste.

```python
import pandas as pd; from sklearn.model_selection import train_test_split, KFold, cross_val_score
from lazypredict.Supervised import LazyRegressor; from sklearn.linear_model import Ridge; import optuna

df = pd.read_csv("beneficiarios_comedor_2023_unheval_limpio_final.csv").query('N_RAC_ALMUERZO > 0')
X, y = df[['EDAD', 'ANIO_MAT1', 'N_RAC_DESAYUNO', 'N_RAC_CENA']], df['N_RAC_ALMUERZO']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

print("Modelos evaluados:\n", LazyRegressor(verbose=0, ignore_warnings=True).fit(X_train, X_test, y_train, y_test)[0])

def objective(trial): return -cross_val_score(Ridge(alpha=trial.suggest_float('alpha', 1e-4, 100.0, log=True)), X, y, cv=KFold(5, shuffle=True, random_state=42), scoring='neg_root_mean_squared_error').mean()

study = optuna.create_study(direction="minimize").optimize(objective, n_trials=30)
print("\nOptuna - Ridge:\n Mejor alpha:", study.best_params, "\n Mejor RMSE:", study.best_value)
```

Listing 3: Uso del Optuna junto con el modelo Ridge

Datos obtenidos del modelo Ridge

Modelo	R^2	RMSE	Tiempo
Ridge	0.71	3.02	0.03 s

- R-Squared: 72% de la variabilidad en la variable objetivo ($N\_RAC\_ALMUERZO$) puede explicarse por las variables independientes (EDAD, $ANIO\_MAT1$, $N\_RAC\_DESAYUNO$, N_RAC_CENA) Esto indica un buen ajuste del modelo.
- RMSE (Root Mean Squared Error): el error promedio de predicción es de aproximadamente 3 raciones de almuerzo. Esto te indica cuán lejos están las predicciones, en promedio, de los valores reales.
- Time Take: el modelo se entrenó muy rápidamente — solo 0.03 segundos, lo que sugiere eficiencia computacional.

6 Resultados optenidos

Se utilizó Optuna para optimizar el valor del hiperparámetro alpha en un modelo de regresión Ridge, con el objetivo de minimizar el RMSE (Root Mean Squared Error). Este indicador mide el error promedio entre las predicciones del modelo y los valores reales observados, siendo más sensible a errores grandes. El parámetro alpha controla el grado de regularización en la regresión Ridge: valores más altos imponen una penalización mayor a los coeficientes del modelo, ayudando así a prevenir el sobreajuste y mejorar la generalización. En este caso, Optuna encontró el valor óptimo de alpha que proporciona el menor RMSE en el conjunto de validación.

Resultado Obtenido

Mejor valor de α: 71.32
Mejor RMSE: 2.49

Esto significa que el mejor modelo Ridge que Optuna encontró tiene una penalización moderadamente alta (valor de alpha grande) y que predice la cantidad de almuerzos servidos con un error promedio de aproximadamente 2.49 raciones.

- El valor óptimo de alpha es alto (71.32), lo cual indica que el modelo se beneficia de una fuerte regularización. Esto puede deberse a que algunas variables están correlacionadas (como $N\_RAC\_DESAYUNO$ y N_RAC_CENA con $N\_RAC\_ALMUERZO$), y el modelo necesita evitar darles demasiado peso.
- El RMSE de 2.485 significa que, en promedio, el modelo se equivoca en aproximadamente 2.5 almuerzos cuando predice el número de almuerzos servidos. Dependiendo del rango total de $N\_RAC\_ALMUERZO$, este error puede considerarse pequeño o no (por ejemplo, si los valores típicos son entre 20 y 100, sería aceptable).
- Los valores de alpha muy pequeños (por ejemplo, cerca de 0) dieron peores resultados, lo cual confirma que sin regularización el modelo no generaliza tan bien.

7 Conclusión

Durante este estudio, se examinó la aplicación de métodos de regresión y optimización de modelos en un dataset auténtico, empleando herramientas como Lazypredict y Optuna. Mediante la purificación y elección de variables pertinentes del dataset de beneficiarios del comedor universitario de la UNHEVAL, se pudo elaborar modelos predictivos capaces de calcular el número de raciones de comida que un estudiante recibe.

Con Lazypredict, se detectaron modelos con un rendimiento óptimo, sobresaliendo el KNeighborsRegressor con un R2 de 0.87 y un RMSE de 2.03, lo que indica una predicción bastante precisa dentro del rango detectado. Estos hallazgos demuestran la eficacia de modelos más sofisticados en comparación con una regresión lineal sencilla.

En cambio, al emplear Optuna, se demostró que su eficacia varía dependiendo del tipo de modelo. Para LinearRegression, no se logró optimizar ninguna característica significativa debido a la falta de hiperparámetros. No obstante, al aplicar Ridge Regression, se utilizó la habilidad de Optuna para modificar el hiperparámetro alpha, lo que posibilitó incrementar la estabilidad y la capacidad de generalización del modelo ante desafíos como la multicolinealidad y el sobreajuste.

En conclusión, el trabajo demuestra que:

- La calidad de los datos y su adecuada preparación son fundamentales para construir modelos predictivos confiables.
- Herramientas automáticas como Lazypredict permiten evaluar rápidamente múltiples modelos y seleccionar los más prometedores.
- El uso de técnicas de regularización, como Ridge, y su ajuste mediante algoritmos de optimización como Optuna, resulta esencial cuando se trabaja con datos reales, especialmente si existen variables correlacionadas o se desea mejorar la robustez del modelo.
- Este análisis proporciona una base sólida para aplicar técnicas de regresión y optimización en contextos reales, y sugiere que combinar diferentes herramientas puede conducir a resultados más precisos y eficientes.