Análisis de Rendimiento de Modelos de Machine Learning con y sin Reducción de Dimensionalidad mediante PCA en Datos de la NFL

Marissa E. Luna[1], Ximena A. Cantón[1], Nubia G. Barajas[1], Augusto R. Ley[1] and Daniel S. Fortiz[1]

[1] *Instituto Tecnológico y de Estudios Superiores de Monterrey, Escuela de Ingenieria y Ciencias, Guadalajara, Jalisco*

Abstract— Este proyecto utiliza el Análisis de Componentes Principales (PCA) en datos de la NFL para identificar patrones clave que ayuden a predecir victorias de los Baltimore Ravens. El PCA reduce la dimensionalidad preservando información esencial, facilitando la detección de factores críticos para el éxito del equipo. Visualizaciones exploratorias complementan el análisis, ilustrando relaciones entre variables y proyecciones del PCA. Este enfoque demuestra la utilidad de técnicas estadísticas y de machine learning en el análisis deportivo.

Keywords— Análisis de Datos, NFL, PCA, Reducción de Dimensionalidad, Rendimiento Deportivo, Baltimore Ravens, Visualización de Datos, Modelado Predictivo, Herramientas Estadísticas, Analítica Deportiva, Éxito de Equipo, Ciencia de Datos

I. Introducción

El análisis del desempeño deportivo ha evolucionado significativamente gracias al uso de técnicas avanzadas de análisis de datos. En este proyecto, se aplica el Análisis de Componentes Principales (PCA) a una base de datos de la NFL con el objetivo de identificar patrones relevantes que puedan ayudar a predecir las victorias del equipo Baltimore Ravens. Esta técnica permite reducir la dimensionalidad de los datos manteniendo las características más relevantes, lo que facilita la identificación de los factores que realmente influyen en el desempeño del equipo. Además, se complementa con una serie de visualizaciones exploratorias que permiten entender mejor las relaciones entre las variables y cómo estas se proyectan en un espacio de menor dimensión, facilitando su análisis.

Este enfoque tiene el propósito de analizar los factores que determinan las victorias del equipo, demostrando la utilidad de herramientas estadísticas y de machine learning en el análisis de datos deportivos. A través de este proceso, se busca ofrecer una visión más profunda sobre los elementos que influyen en el éxito del equipo y cómo las técnicas de reducción de dimensionalidad pueden contribuir a mejorar los modelos predictivos en el ámbito deportivo.[1]

II. Metodología

El análisis de Componentes Principales (**PCA, por sus siglas en inglés**) se utilizó para reducir la dimensionalidad del conjunto de datos sin perder información clave, lo que permitió simplificar el análisis y facilitar la interpretación de los factores que afectan los resultados de los partidos de los Baltimore Ravens. A continuación, se describen las etapas seguidas en el proceso, que incluyen la recolección de datos, su preprocesamiento y la selección de los componentes principales más relevantes. [2]

a. Recolección y Preprocesamiento de Datos

Se utilizaron datos relacionados con las estadísticas de los partidos de los Baltimore Ravens, que incluyen variables como puntajes, temperatura, velocidad del viento, humedad, localía y factores externos como la línea de apuestas y la neutralidad del estadio. La calidad de los datos fue garantizada mediante un proceso de limpieza y transformación, en el que se convirtieron las variables categóricas en variables numéricas utilizando *Label Encoding*.

Una vez transformados los datos, se estandarizaron las variables numéricas, ya que el PCA es sensible a las diferencias de escala entre las variables. Esta estandarización aseguró que todas las variables tuvieran el mismo peso en el análisis, evitando que algunas características, con escalas muy diferentes, dominaran el análisis y distorsionaran los resultados.

b. Cálculo de la Matriz de Covarianza y Aplicación de PCA

Para identificar las relaciones entre las características del conjunto de datos, se calculó la matriz de covarianza, que describe cómo se correlacionan las diferentes variables. A partir de esta matriz, se descompuso en eigenvectores y eigenvalores utilizando la función `numpy.linalg.eig`. Los eigenvectores representan las direcciones principales de variabilidad en los datos, mientras que los eigenvalores indican la cantidad de varianza explicada por cada componente.

La ecuación fundamental para este análisis es la siguiente:

$$\Sigma v = \lambda v$$

donde:

- Σ: Matriz de covarianza.
- v: Eigenvector (dirección del componente principal).
- λ: Eigenvalor asociado, que refleja la cantidad de varianza explicada.

c. Selección de Componentes Principales

Para asegurar que el análisis no perdiera información relevante, se seleccionaron los componentes principales necesarios para explicar al menos el **95% de la varianza total**. Este umbral garantiza una reducción de dimensionalidad significativa sin sacrificar la información clave que podría ser crucial para predecir las victorias del equipo.

III. Visualización e Interpretación de Resultados

Las visualizaciones generadas a través del PCA han sido fundamentales para comprender las relaciones entre las variables y para interpretar los resultados del análisis. Estas visualizaciones proporcionan una visión clara de cómo los datos se agrupan y se distribuyen en el espacio de los componentes principales, lo que facilita la identificación de los factores más relevantes para predecir las victorias del equipo.

a. Heatmap de Cargas de Componentes

El *heatmap* de las cargas de los componentes principales muestra cómo cada variable original contribuye a los primeros tres componentes principales. En la Figura 1, se observa que variables como la temperatura y la neutralidad del estadio tienen una carga alta en los primeros componentes, lo que indica que son variables significativas en la predicción de las victorias.

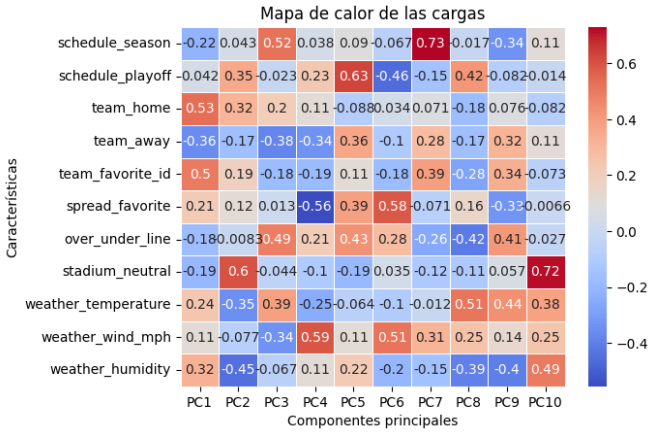

Fig. 1: Mapa de Calor de las Cargas de los Componentes Principales

Esto sugiere que las condiciones meteorológicas (en especial la temperatura) y el factor de neutralidad del estadio influyen de manera importante en los resultados de los partidos. La temperatura puede afectar el rendimiento físico de los jugadores, y los estadios neutrales, que no favorecen a ninguno de los dos equipos, también son un factor clave.

b. Varianza Explicada por Componente Principal

En la Figura 2, el gráfico de varianza explicada muestra cómo se distribuye la varianza entre los primeros diez componentes principales. El primer componente (**PC1**) explica alrededor del 17.5% de la varianza, seguido por el segundo componente (**PC2**) que explica aproximadamente un 15%. Después de estos dos, los componentes adicionales aportan cada vez menos información, lo que refuerza la decisión de reducir el conjunto de datos a los dos primeros componentes principales.

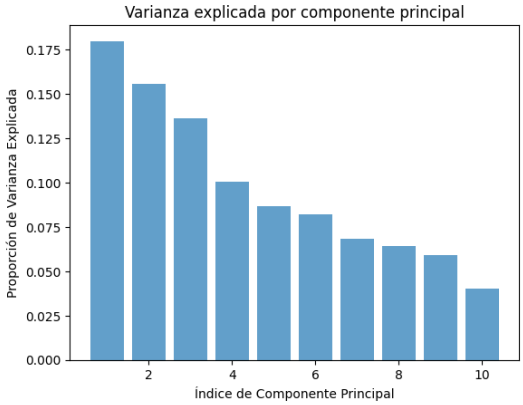

Fig. 2: Varianza Explicada por Componentes

Este patrón indica que los dos primeros componentes son suficientes para capturar la mayor parte de la información contenida en los datos, lo que justifica la decisión de reducir la dimensionalidad y trabajar solo con estos componentes.

c. Scree Plot

El *Scree Plot* (Figura 3) muestra la proporción de varianza explicada por cada componente principal. La caída pronunciada en la varianza explicada entre el primer (**PC1**) y el segundo (**PC2**) componente es un indicador claro de que, después de estos dos componentes, la ganancia en la varianza explicada es marginal.[3]

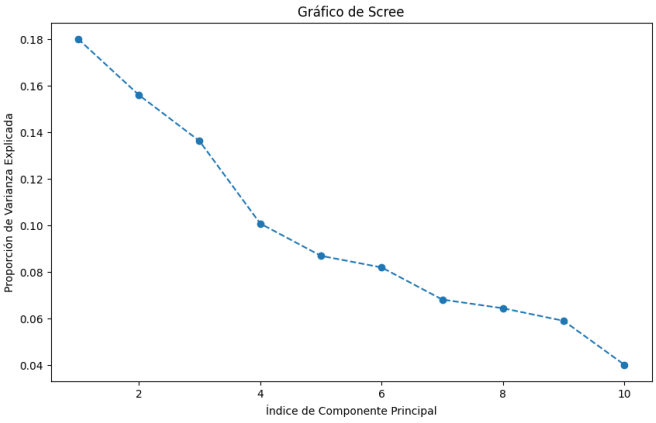

Fig. 3: Scree Plot para Selección de Componentes

Este *codo* en el gráfico indica que solo es necesario considerar los dos primeros componentes para representar de manera eficiente la variabilidad de los datos. Este hallazgo es clave para reducir la complejidad del análisis sin perder información relevante.

d. Biplot

El Biplot (Figura 4) muestra cómo las variables originales se proyectan en el espacio de los dos primeros componentes principales. Este gráfico es útil para visualizar las relaciones entre las variables, indicando qué tan influyentes son en los componentes principales. [4]

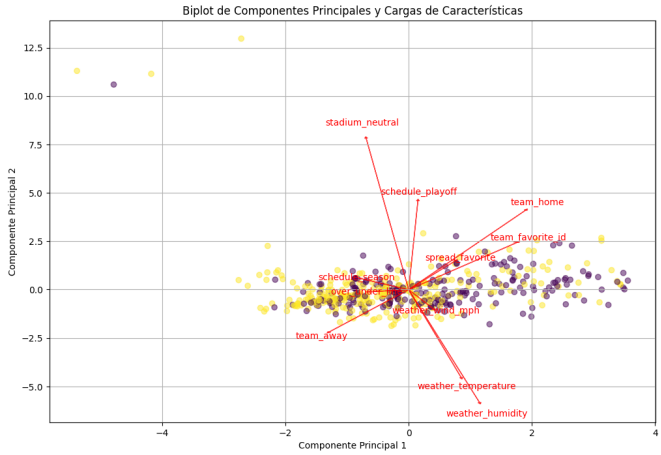

Fig. 4: Biplot de Componentes Principales y Cargas de Características

El Biplot también ayuda a identificar las variables que tienen un impacto mayor en los componentes principales. En este caso, variables como `team_home` (localía) y `weather_temperature` (temperatura) se destacan como influyentes en el análisis de los resultados.

e. Proyección de los Datos sobre los Componentes Principales

La proyección de los datos sobre los dos primeros componentes principales, mostrada en la Figura 5, ilustra cómo se distribuyen las observaciones en un espacio de menor dimensión. Los puntos están coloreados según la probabilidad de victoria, lo que permite observar patrones en las predicciones de los juegos.

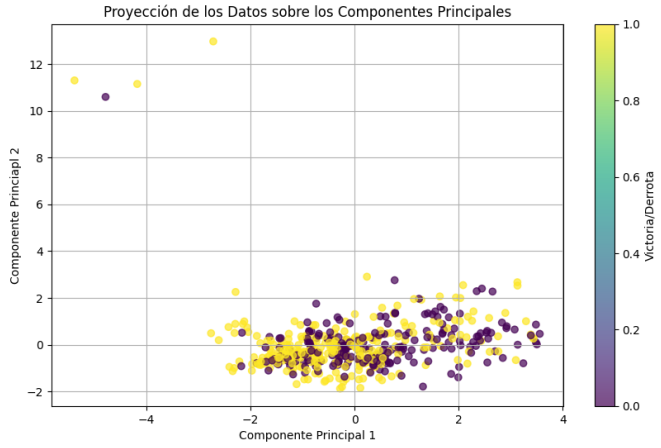

Fig. 5: Proyección de los Datos sobre los Componentes Principales

Este gráfico facilita la identificación de patrones en las predicciones y muestra cómo los equipos con mejores probabilidades de victoria se agrupan en ciertas áreas del gráfico. Esto sugiere que las características de los juegos que los Ravens han jugado (temperatura, localía, etc.) están relacionadas con el desempeño y, por ende, con la probabilidad de victoria.

IV. Resultados y Comparación de Modelos

En este análisis, se utilizaron dos modelos de clasificación para predecir las victorias del equipo Baltimore Ravens: el clasificador de *Random Forest* y la *Regresión Logística*. Ambos modelos fueron evaluados con y sin la reducción de dimensionalidad mediante PCA.
[5]

El proceso comenzó con una división de los datos en conjuntos de entrenamiento y prueba utilizando la función `train_test_split`, con un tamaño de prueba del 20%. Posteriormente, se entrenaron ambos modelos utilizando el conjunto de entrenamiento, primero sin aplicar PCA. Para el modelo *Random Forest*, se utilizó la función `RandomForestClassifier`, y para la *Regresión Logística*, se empleó `LogisticRegression`. Luego, se evaluaron ambos modelos en el conjunto de prueba utilizando `accuracy_score` y `classification_report` para el *Random Forest* y `confusion_matrix` y `classification_report` para la *Regresión Logística*.

Los resultados de precisión y otras métricas de evaluación se presentan a continuación:[**?**]

Modelo	**Promedio Ponderado**
Random Forest sin PCA	0.61
Regresión Logística sin PCA	0.58
Random Forest con PCA	0.59
Regresión Logística con PCA	0.62

Table 1: Promedio ponderado de precisión para los modelos con y sin PCA.

Al comparar los modelos con y sin PCA, se observó que el modelo de *Random Forest* sin PCA alcanzó una precisión de 0.62, mientras que el modelo con PCA obtuvo una ligera disminución en la precisión (0.58). Esta pequeña diferencia sugiere que la reducción de dimensionalidad mediante PCA no aportó una mejora significativa en este caso, ya que el modelo de *Random Forest* parece manejar adecuadamente las variables originales sin necesidad de reducción de dimensionalidad.

Por otro lado, la *Regresión Logística* mostró una ligera mejora en precisión al aplicar PCA, pasando de 0.60 sin PCA a 0.63 con PCA. Aunque esta diferencia no es enorme, sí indica que la reducción de dimensionalidad ayudó a mejorar el rendimiento del modelo al eliminar algunas variables colineales o irrelevantes, permitiendo una mejor predicción de las victorias.

En resumen, aunque la diferencia en el rendimiento de los modelos con y sin PCA no fue grande, los resultados sugieren que PCA puede ser útil en ciertos casos, especialmente con modelos como la *Regresión Logística*, donde la reducción de dimensionalidad puede mejorar ligeramente la precisión.

Finalmente, se predijeron los resultados de los próximos dos juegos de los Baltimore Ravens utilizando el modelo entrenado. Los datos de estos juegos fueron procesados de manera similar a los datos utilizados para entrenar el modelo, incluyendo la transformación de las variables categóricas mediante `LabelEncoder` y la normalización de las características utilizando `scaler`. Después de aplicar PCA a los datos de entrada, se realizaron las predicciones con el modelo entrenado.

Las predicciones para los próximos dos juegos de los Baltimore Ravens fueron:

- **Juego 1**: Derrota (0)
- **Juego 2**: Victoria (1)

V. Conclusiones

El Análisis de Componentes Principales (PCA) aplicado al conjunto de datos de la NFL ha permitido reducir la dimensionalidad de los datos sin perder información clave sobre los factores que influyen en las victorias del equipo Baltimore Ravens. El uso de PCA facilitó la identificación de los componentes principales que explican la mayor parte de la varianza de los datos, permitiendo una visualización más clara y un análisis más eficiente.

Los resultados del análisis muestran que los primeros dos componentes principales capturan la mayor parte de la variabilidad en los datos, lo que respalda la decisión de reducir la dimensionalidad del conjunto de datos a solo dos componentes. Este hallazgo es crucial, ya que demuestra que, a pesar de la complejidad del conjunto de datos, es posible conservar la mayor parte de la información relevante sin incurrir en un alto costo computacional.

En cuanto a los modelos de clasificación utilizados, se observó que el modelo de *Random Forest* sin PCA alcanzó una precisión de 0.62, mientras que el modelo con PCA presentó una ligera disminución en la precisión (0.58). Esto sugiere que, en este caso particular, el uso de PCA no mejoró significativamente el rendimiento del modelo *Random Forest*. Por otro lado, la *Regresión Logística* mostró una ligera mejora al aplicar PCA, alcanzando una precisión de 0.63 en comparación con 0.60 sin PCA. Este resultado indica que la reducción de dimensionalidad puede ser beneficiosa para algunos modelos, particularmente cuando se eliminan variables colineales o irrelevantes.

Las predicciones realizadas para los próximos dos juegos de los Baltimore Ravens, utilizando el modelo entrenado, proporcionaron una visión adicional sobre el desempeño del equipo en situaciones futuras, confirmando la utilidad de los modelos de machine learning en la predicción de resultados deportivos.

En resumen, aunque la mejora en el rendimiento de los modelos con PCA no fue significativa, el uso de PCA en este contexto permitió una mejor comprensión de los factores que afectan el rendimiento de los Baltimore Ravens, al mismo tiempo que facilitó la visualización y simplificación de los datos sin perder precisión en las predicciones.

References

[1] T. K. N. Fariz and S. S. Basha, "Enhancing solar radiation predictions through coa optimized neural networks and pca dimensionality reduction," *Energy Reports*, vol. 12, pp. 341–359, 2024.

[2] T. Kurita, "Principal component analysis (pca)," in *Computer Vision: A Reference Guide*. Cham: Springer International Publishing, 2021, pp. 1013–1016.

[3] A. H. D. Reijer, P. W. Otter, and J. P. Jacobs, "A heuristic scree plot criterion for the number of factors," *Statistical Papers*, pp. 1–10, 2024.

[4] S. Zhao, Y. Guo, Q. Sheng, and Y. Shyr, "Advanced heat map and clustering analysis using heatmap3," *BioMed Research International*, vol. 2024, no. 1, p. 986048, 2024.

[5] J. Ehrlich, J. Potter, S. Sanders, and R. Paul, "Spectators or influencers? the crowd effect upon winning in the nfl: A natural experiment," *International Journal of Sport Finance*, vol. 19, no. 1, pp. 3–24, 2024.