Data Challenge: Monster Hunt en el ITAM

Econometría Aplicada I

Andres Bermudez* Esteban Degetau† Andrea Rancaño‡

2023-11-08

Para resolver la tarea de clasificación de los *monstruos* que invadieron el ITAM, implementamos una serie de ejercicios de clasificación supervisados y no supervisados y seleccionamos el de mayor precisión y menor varianza. Este enfoque lo basamos en los ejercicios de clasificación de Thomas (2016). También buscamos replicar el ejercicio de Pérez Herrero (2017), pero su procedimiento nos resultó poco claro. Adicionalmente, consultamos el libro de Irizarry (2019) para guiar el entrenamiento supervisado.

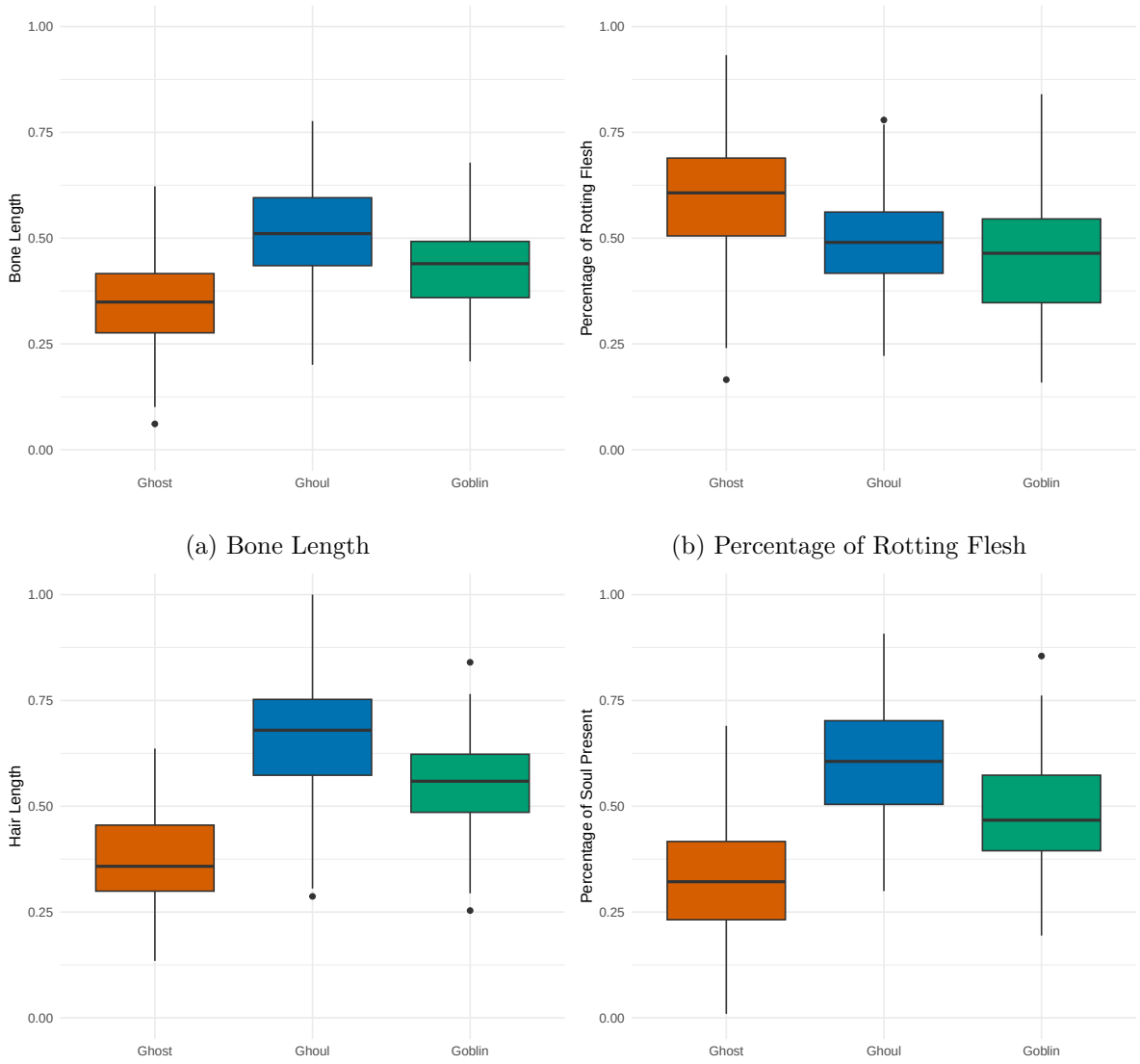
Esta nota está organizada de la siguiente manera. Primero presentamos un breve análisis exploratorio de los datos. Después, presentamos los ejercicios de entrenamiento no supervisado y supervisado. Finalmente, presentamos el procedimiento de creación de nuestro monstruo, Wenceslao.

Datos

En este apartado, seguimos el análisis de Thomas (2016) para darnos una idea de la distribución de los datos. La Figura 1 muestra que los monstruos que acechan al ITAM son muy similares a los que se estudiaron en Thomas (2016). Resalta que, cada una de las características de los monstruos no permite al lector categorizar por tipo a simple vista. Es decir, no hay una característica que *por sí misma* distinga a los monstruos de cada tipo.

La Figura 2 muestra que la distribución de colores por tipo de monstruo es bastante homogénea, por lo que no se puede clasificar con base en ella por sí misma.

*Contribuciones: Implementación independiente de entrenamiento por redes neuronales y Bayesian GLM.
†Contribuciones: Replicación de Thomas (2016); entrenamiento no supervisado; entrenamiento supervisado por Random Forest y GLM.
‡Contribuciones: Replicación de Pérez Herrero (2017); entrenamiento supervisado por KNN siguiendo a Irizarry (2019); creación de Wenceslao.



(a) Bone Length

(b) Percentage of Rotting Flesh

(c) Hair length

(d) Percentage of Soul Present

Figura 1: Diagramas de caja y brazo por tipo de monstruo

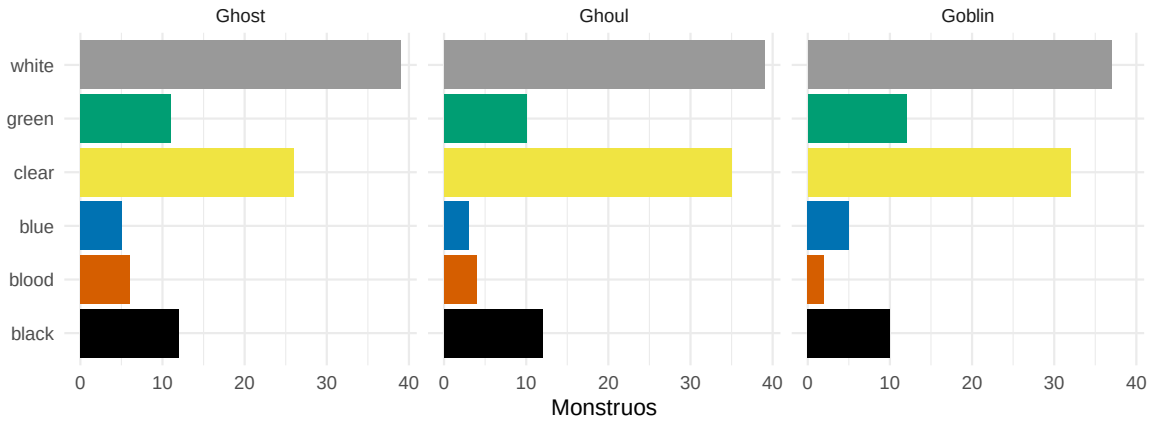


Figura 2: Distribución de colores por tipo de monstruo

Componentes principales

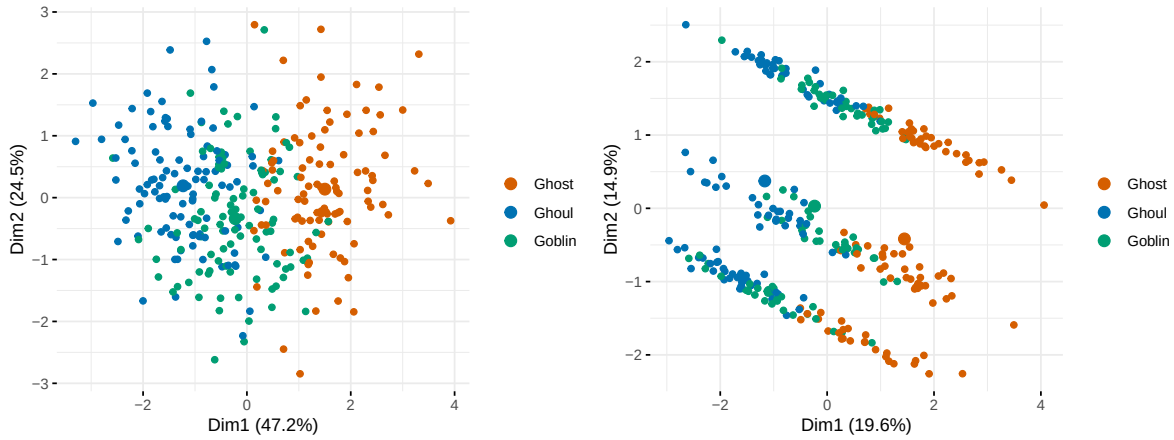
En esta sección expandimos el análisis de componentes de Thomas (2016) para indagar en la segmentación de los datos. En particular, buscamos encontrar qué combinación de variables permite segmentar de manera más eficiente a los monstruos. Consideramos cuatro combinaciones de variables:

a. Solo variables numéricas
b. Numéricas y dummies de color
c. Numéricas con interacciones
d. Todas las variables

La Figura 3 muestra la segmentación de los monstruos a lo largo de dos componentes principales, para cada uno de las combinaciones de variables consideradas. Encontramos que, al menos usando dos componentes principales, no se puede segmentar claramente a los monstruos bajo ninguna combinación de variables.

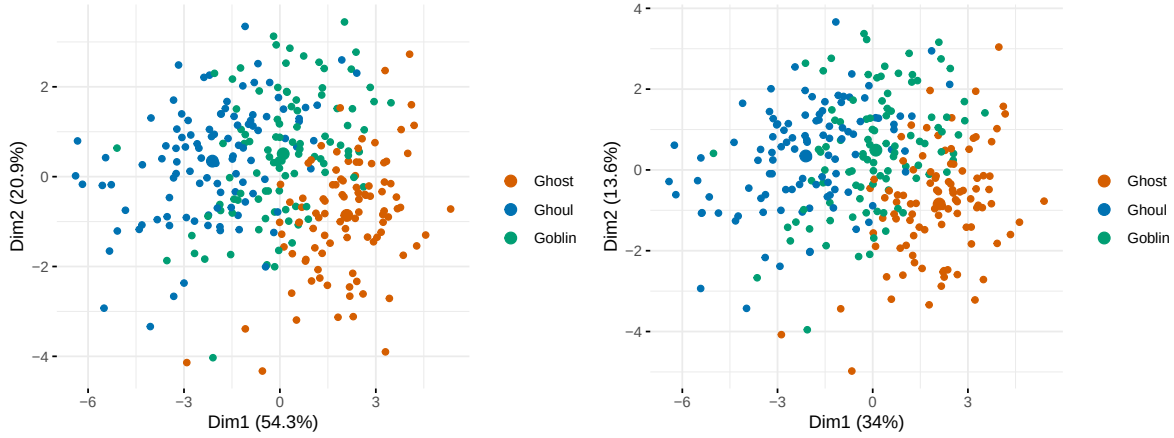
Entrenamiento no supervisado

Esta sección busca expandir el entrenamiento no supervisado en Thomas (2016). Ampliamos el análisis al considerar las cuatro combinaciones de variables en la sección anterior y 3 modelos no supervisados; *Hierarchical*, *K-means* y *PAM*. A cada uno de los modelos se le asignó la tarea de encontrar 3 categorías en los datos *sin incluir la categoría real*. El objetivo de este ejercicio es contar con un *benchmark* de precisión y varianza para la selección del mejor modelo de predicción.



(a) Solo variables numéricas

(b) Numéricas y dummies de color



(c) Numéricas con interacciones

(d) Todas las variables

Figura 3: Análisis de componentes principales

Tabla 1: Prueba de Hopkins

Combinación de variables	Hopkins
Numeric	0.96
Numeric + Color	1.00
Numeric + Interactions	1.00
All variables	1.00

El primer paso de todo entrenamiento no supervisado es determinar si los datos muestran agrupamiento. La Tabla 1 muestra resultados de la prueba de Hopkins cercanos a 1, que es consistente con datos altamente segmentados.

El siguiente paso en este ejercicio fue calcular la precisión y varianza de cada método en cada conjunto de datos. Puesto que los modelos son determinísticos, i.e. siempre generan la misma segmentación para el mismo conjunto de datos, hicimos un ejercicio de remestrueo con 1,000 submuestras aleatorias con reemplazo del mismo tamaño de la muestra original de monstruos. El ejercicio de remuestreo nos permite calcular la precisión y varianza de cada modelo.

La Figura 4 muestra los resultados de remuestreo de los modelos no supervisados. En todos los modelos que consideramos, la inclusión de variables adicionales a las numéricas empeora la precisión sin mejorar la varianza de la predicción. Los modelos no supervisados producen una predicción correcta en un poco más del 70 por ciento de los monstruos en la muestra.

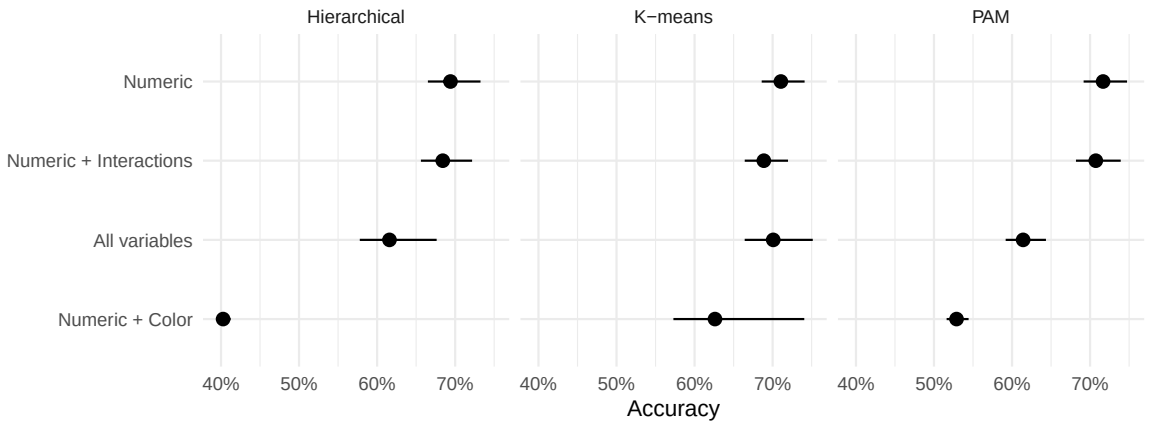


Figura 4: Resultados del remuestreo de entrenamiento no supervisado

La Tabla 2c muestra que el mejor modelo no supervisado segmenta bastante bien a los Ghost y Ghoul, pero confunde a los Goblin con Ghoul. Este resultado nos ayudará para la creación de nuestro monstruo.

Tabla 2: Matriz de confusión para los mejores modelos. Probabilidad de clasificación dada la categoría verdadera (%).

(a) GLM Net - Numeric + interactions

Predicted	Ghost	Ghoul	Goblin
Ghost	89.90	1.01	9.09
Ghoul	1.94	81.55	16.50
Goblin	19.39	24.49	56.12

(b) Neural Networks - Numeric

Predicted	Ghost	Ghoul	Goblin
Ghost	92.93	0.00	7.07
Ghoul	0.00	77.67	22.33
Goblin	12.24	22.45	65.31

(c) PAM - Numeric

Predicted	Ghost	Ghoul	Goblin
Ghost	80.81	0.97	9.18
Ghoul	1.01	80.58	25.51
Goblin	18.18	18.45	65.31

Entrenamiento supervisado

Los ejercicios anteriores sirvieron para guiar el entrenamiento supervisado de esta sección. Entrenamos modelos utilizando los métodos *Random Forest* y *GLM net*;[1] *K-Nearest Neighbors*;[2] así como *Neural Networks*, *Averaged Neural Networks* y *Bayesian GLM*.

Para encontrar el modelo óptimo, entrenamos un modelo con cada método en cada una de las cuatro combinaciones de variables utilizadas anteriormente. La ventaja de este procedimiento es que, a través del remuestreo, obtendremos una medida de precisión y varianza en cada modelo, y tendremos un repertorio amplio de dónde seleccionar el mejor modelo.

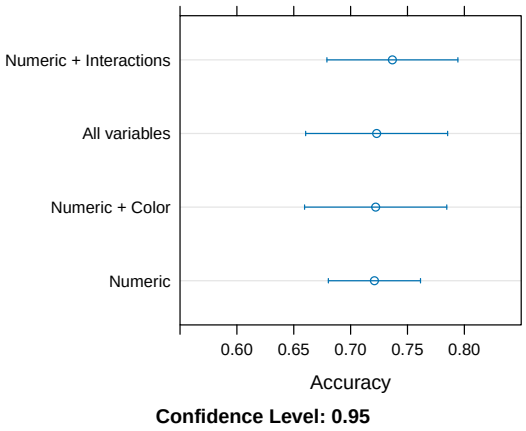
La Figura 5 muestra los resultados de los ejercicios de entrenamiento supervisado. Resalta que los mejores modelos son el entrenado con GLM net con variables de interacción, así como el modelo de redes neuronales entrenado con solo variables numéricas.

La Tabla 2 muestra la matriz de confusión para estos modelos. Decidimos utilizar el modelo de redes neuronales para la competencia porque es mejor en detectar a los Goblin que el modelo de GLM net, sin perder precisión. Esta decisión nos dará una ventaja al predecir a los monstruos que los otros equipos generen basándose en GLM Net.

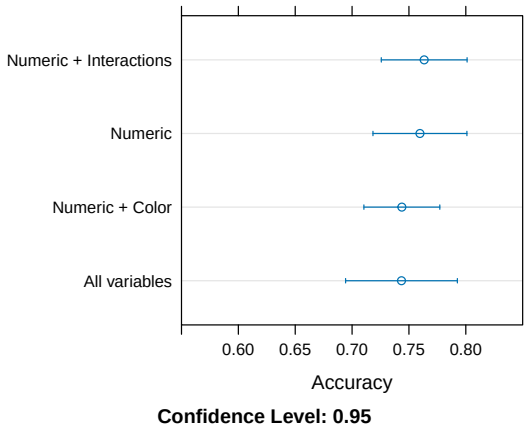
Creación del monstruo

Una vez que seleccionamos a nuestro modelo para el concurso, diseñamos a nuestro monstruo siguiendo el procedimiento descrito a continuación:

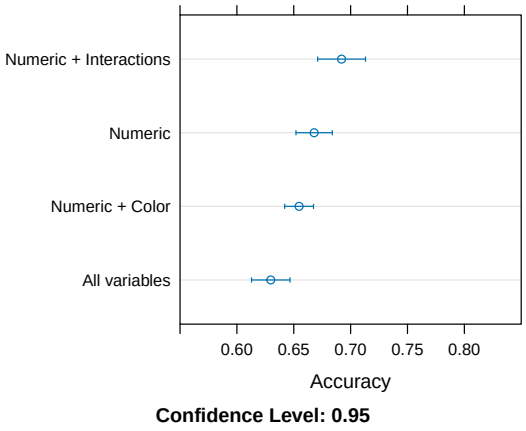
[1] Siguiendo a Thomas (2016).
[2] Siguiendo a Irizarry (2019).



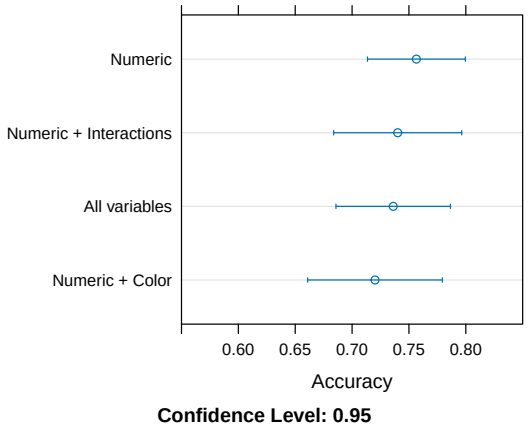
(a) Random Forest



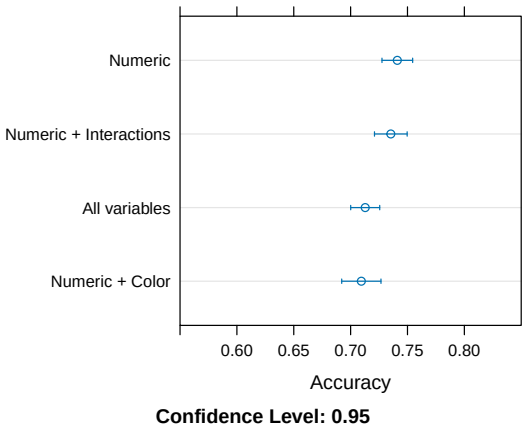
(b) GLM net



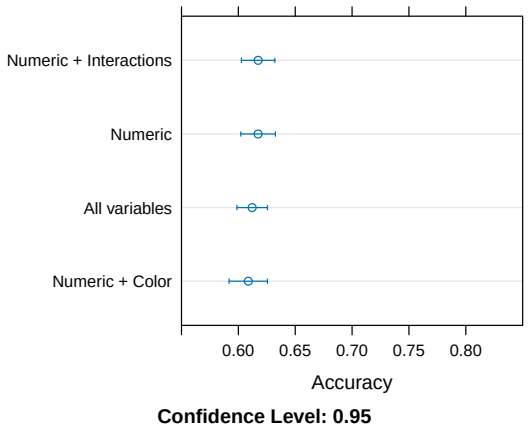
(c) Weighted KNN



(d) Neural Networks



(e) Averaged NN



(f) Bayesian GLM

Figura 5: Resultados del entrenamiento supervisado

1. Creamos un caldero de 100,000 *monstruos*, con características generadas de forma aleatoria siguiendo una distribución uniforme continua.
2. Predecimos la categoría del monstruo de acuerdo al modelo de redes neuronales entrenado en la sección anterior. Además, calculamos la confianza con la que se hizo la predicción.
3. Desechamos del caldero a todos los monstruos cuya categoría predicha no fuera Goblin. Escogimos esta categoría porque es la más difícil de detectar en los modelos de mayor precisión que encontramos.
4. Encontramos al Goblin cuya confianza de predicción fuera mínima. En nuestro caso, fue de 0.5090. De esta manera, maximizamos la incertidumbre de clasificación que enfrentarán los otros equipos, asegurando que nuestro modelo lo categorizará correctamente.
5. Bautizamos a nuestro monstruo como Wenceslao.

Referencias

Irizarry, Rafael A. 2019. *Introduction to Data Science: Data Analysis and Prediction Algorithms with R*. Boca Raton London New York.

Pérez Herrero, Enrique. 2017. «Hyperparameter Random Search with 'Mlr'». https://kaggle.com/code/enrique1500/hyperparameter-random-search-with-mlr.

Thomas, Amber. 2016. «Ghosts, Goblins, and Ghouls. Oh My!» https://kaggle.com/code/amberthomas/ghosts-goblins-and-ghouls-oh-my.