La Ciencia de los Datos Aplicada. Caso: "La Relevancia de las Pruebas en el Combinado Scout NFL en la Posición de "Linebackers"(Apoyadores) 1987 al 2022" [1].

Data Science Applied. Case: 'The relevance of the tests in the NFL Scout Combine for the Linebacker Position 1987 to 2022' [1].

Autor: Mg. Mauricio R. Arriaga Guajardo [2]
mauricio.arriaga@icloud.com
Doctorado en Ciencias en Inteligencia de Negocios. **TECANA AMERICAN UNIVERSITY**

RESUMEN

El objetivo de esta investigación es verificar la relevancia de las pruebas en el Combinado Scout NFL y predecir el éxito de los "Linebackers" (Apoyadores) desde 1987 hasta 2022 utilizando minería de datos y machine learning en R. La pregunta central es si estas pruebas pueden describir dicha probabilidad de éxito. Basado en la bibliografía de autores como Kelleher & Tierney y Lantz, se emplearon modelos de regresión y técnicas de machine learning (KNN, Naive Bayes, Árboles de decisión). Los resultados indican que los jugadores de élite son más ligeros, rápidos y fuertes. Se concluye que las pruebas son relevantes y se recomienda utilizar los modelos para futuras selecciones de jugadores.

Palabras Clave: "Combinado `NFL´ ", "Data Science," "Machine learning," "RStudio."

ABSTRACT

The objective of this research is to verify the relevance of the NFL Scouting Combine tests and predict the success of linebackers from 1987 to 2022 using data mining and machine learning in R. The central question is whether these tests can describe this probability of success. Based on the bibliography of authors such as Kelleher & Tierney and Lantz, regression models and machine learning techniques (KNN, Naive Bayes, Decision Trees) were employed. The results indicate that elite players are lighter, faster, and stronger. It is concluded that the tests are relevant, and it is recommended to use the models for future player selections.

Keywords: "NFL Scouting Combine", "Data Science", "Machine Learning", "RStudio"

[1] Trabajo de investigación 2022, parte del programa y plan de estudios, para el Doctorado en Ciencias en Inteligencia de Negocios de **TECANA AMERICAN UNIVERSITY (TAU)** of USA.

[2] Profesional en Inteligencia de Negocios y TIC, actualmente laborando en Microsoft. Concluyendo el Doctorado en Ciencias en Inteligencia de Negocios de TAU. Dos maestrías, una en IoT de la Universidad de Curtin, Australia y otra en Comercialización de Conocimientos Innovadores (UAEM). Dos carreras tanto en Administración como Ingeniería Computacional. Posee más de 65 certificaciones técnicas. Exjugador de Football Americano (AFAIMAC, FADEMAC y ONEFA) y entrenador por más de 20 años. mauricio.arriaga@icloud.com

INTRODUCCIÓN

La Ciencia de los Datos se ha establecido como una herramienta invaluable en el ámbito deportivo profesional, como lo ilustra el libro de Lewis (2004), donde se explica cómo los Oakland Athletics utilizaron datos para mejorar el reclutamiento de jugadores (Kelleher & Tierney, 2018, pág. 28). En esta investigación, se aplica la Ciencia de los Datos de manera específica al fútbol americano, centrándose en el área del machine learning y sus herramientas, particularmente mediante modelos estadísticos y algoritmos en R. Se busca responder a la pregunta de investigación: ¿Es posible verificar la relevancia de las pruebas en el Combinado Scout NFL para describir la probabilidad de éxito en la posición de Linebackers (Apoyadores) desde 1987 hasta 2022, utilizando minería de datos y machine learning con la herramienta de R?

La Ciencia de los Datos se define como el estudio de los datos con el objetivo de extraer información significativa para las organizaciones, combinando principios y prácticas de matemáticas, estadística, inteligencia artificial y/o machine learning, así como ingeniería informática para analizar grandes volúmenes de datos (Herbert, 2020). Este enfoque multidisciplinario permite a los científicos de datos formular y responder preguntas como las planteadas en esta investigación.

El documento está estructurado en una introducción breve que especifica el objetivo general y los objetivos específicos, seguido por cuatro capítulos. El Capítulo 1 abarca los antecedentes que motivaron la investigación, explicando los orígenes y la problemática identificada en las pruebas del Combine Scout NFL, justificando la realización del estudio, delimitando el problema y formulando la pregunta de investigación. Se describe la metodología descriptiva (mixta) aplicada de manera transversal, junto con las técnicas e instrumentos utilizados para la recopilación de datos. En el Capítulo 2 se presenta el marco teórico fundamental que sustenta la investigación, incluyendo el proceso de minería de datos, el aprendizaje automático (machine learning) y la aplicación de estadística avanzada en RStudio, con el objetivo de generar conocimientos reproducibles para investigaciones futuras. El Capítulo 3 detalla los resultados obtenidos y la principal contribución del estudio, explicando los algoritmos estadísticos y modelos de machine learning empleados en R para abordar la pregunta de investigación y presenta el desarrollo del estudio junto con los resultados. Finalmente, el Capítulo 4 presenta las conclusiones del estudio y ofrece recomendaciones para trabajos futuros.

Objetivo General

El objetivo principal de este informe es aplicar la Ciencia de Datos para verificar la relevancia de las pruebas en el Combinado Scout NFL y describir la probabilidad de éxito en la posición de Linebackers (Apoyadores) con información desde 1987 hasta 2022, utilizando minería de datos y machine learning con la herramienta R.

Objetivos específicos

- Describir los antecedentes de la investigación, así como el alcance, el problema identificado, la justificación del estudio y la metodología empleada.
- Disertar los fundamentos de la Ciencia de Datos que sustentan esta investigación
- Verificar la relevancia de las pruebas en el Combinado Scout NFL para evaluar la probabilidad de éxito en la posición de Linebackers (Apoyadores) desde 1987 hasta 2022 utilizando minería de datos y machine learning con la herramienta R.

CAPÍTULO 1

PLANTAMIENTO Y FORMULACIÓN DEL PROBLEMA

Proverbio Coreano, la mitad del viaje se logra con el primer paso…

Dada la relevancia del informe, es importante dejar acotado y claro, la descripción de los antecedentes de la investigación, el alcance, el problema, la justificación y la metodología

Antecedentes que inician la investigación

El Departamento de Operaciones de la National Football League (NFL) organiza anualmente desde 1987 un evento crucial para la selección de jugadores colegiales, conocido como el Combine Scout NFL. En este evento, los jugadores invitados son evaluados mediante pruebas físicas y mentales por scouts, gerentes generales y entrenadores de la NFL. Según Casan (2022) estas pruebas incluyen mediciones físicas, exámenes médicos, entrevistas, y diversas evaluaciones como la velocidad en las 40 yardas, salto de altura, salto de longitud, fuerza y agilidad. La proliferación mediática del evento a través de

canales como Fox Sports, NFL Channel y ESPN ha llevado a una creciente participación de analistas deportivos y fanáticos que cuestionan la utilidad y efectividad de estas pruebas en relación con el desempeño real de los jugadores profesionales.

Descripción del problema

Varios autores, incluyendo a Casan (2022) y el estudio de Ben-Ishay (2020) sugieren que las mediciones obtenidas en el Combine Scout NFL no siempre correlacionan de manera significativa con el éxito de los jugadores en sus carreras profesionales en la NFL. Esto plantea la incertidumbre sobre la efectividad de estas pruebas para seleccionar adecuadamente a los jugadores más prometedores. Es crucial determinar si existe una relación cuantificable y rastreable entre las pruebas realizadas en el NFL Scouting Combine y el rendimiento exitoso de los linebackers seleccionados en estos eventos, dado el significativo gasto de recursos y la atención que genera este proceso de selección.

Pregunta de investigación

¿Es posible verificar la relevancia de las pruebas en el Combinado Scout NFL para predecir la probabilidad de éxito en la posición de linebackers (apoyadores) con datos desde 1987 hasta 2022 utilizando minería de datos y machine learning con la herramienta R?

Justificación

Al analizar a los jugadores con los resultados obtenidos en el “Combinado NFL” y el desempeño actual (como jugador profesional) en la liga, se podría crear un modelo para identificar las características más relevantes en la posición de linebacker, aplicando la Ciencia de los Datos, podría ayudar a los tomadores de decisiones en los equipos y buscar eficiencia y eficacia al observar lo importante, al reclutar a los jugadores; así tanto la NFL, como los Gerentes Generales, y los dueños de los equipos podrían enfocar su atención al indicador correcto; acorde a Kelleher (Ciencia de Datos, 2018) la ciencia de los datos es comúnmente usada en los deportes profesionales, como se refleja en el libro (Lewis, 2004) el cual muestra como el equipo de baseball los Oakland Athletics utilizaron la ciencia de los datos, para mejorar el reclutamiento de jugadores, el científico identificó que en las estadísticas de porcentaje de llegar a la primer base y el poder de un bateador eran indicadores más informativos del éxito ofensivo, que las estadísticas regularmente utilizadas, esto lo usaron para seleccionar a los jugadores más adecuados y de mejor adhesión a los nuevos objetivos, resultando a su favor y llevando al equipo a resultados excelentes; de tal forma que acorde a Kelleher es justificable aplicar la Ciencia de los datos y sus herramientas a la

investigación del problema acotado en este estudio (p. 28,29); adicionalmente, la NFL incita a encontrar nuevas y fascinantes perspectivas desde el punto de vista de analítica avanzada.

Alcance

El alcance de esta investigación, abarca los siguientes elementos, definidos por Keller & Tierney (2018) para la ***"Ciencia de los Datos",*** se deben contemplar, el conjunto de principios, la definición del problema, los algoritmos y los procesos para extraer patrones no obvios y útiles de grandes conjuntos de datos y puntualizan que muchos de los elementos de la ciencia de los datos están relacionada al uso de ***aprendizaje automático*** **(**Machine learning) y la "***Minería de Datos"*** (p. 13); así que de primera instancia, el alcance cubre el conjunto de principios de la investigación, con fuentes válidas y oficiales, libros publicados y autores reconocidos en el ámbito de la Ciencia de los Datos y las estadísticas, se mantiene la perspectiva desde varios puntos de vistas de diferentes autores, para **fincar los fundamentos, conceptos generales** y se defina claramente el proceso requerido para trabajar c**on la ciencia de los datos, la cual cuenta con las herramientas suficientes** para poder abordar el problema desde varias perspectivas; en consecuencia; el segundo elemento (que incluye el alcance) es la definición del problema **a comprender, siendo la** pregunta de investigación ¿ Es posible verificar la relevancia de las pruebas en el Combinado Scout NFL para describir la probabilidad de ser exitoso en la posición del "Linebackers" (Apoyadores) 1987 al 2022 usando, minería de datos & machine learning, con la herramienta de R?; el conjunto de elementos adicionales que competen al alcance, es la inclusión del uso de (Machine Learning) que acorde detalla Lantz (2019) es posible profundizar y desarrollar grupos de algoritmos, que transformen datos en conocimiento accionable usando estadística avanzada con herramientas como RStudio y lenguaje "R" (p. 1) y para obtener y preparar los datos se deben gestionar a través de los procesos contenidos en la "Minería de Datos".

La investigación cubre puntos de análisis apoyados de la matemática, la probabilidad y la estadística; sin embargo, la investigación no pretende ser un curso de matemáticas, de tal forma que lo abarcado aquí, es la descripción, utilización y las aplicaciones de algunas de las herramientas para la ejecución de la estadística, así como generar modelos en lenguaje de ML específicamente "R" que pueda reproducir la investigación.

Metodología empleada

Hernández, Fernández y Baptista (2010) explican que, dada la naturaleza fundamental de este estudio, se enmarca en un tipo de investigación descriptiva. Esta investigación se llevará a cabo utilizando tanto enfoques cualitativos como cuantitativos. Münch (2009) especifica que la

investigación descriptiva se centra en características como la frecuencia, la ocurrencia y el desarrollo de fenómenos (p. 22).

Alcance de la Investigación metodológica

(Bernal, 2010) “La delimitación o el alcance en investigación se refiere a la dimensión o al cubrimiento que ésta tendrá en el espacio geográfico, período de tiempo y perfil sociodemográfico del objeto de estudio” (p 109); (Espacio geográfico) La investigación, se constituye del siguiente alcance, contiene datos obtenidos exclusivamente de la NFL de los EEUU, en referencia al periodo de tiempo en los años 1987 -2022

Diseño de la Investigación

Explica Hernández (2010)“con la finalidad de responder a las preguntas de investigación planteadas y para conseguir el objetivo del estudio, es necesario que el investigador diseñe tanto, un plan y la estrategia para conseguir la información necesaria en la investigación; en esas circunstancias en la presente investigación se llevará a cabo un diseño de investigación transversal, el cual consiste en recolectar los datos en un solo momento y tiene como finalidad describir las variables existentes y analizar su incidencia e interrelación en un momento dado” (p-45); en la tabla 1 se muestra las unidades de análisis, la población, la muestra, las técnicas, el diseño de investigación y el alcance metodológico.

Tabla 1

Síntesis de las Unidades de Análisis, la Población, la Muestra, las Técnicas, el Diseño de Investigación y el Alcance Metodológico

Enfoque	Descripción
Unidad de análisis	- Personas: Jugadores que oficialmente se presentaron al Combinado de la NFL (registrados 1987 al 2022); Así como, lista de los mejores 50 jugadores de la NFL era moderna (Wedell, 2010); lista de los 100 mejores jugadores de la temporada en curso 2022 (PFF, 2022). - Documentos: Datos estructurados y no estructurados disponibles en formato CSV u otros similares, oficialmente aprobados por la NFL, con los resultados de las pruebas del "Combine Scout NFL" desde 1987 hasta 2022.
Población	Jugadores que oficialmente se presentaron al Combinado de la NFL (registrados 1987 al 2022
Muestra	La muestra se limita al 100% de los jugadores en la posición de linebackers (apoyadores) participantes registrados en el "Combine Scout NFL" desde 1987 hasta 2022. Además, se incluye la lista de los 100 mejores apoyadores de la NFL de la temporada actual.
Técnica	Revisión documental digital a través de internet (secundarios).
Herramientas	Machine Learning con “R;” Excel y RStudio

Fuente: Elaboración del autor, con datos de la investigación

Dificultades y limitaciones confrontadas

Los datos recopilados tienen información importante pero incompleta; varias de las fuentes de datos con información relevante, nos son abiertas y tienen costos involucrados; la muestra de entrenamiento es pequeña y dispersa, lo que fracciona y puede sesgar los resultados si no son analizados holísticamente; las estadísticas disponibles de la NFL Big Data tienen restricciones para ser accesadas. La información proveniente de dispositivos del internet de las cosas (IoT) existen y son relevantes por que contienen información del movimiento real en el campo, sin embargo, no están disponibles por la NFL para ser analizadas.

CAPÍTULO 2

MARCO TEORICO CONCEPTUAL

Si crees en ti mismo y tienes el coraje, la determinación, la dedicación, el impulso competitivo y si estás dispuesto a sacrificar las pequeñas cosas de la vida y pagar el precio por las cosas que valen la pena, entonces "se puede lograr"… Entrenador Vince Lombardi.

En el presente capítulo se explora el marco teórico de la Ciencia de los Datos, según Keller & Tierney (2018) que abarca principios, definición de problemas, algoritmos y procesos para extraer patrones útiles de grandes conjuntos de datos. Este campo se relaciona estrechamente con el Machine Learning y la Minería de Datos (p. 13).

La Ciencia

Münch (2009) define la ciencia como un conjunto sistemático de conocimientos que permite al hombre explicar y transformar el mundo. Bernal (2010) añade que, históricamente, la ciencia se interpreta y utiliza de manera racional y sistemática en la sociedad (p. 13, 286).

Los Datos

Los Datos se refieren a una representación abstracta de la realidad que puede incluir sucesos, personas y objetos, dotados de atributos y variables que permiten su agrupación en conjuntos con similitudes (Kelleher & Tierney, 2018, pág. 35). Herbert (2019) señala que estos datos, al ser procesados mediante actividades de ciencia de datos como la transformación y la limpieza, generan información valiosa, representada en la pirámide de Han, Kanber y Pei (2011, pp. 9-10). Según Kelleher (2018), los datos se clasifican en numéricos, nominales y ordinales, con los numéricos admitiendo cálculos

matemáticos y los ordinales reflejando orden sin admitir operaciones aritméticas. Además, se identifican siete categorías de datos, incluyendo estructurados y no estructurados, cada uno con usos específicos en función de su estructura y tipo (Smith, 2022, págs. 8-11).

Figura 1. *La Pirámide Muestra la Evolución "Data Information Knowledge Wisdom", en la Parte Inferior, los Datos Crudos, en la Cima el Conocimiento*
Nota. Fuente: (Kelleher, 2018, págs. 39,40)

La Ciencia de los Datos

En el seminario "Statistical Thinking for Data Science and Analytics" de la Universidad de Columbia en 2018, diversos profesores definieron la Ciencia de los Datos. Según la Dra. Kathleen, implica la colaboración entre tecnólogos y solucionadores de problemas (Mckeown, 2018). El Dr. Blais enfatizó la construcción de herramientas con datos para resolver problemas, mientras que el Dr. Dubois Bowman describió un proceso que va desde la extracción inicial de datos hasta su integración y análisis a través de la minería de datos (Bowman, 2018). La directora de programas estratégicos de Ciencia de Datos predijo una revolución tecnológica con Machine Learning y AI, que transformará diversas áreas del conocimiento (Seminario, Universidad de Columbia, 14 de enero de 2018).

Herbert (2020) explica que el Deep Learning y el Machine Learning son aplicaciones de la Inteligencia Artificial, y destaca que los científicos y analistas de datos trabajan con información desde distintas perspectivas y campos de cobertura (p. 143, 9).

Los Científico de Datos

Herbert (2020) describe que los científicos de datos utilizan diversas perspectivas y herramientas especializadas para comprender la historia que cuentan los datos. Ellos crean modelos inteligentes que pueden ser alimentados constantemente con datos para mejorar y ajustar predicciones, como en el caso del automóvil autónomo de Google, que se vuelve más eficiente con el uso continuo (p. 11). Mientras los analistas de datos explican la historia de los datos, los científicos de datos emplean algoritmos avanzados como estadística, probabilidad y matemáticas para profundizar aún más (Figura 2).

Figura 2. *Campos de Acción del Analista de Datos y del Científico de Datos*
Nota. Fuente: Figura adaptada por el autor, basada en la obra original de Herbert (2020, pág. 7), quien a su vez la recuperó de Chapman y otros (2000).

La Minería de Datos

Dentro de la ciencia de los datos existe un proceso fundamental que es la “Minería de datos”, quien gestiona y colabora a través de seis fases que no se comportan de manera rígida y al terminar cada fase, puede iniciar la siguiente; las fases son, “la comprensión del negocio”, “la comprensión de los datos”, “la preparación de los datos”, “el modelado”, “la evaluación y despliegue” explica Chapman y otros (2000); las seis fases se muestran en la siguiente figura 3

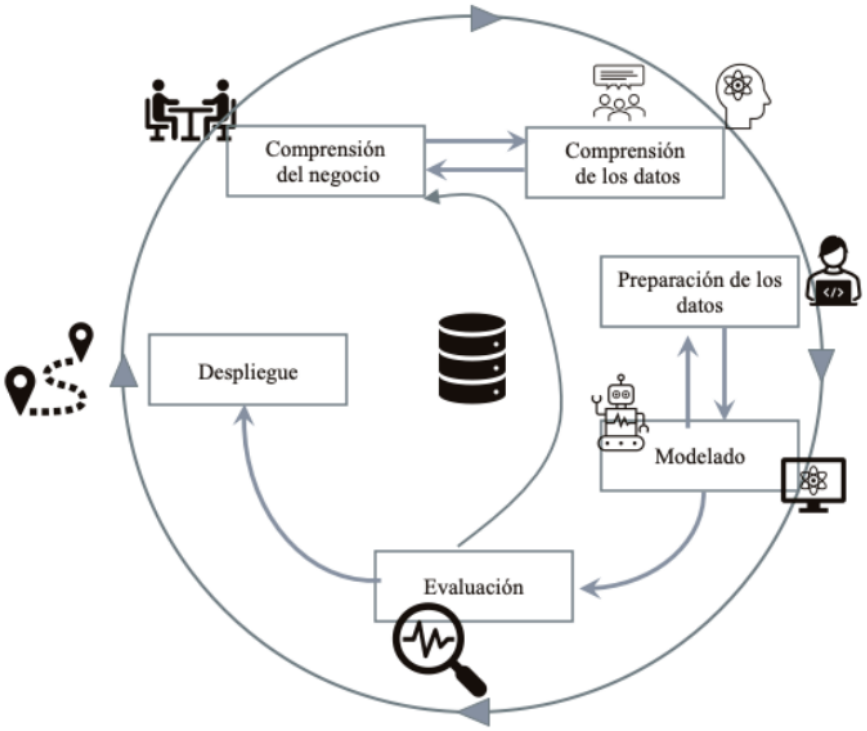

Figura 3. *El Ciclo de Vida CRISP-DM*
Nota. Fuente: Tomada de (Kelleher & Tierney, 2018, pág. 47) quien lo basaron en (Chapman, y otros (2000).

Fase I & II, Comprensión del negocio y de los Datos

Estas **dos fases** permiten al científico de datos investigar **qué hace el negocio** y **qué resultados desea obtener**, además de explorar los tipos de datos disponibles y decidir su utilidad (Kelleher, 2018, pág. 46). Coincide con Sherman (2015) quien enmarca esta fase en la **arquitectura de la información** en Business Intelligence i Inteligencia de Negocios (BI) y aborda preguntas clave **sobre análisis, decisiones, acceso a los datos y su ubicación**.

Fase III, Preparación de los datos

Incluye **la identificación de fuentes de datos**, **la recolección y limpieza de datos** utilizando herramientas específicas Chapman y otros (2000) Herbert (2019, págs. 42-52). Sherman (2015) destaca la importancia de construir **la franquicia de datos** para garantizar su utilidad a largo plazo.

Fase IV, El Modelado

Se refiere a la representación de **problemas del mundo real en términos matemáticos** para hacer predicciones y pronósticos, además,

definen el algoritmo como un **conjunto de instrucciones** o **cálculos diseñados para ser ejecutados por un ordenador**, e**scribir algoritmos** se conoce como **codificación o programación informática** Ahmed (2021). Herbert (2019), se centra en los modelos descriptivos y predictivos mediante **técnicas como la agrupación**, el **aprendizaje de reglas de asociación**, **el análisis de componentes principales** y la agrupación de afinidades (pág. 15).

Fase V, Evaluación

Implica probar **la calidad y validez del modelo** utilizando conjuntos de datos conocidos y de prueba comparte Chapman y otros (2000, págs. 14- 16).

Fase VI, Despliegue.

Consiste en implementar el modelo en producción para que los usuarios puedan beneficiarse de él (Herbert, 2019). Finalmente es importante considerar, que los procesos son iterativos, existe un círculo externo a todas las fases, lo cual implícitamente resalta que todo el proceso es iterativo, puede ser por mejoras, obsolescencias, actualizaciones o cualquier otra situación (Kelleher & Tierney, 2018).

Machine Learning

Explica (Lantz, 2019) que el hermano de la minería de datos es el aprendizaje automático (p.3); adiciona Nwanganga & Chapple (2020) que el aprendizaje automático (ML) es un subconjunto de técnicas de inteligencia artificial que, aplica estadísticas a problemas de datos en un esfuerzo por descubrir nuevos conocimientos a través de generalizar ejemplos y lo ejecuta con apoyo de las computadoras y los algoritmos (pp. 7,8). En la figura 4, se detalla como machine learning está contenida dentro de la IA.

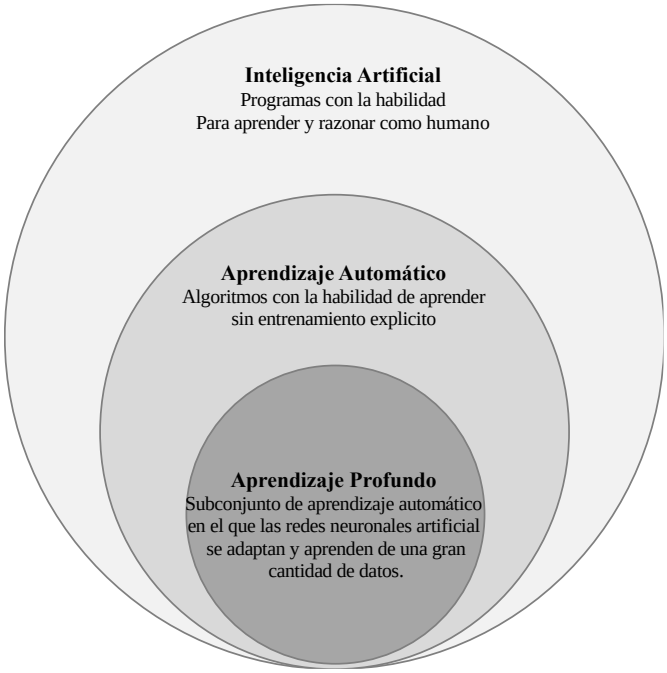

Figura 4. *Inteligencia Artificial y sus Aplicaciones, Machine Learning, Deep learning*
Nota. Fuente: Tomada de (Información, 2021) recuperado de https://www.ufsm.br/pet/sistemas-de-informacao/2021/05/11/introducao-a-machine-learning/

Machine learning llevado a la práctica

Lantz (2019) propone un proceso de **cinco pasos** para aplicar Machine Learning en el mundo real. Primero, **la recopilación de datos** implica reunir datos de aprendizaje en una fuente única como archivos o bases de datos. Segundo, **la exploración y preparación de datos** incluye limpiar y ajustar los datos para asegurar calidad. Tercero, se **selecciona un algoritmo** adecuado y **se entrena el modelo** con los datos preparados. Cuarto, **se evalúa la precisión del modelo** y se ajusta según sea necesario para **mejorar** su rendimiento. Este proceso asegura que el Machine Learning pueda aplicarse efectivamente a diversas tareas (p. 18).

Técnicas de aprendizaje Supervisado y No supervisado

Lantz (2019) distingue entre aprendizaje **supervisado y no supervisado** en técnicas de Machine Learning. El aprendizaje **supervisado** utiliza datos etiquetados para aprender patrones y realizar predicciones, como clasificar correos electrónicos como spam. En contraste, explican Nwanganga & Chapple el aprendizaje **no supervisado** descubre patrones sin datos etiquetados, útil por ejemplo en la detección de comportamientos fraudulentos

o en la agrupación de datos para análisis de segmentación (2020, págs. 12, 21).

Técnicas de Clasificación

En técnicas de Machine Learning, existen **tres principales tipos de aprendizaje**: **clasificación, regresión y aprendizaje de similitud** (clustering o agrupación). La clasificación predice la pertenencia a categorías no numéricas, como niveles educativos. La regresión predice resultados numéricos específicos, mientras que el clustering revela cómo las observaciones en un conjunto de datos se parecen entre sí (Nwanganga & Chapple, 2020, pág. 14);. Lants (2019) subraya la importancia de identificar qué técnica es adecuada para cada tipo de datos en proyectos de Machine Learning: clasificación, regresión, clustering y detección de patrones (p. 23, 24).

Tabla 2

Algoritmo de Aprendizaje No Supervisado

Modelo	Tarea de aprendizaje
Reglas de asociación	Detección de patrones
k-significa agrupamiento	Agrupamiento

Fuente: Tomada de (Lantz, 2019, pág. 23)

Tabla 3

Algoritmos de Aprendizaje Supervisados

Modelo	Tarea de aprendizaje
k-vecinos más cercanos	Clasificación
Naive Bayes	Clasificación
Árboles de decisión	Clasificación
Regresión lineal	Predicción numérica
Árboles de regresión	Predicción numérica

Fuente: Tomada de (Lantz, 2019, pág. 23)

Machine learning con “R”

Acorde a Nwanganga y Chapple (2020) explican que el lenguaje de programación R, iniciado en 1992, es ampliamente utilizado en estadística y análisis de datos. Como un lenguaje libre y de código abierto, su desarrollo es impulsado por una comunidad activa. La mayoría de las técnicas modernas de aprendizaje automático están prontamente disponibles para sus usuarios a través de paquetes en el Comprehensive R Archive Network (CRAN). Además, RStudio, un entorno de desarrollo integrado ha mejorado su accesibilidad y funcionalidad para realizar tareas en un ambiente gráfico (p. 26).

Estructura de Datos en “R”

En "R", Lantz (2019) explica que existen varias estructuras de datos fundamentales para el análisis estadístico y el aprendizaje automático. Estas **incluyen vectores, factores, listas, arreglos, matrices y marcos de datos** (data frames). **Los vectores** son conjuntos ordenados de elementos del mismo tipo, mientras que **los factores** representan categorías o variables ordinales dentro de vectores. **Las listas** permiten elementos de diferentes tipos, **los arreglos** son matrices bidimensionales, y **las matrices** contienen elementos numéricos para operaciones matemáticas. Los marcos de datos son estructuras clave similares a hojas de cálculo o bases de datos, combinando vectores y listas (Nwanganga & Chapple, 2020, pp. 46-50).

“R” estadísticas básicas

Profundizan Nwanganga y Chapple (2020) destacan que las estadísticas descriptivas facilitan la exploración y comprensión de datos, describiendo características como la moda y la frecuencia para datos categóricos, y la media y la mediana para datos continuos. También mencionan la correlación, que mide la relación entre dos variables, donde cambios en una afectan a la otra (p. 63). Lantz (2019) explica los cuartiles como herramientas analíticas básicas que dividen una serie de datos en porcentajes del 25%, con el segundo cuartil coincidiendo con la media (pp. 48-50)

K-vecino más cercanos (K-Nearest Neighbors)

Lantz (2019) explica que para desarrollar el **modelo K-NN** solo se necesitan los datos de entrenamiento y elegir el número de vecinos (k). Sus fortalezas incluyen su simplicidad, efectividad y velocidad en el entrenamiento, sin hacer suposiciones sobre la distribución de la data. Sin embargo, sus debilidades son la falta de transparencia del modelo, la necesidad de seleccionar un parámetro k y la lentitud en la fase de clasificación, especialmente con características nominales y datos incompletos (pp. 66,67). El siguiente es un ejemplo de clasificación en la figura 5.

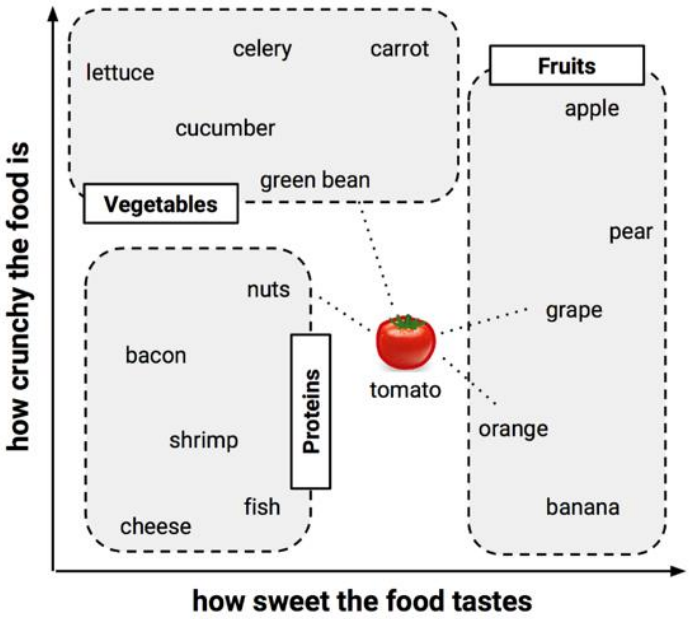

Figura 5. *K-Nearest Neighbors*
Nota. Fuente: Tomada de (Lantz, Gráficos Machine Learning with R)

Naive Bayes o Redes Bayesianas

Según Nwanganga & Chapple (2020), el enfoque de Bayes se basa en utilizar la probabilidad de eventos anteriores para predecir eventos futuros. Por ejemplo, al estimar la probabilidad de lluvia hoy, se considera la proporción de días similares en los que llovió en el pasado. Este método es aplicado en diversas áreas, como el filtrado de spam, donde se etiquetan correos electrónicos no vistos basándose en correos similares previamente etiquetados (p. 250). vease la figura 6 la formula representada de Bayes.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Figura 6. *La Probabilidad de que el Evento B Suceda, Dado que Hay un Evento A, Desarrollada por Naive Bayes*
***Nota. Fuente**: Tomada de (Lantz, 2019, pág. 94)*

Árboles de Decisión

Definen Kelleher & Tierney (2018) que los Árboles de Decisiones son “un tipo de modelos de predicciones que codifican las reglas [si-entonces] en una estructura de árbol, es decir, predice el resultado de la probabilidad que sucede al evento dado en un camino dado” (p. 159). Vease ejemplo figura 7.

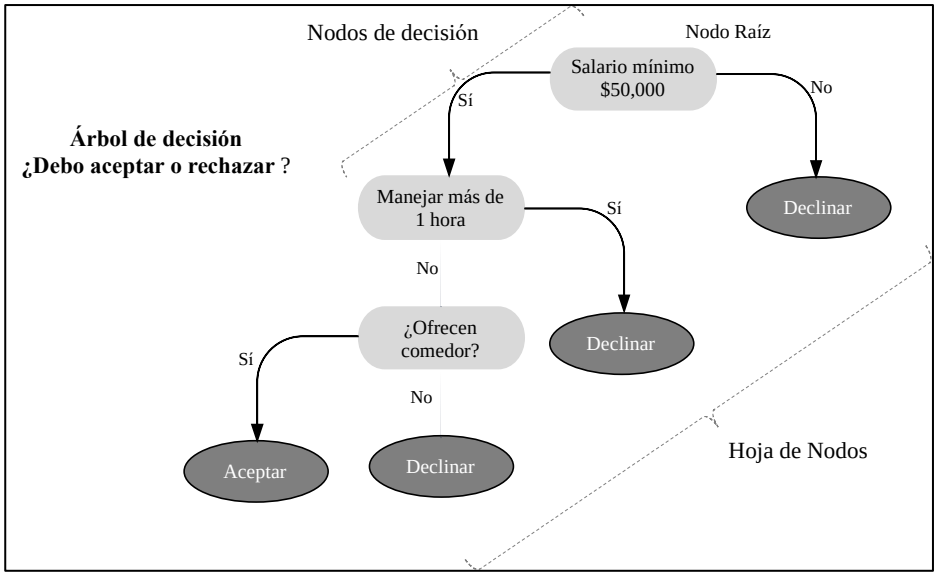

Figura 7. *Árbol de Decisión*
Nota. Fuente: (Lantz, Gráficos Machine Learning with R) https://static.packt-cdn.com/downloads/9781788295864_ColorImages.pdf

Árboles de Regresión. CART (Classification and Regression Trees)

Los Árboles de Regresión son una herramienta útil para predecir valores numéricos, como los precios de las viviendas basados en características como tamaño y ubicación. El algoritmo, como CART, aprende a dividir los datos en subconjuntos basados en características específicas para hacer predicciones precisas. Hay dos tipos principales de árboles para predicción numérica: los árboles de regresión, que predicen utilizando el valor promedio de los ejemplos en cada hoja, y los árboles modelo, que construyen modelos de regresión lineal en cada hoja. Aunque los árboles modelo pueden ser más difíciles de entender, tienden a generar modelos más precisos. La figura 8 muestra un ejemplo de un árbol de decisión, que es una representación visual de un árbol de regresión, ayudando a comprender cómo se toman las decisiones y qué características influyen en las predicciones (Lantz, Machine Learning with R, 2019, págs. 187, 188).

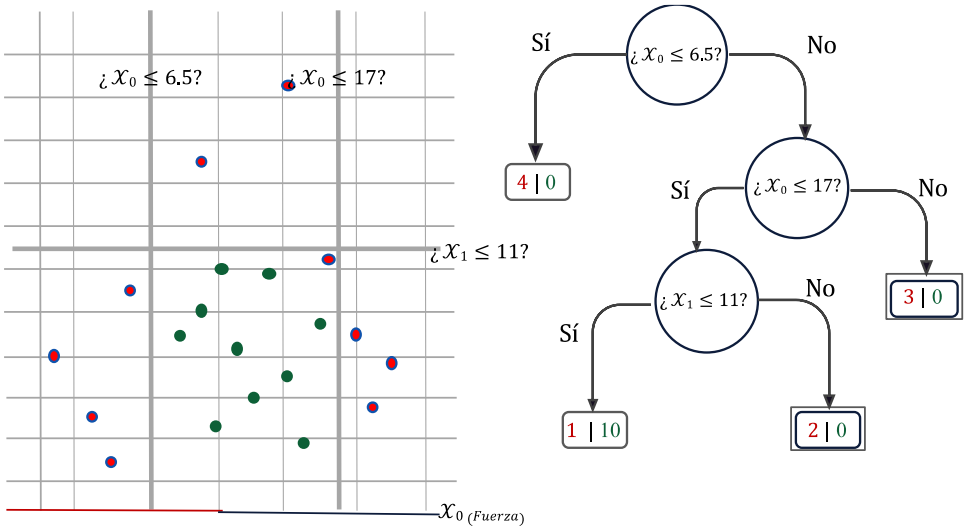

Figura 8. *Árboles de Clasificación y Regresión*
Nota. Fuente: (Decisión, https://www.codificandobits.com/img/posts/2021-04-11/arbol-decision-representacion-grafica.png2021)

Regresión lineal

Según Nwanganga & Chapple (2020), las técnicas de regresión en el aprendizaje automático buscan predecir una respuesta numérica al cuantificar la relación entre valores numéricos, utilizando la correlación para describir y cuantificar esta relación. La correlación se expresa mediante el coeficiente de correlación (pp. 103,106); por otra parte, Lantz (2019) explica la regresión lineal como la relación entre una variable dependiente y una variable predictora independiente, representada por una ecuación de pendiente-intersección. La pendiente describe el cambio en la variable dependiente dado un cambio en la variable predictora, mientras tanto la intersección indica dónde la línea cruza el eje y (p.172). Se presenta una esquematización de este concepto en la figura 9.

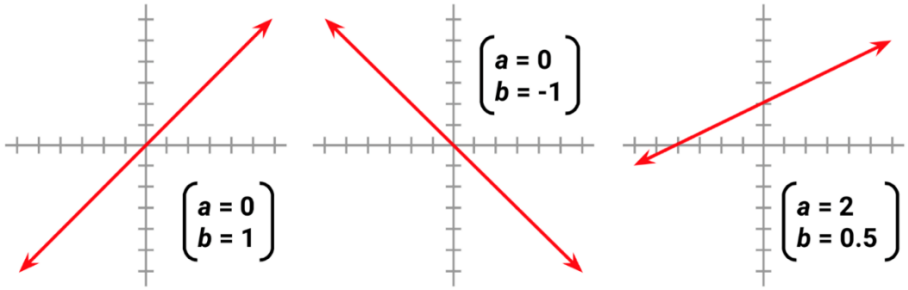

Figura 9. *Regresión lineal*
Nota. Fuente: (Lantz, Gráficos Machine Learning with R)

Ética y privacidad

El uso ético de los datos es crucial en este estudio. Spiegelhalter (2021) enfatiza la importancia de manejar los datos con responsabilidad, especialmente en contextos como las redes sociales y datos gubernamentales, donde la veracidad y las restricciones son fundamentales. Lantz (2019) destaca que el aprendizaje automático, aunque poderoso, debe utilizarse con principios éticos para evitar posibles abusos, como la vigilancia no autorizada. Las regulaciones como el GDPR en Europa y la Ley Federal de Protección de Datos en México imponen normas estrictas para proteger la privacidad. Microsoft (2017) promueve principios de legalidad, equidad y transparencia en el uso de datos. Kelleher (2018) menciona técnicas como la privacidad diferencial y el aprendizaje federado para preservar la privacidad mientras se realizan análisis de datos sensibles, evitando la discriminación y garantizando la seguridad de la información.

"NFL National Football League"

"La National Football League (NFL),en castellano es la Liga Nacional de Fútbol Americano, es la liga de mayor embestidura en el deporte practicado de manera profesional en los EEUU. Originalmente constituida en 1920 con once equipos y a la fecha se compone de 32 franquicias en diversas ciudades de los EEUU; se compone de dos conferencias, la Conferencia Nacional (NFC) y la Conferencia Americana (AFC) a su vez, cada conferencia se integra por cuatro divisiones (la del Norte, el Sur, el Este y el Oeste) y cada una de ellas, por cuatro equipos" (Carroll & Neft, 1999).

Pruebas en Combinados de la NFL

Explica el "Departamento de Operaciones de la "National Football League" (NFL), que lleva a cabo la organización de un evento formal (desde 1987) para el proceso de la selección de jugadores colegiales, con la intención de integrarles a las filas de los equipos de la máxima categoría de éste deporte; así que cada año se lleva a cabo el conocido evento "Combinado de la NFL" o por su nombre en inglés "NFL Scouting Combine"; la cita dura una semana, en el mes de febrero en el estadio "Lucas Oil Stadium" en la ciudad de Indianápolis (Operations, 2022), véase también

El Combinado Scout NFL incluye varias pruebas físicas clave para evaluar a los atletas:

Carrera de 40 yardas, mide velocidad, aceleración y explosividad desde un inicio estático. Los atletas son cronometrados en intervalos de 10, 20 y 40 yardas, tomada la idea de Coauch Paul Brown tiempo que requieren para las patadas de despeje (Casan, 2022).

Levantamiento de pesas (Bench press), evalúa la fuerza del tren superior mediante repeticiones máximas con 225 libras, indicando la preparación física del jugador en la universidad. Idea tomada por el esfuerzo de bloqueo cuerpo a cuerpo (Casan, 2022).

Salto vertical, Demuestra la explosión de la parte inferior del cuerpo, midiendo la potencia y la fuerza durante un salto desde posición estática (Ben-Ishay, 2020).

Salto de longitud, similar al salto vertical, prueba la potencia y explosividad en dirección horizontal, evaluando fuerza, equilibrio y coordinación en el salto y aterrizaje.

Ejercicio de explosividad a 3 conos, evalúa la agilidad y cambio de dirección rápida con giros de 90° alrededor de conos dispuestos en forma de L.

Carrera de ida y vuelta, prueba la rapidez lateral, aceleración, cambio de dirección, detención, equilibrio, agilidad y explosividad en distancias cortas (Casan, 2022).

Estas pruebas ayudan a los equipos de la NFL a evaluar diferentes habilidades físicas y técnicas de los prospectos.

La posición de Linebacker (apoyador)

El linebacker es clave en la defensa, encargado de detener el avance de la ofensiva, cubrir pases y liderar el equipo defensivo. Hay tres posiciones comunes: el "Mike" como apoyador central, el "Sam" en el lado fuerte y el "Will" en el lado débil. Estos nombres ayudan a identificar sus roles en la defensiva, recordando "Strong" para Sam y "Weak" para Will (Carroll & Neft, 1999)

.

CAPÍTULO 3

DESARROLLO

Este informe utiliza una congruencia metodológica para evaluar la relevancia de las pruebas del Combinado Scout NFL en la predicción del éxito en la posición de Linebackers. Se aplica minería de datos y machine learning con R, basándose en el proceso de cinco pasos de Lantz (2019) para aplicaciones del mundo real. Además, se sigue el modelo de seis fases de Chapman et al. (2000) en minería de datos. Se inicia con la comprensión del negocio y la recopilación de datos, seguido por la exploración, limpieza y preparación de datos. Posteriormente, se construyen modelos estadísticos como KNN, Regresión Lineal, Árbol de Decisión y Naive Bayes, los cuales se entrenan, ejecutan y evalúan utilizando técnicas de machine learning.

La comprensión del negocio

Acorde a la NFL explica que la relevancia de la posición del apoyador, puede marcar la diferencia entre perder y ganar el partido, ya que el trabajo que realiza, como parte integral de la defensiva (independientemente de ser un **líder**) es el ser **un estratega** para evitar que avance el equipo ofensivo; se colocan detrás de los linieros defensivos (tackles), es decir, detrás de aquellos jugadores que se colocan en primera línea y casi siempre agachados (manos al piso); se les llaman **apoyadores** porque apoyan a los linieros defensivos para cerrar las brechas, detener a los corredores, lo cual les implica una reacción rápida y **fortaleza física**, así también deben apoyar a los jugadores del perímetro en las jugadas de pase, lo que les demanda **gran velocidad y cambios de direcciones**; aunque es común que inicien con un paso de ajuste y comience la **lectura de sus jugadores llaves**, ya sean los guardias, QB, HB, TE, entre multiples opciones, con ello comienza su cobertura de apoyo, muchas veces se les indica que atraviesen la línea de golpeo (**penetración**) para capturar al mariscal de campo o al corredor antes de que puedan ejecutar la jugada; dependiendo de la formación defensiva puede haber prácticamente cualquier número de jugadores colocados como apoyadores, aunque lo común es usar de dos a cuatro en el terreno de juego (Grades, 2022); acorde a Cassan (2022) algunas de las características más importantes en este tipo de jugador, **es la velocidad**, ya que requieren perseguir a los corredores de bola, que son considerados ligeros y muy ágiles, así también deben cubrir y evitar que reciban **pases tanto las alas cerradas**, como los corredores que salen de escape (es decir hacia la banda del terreno de juego); se suma a la complicación que requiere **gran fortaleza física** para contener en **la línea de golpeo y no ceder espacio**, ni perder de vista el balón en todo momento; gran parte de sus movimientos son **explosivos, laterales y cambios de direcciones con ángulos radicalmente opuestos**, una característica muy importante es el **reconocimiento de patrones ofensivos**, que le permita leer, ajustar y anticipar lo que hará la ofensiva (p. 12). Con base en estas

responsabilidades, hay una serie de criterios que se utilizan para elegir a la elite de apoyadores que han participado en el juego desde los inicios de la historia en la NFL; estos incluyen sus logros personales y de equipo, las estadísticas de su carrera y el grado de influencia en el juego, todos evaluados a través de la lente de la era en la que jugaron; estas son algunas de las selecciones con los mejores apoyadores de la NFL de todos los tiempos, Lawrence Taylor, PatrickWillis, Ray Lewis, Jack Lambert, Jack Ham, Sam Huff, Mike Sigletary, Randy White, Dick Butkus, Junior Seau, Ray Nitschke, entre muchos otros (vease Apendice A, se incluye la lista de los mejores jugadores identificados); así que para los scouts estos son algunos prototipos de lo que deben buscar al identificar posibles prospectos; en alas de obtener más información de los jugadores aspirantes del colegial, en el combinado de la NFL incluyen, acorde a Casan (2022) la carrera de 40 yardas, la fuerza en levantamiento de pesa(en banquillo), el salto vertical, el salto de longitud, el ejercicio de correr a 3 conos y la prueba de correr ida y regreso.

La Minería de Datos

Este estudio es conformado, por un lado, con la información proveniente de los resultados en las pruebas del combinado en el periodo de estudio definido y en segundo término se han enriquecido los datos al identificar en ellos a los jugadores que han sido exitosos; de tal forma ahora, con estos datos cruzados, es posible conocer el desempeño de la elite, en su paso en las pruebas del combinado.

Al revisar la estructura y la integridad, se identifican datos faltantes, se revela información interesante, algunos de los jugadores más reconocidos de la historia, no cuentan con resultados del Combinado de la NFL; ver la figura siguiente que muestra a Ray Lewis, James Farrior y Keith Booking , sin resultados en las 40 yardas, en la prueba de los 3 conos, entre otros datos inexistentes., vease figura 10.

Year	Name	College	POS	Height (in)	Weight (lbs)	Wonderlic	40 Yard	Bench Press	Vert Leap (in)	Broad Jump (in)	Shuttle	3Cone	RANK
1996	Ray Lewis	Miami (FL)	ILB	72.4	235								1
1997	James Farrior	Virginia	OLB	73.8	234			22	35.5	120	4.4	7.62	1
1998	Keith Brooking	Georgia Tech	OLB	74.4	244			24					1
2016	Myles Jack	UCLA	OLB	73	245			19	40	124			52
2016	Jaylon Smith	Notre Dame	OLB	74	223								44
2018	Rashaan Evans	Alabama	OLB	73.88	232				30	116	4.36	6.95	41
2021	Jeremiah Owusu-Koramoah	Notre Dame	OLB	73.5	221				36.5	124	4.15		24

Figura 10. *Ray Lewis y tres super estrellas sin resultados del Combinado*
Nota. Fuente: Elaboración del autor con la herramienta de RStudio

Resumen Estadístico y Análisis

Al continuar con el análisis estadístico de los datos del combinado de la NFL, se realizó una exploración detallada de la estructura y distribución de las

variables. Resumen del análisis estadístico de los datos del combinado de la NFL (1987-2022):

Distribución de Datos Generales:

Altura: Rango de 68.75 pulgadas (1.746m) a 78.38 pulgadas(1.991m), media de 73.67 pulgadas (1.871m).

Peso: Rango de 203 libras (92.1kg) a 277 libras (125.6kg), media de 238 libras.(108kg).

Wonderlic: Rango de 13 a 32, media de 21.17.

Pruebas Físicas: Rangos y medias de velocidad en 40 yardas, press de banca, saltos vertical y horizontal, y pruebas de shuttle y 3 conos.

Comparación de Jugadores Top 100 vs No Rankeados:

Diferencias mínimas en peso y tiempos de velocidad en 40 yardas entre ambos grupos.

Hallazgos Clave:

Datos Faltantes: Ausencia de registros en pruebas clave para jugadores notables.

Rendimiento de la Élite: Consistentemente superior, destacando la importancia de pruebas físicas en la predicción del éxito.

Ahora bien, una variable interesante **es la velocidad en las 40 yardas**, un ejemplo que ayuda a dar contexto a estas cifras es el siguiente comparativo entre dos personas ajenas a este deporte, el primero (dado que existe la precedencia y resultados en los registros en la NFL) del **atleta olímpico** más rápido de la historia **Usain Bolt**, en esta prueba obtuvo un **impresionante 4.22 segundos**, igualando al receptor más rápido que se presentó en el 2017 John Ross; así su opuesto el famoso comentarista de la NFL **Rich Eisen**, quien año tras año participa en el evento (con fin de recaudar fondos de apoyo social) pero sus registros son guardados por la NFL dentro de sus **mejores tiempos logra un mínimo de 5.98 seg**; bien con este contexto se compara contra todos **los linebackers que promedian (media) +**, lo cual elimina de ser un posible seleccionado a linebacker a Rich Eisen; ahora bien, comparando al apoyador de la elite más veloz con 4.42 (mínimos) segundos, se concluye que no podrá alcanzar al atleta olímpico.

Figura 11. *Rich Eisen y sus tiempos en las 40 yardas*
Nota. Fuente: NFL (NFL, 2022); https://www.nfl.com/videos/run-rich-run-year-by-year-results-of-rich-eisen-s-40-yard-dash

El siguiente gráfico (figura 12) presenta dos histogramas comparativos de la velocidad en 40 yardas entre los jugadores catalogados como "Los Mejores" y "NO rankeados".

Los Mejores
Velocidad predominante: 4.5 segundos.
Distribución: Mayoría entre 4.4 y 4.8 segundos.
Varianza: Baja, indicando tiempos más uniformes.

NO rankeados
Velocidad predominante: 4.8 segundos.
Distribución: Mayoría entre 4.6 y 5.0 segundos.
Varianza: Alta, indicando mayor dispersión.

Comparación
Los Mejores: Más rápidos y uniformes en tiempos.
NO rankeados: Más lentos y mayor dispersión.

La velocidad es un factor clave que diferencía a los mejores jugadores de los no rankeados (o no destacados).

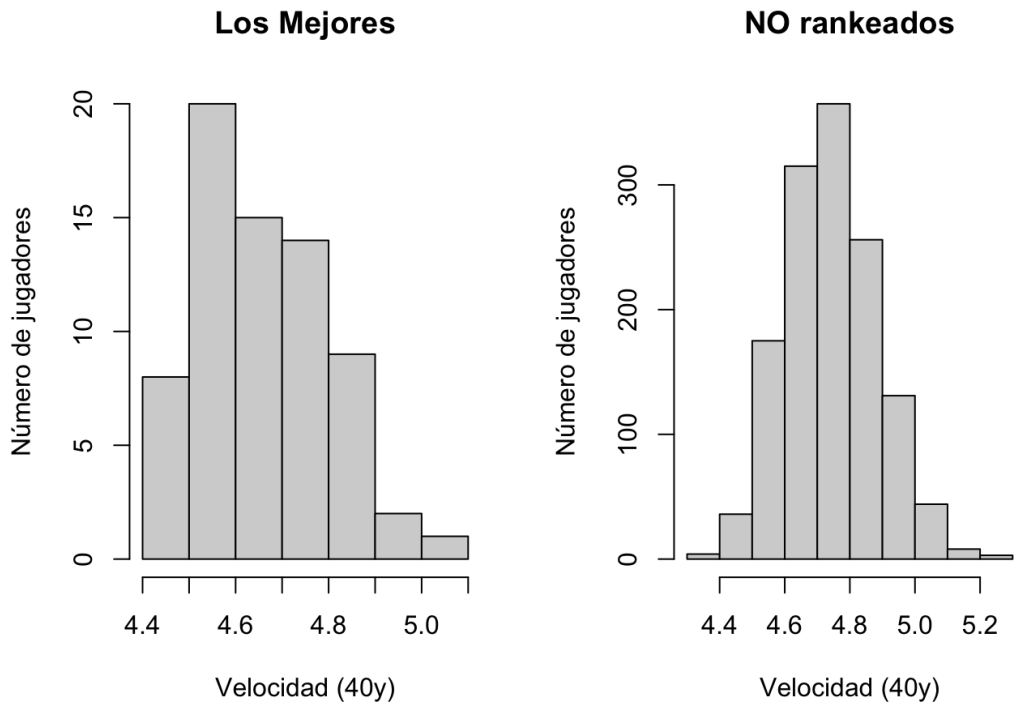

Figura 12. *Gráfico de histograma, la distribución de los jugadores por velocidad*
Nota. Fuente: Elaboración del autor con Studio R y datos de la investigación

La NFL proporciona datos que incluyen una variable conocida como **Wonderlic**, diseñada para identificar a las mejores selecciones en posiciones como la de apoyador. Sin embargo, según este estudio, la métrica de Wonderlic solo ha tenido éxito en predecir correctamente **a 3 de los 24 jugadores** ahora considerados élite, lo que **equivale a una fiabilidad de predicción del 12.5%.**

Machine Learning

Regresión Lineal

Para comprender las relaciones entre una variable dependiente y una o varias variables independientes, se ha aplicado el análisis de regresión lineal. En la figura 13 se presenta una de las varias regresiones realizadas. Se observa que **conforme aumenta el peso de los jugadores**, aumentan los tiempos en las 40 yardas, lo que indica que **son más lentos**. Sin embargo, es

evidente que los jugadores élite son más veloces y ligeros, como muestra la línea roja tenue.

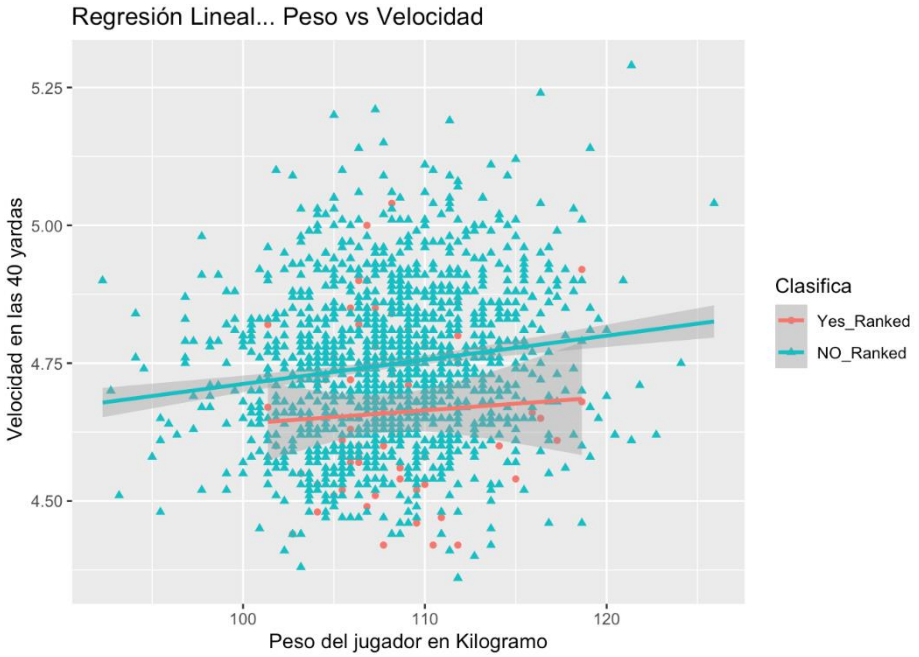

Figura 13

Regresión Lineal Comparativo, Peso del Jugador vs Velocidad

***Nota, Fuente**:Elaboración del autor, desarrollada en los trabajor realizados en RStudio.*

Usando análisis de regresión y considerando la variable del año en que los jugadores participan en el combinado, se observa que tanto la fuerza como la velocidad han mejorado significativamente. En la **figura 14 se muestra únicamente la mejora en la velocidad cada año**, **destacando que los jugadores de élite siguen obteniendo mejores resultados.**

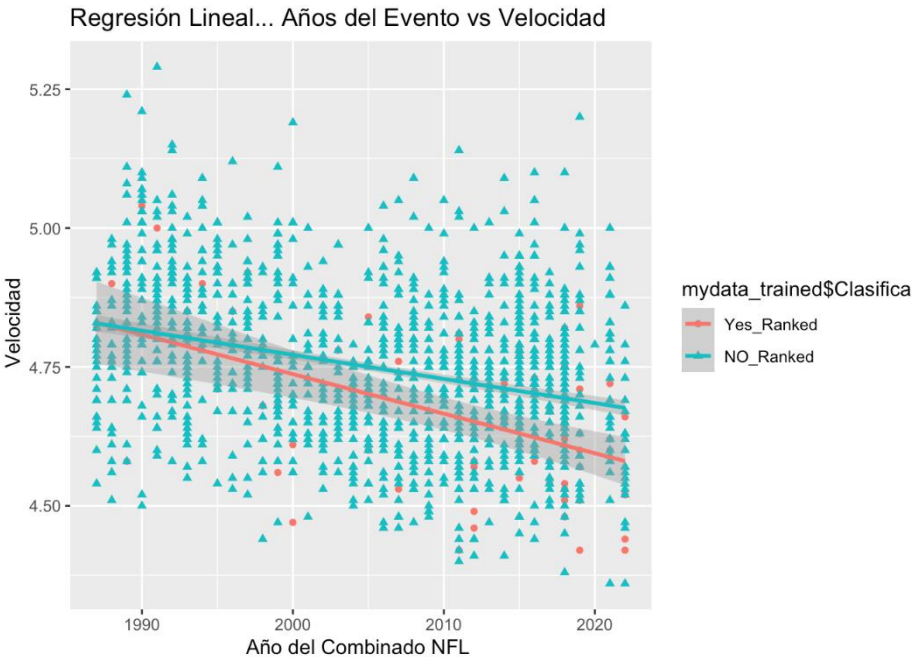

Figura 14
Regresión Lineal Comparativo, Año del Combinado vs Velocidad
Nota, Fuente: Elaboración del autor, en esta investigación con uso de RStudio

En la siguiente figura 15 se compara la **velocidad vs fuerza,** la relación explorada es entre la velocidad del jugador en recorrer 40 yardas y su fuerza, medida como el número de repeticiones que el jugador puede hacer en un bench press de 255 libras (115 kg). El análisis sugiere que **los jugadores que son capaces de hacer más repeticiones (indicando mayor fuerza) tienden a ser más rápidos en las pruebas de 40 yardas**. El color y la forma nuevamente clasifican a los jugadores, con 15 de 18 jugadores elite siendo rápidos y fuertes.

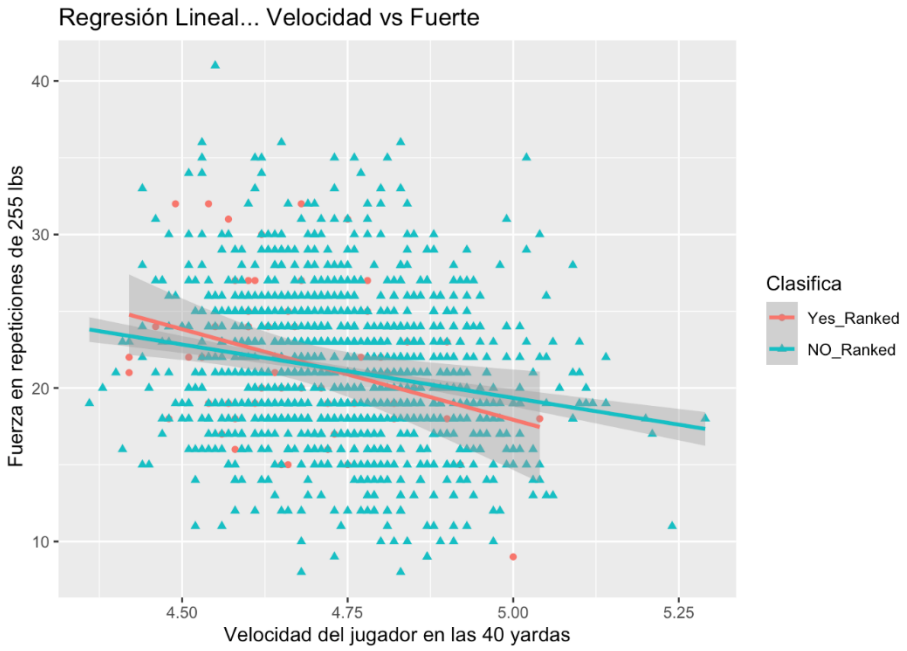

Figura 15

Regresión Lineal Comparativo, Fuerva vs Velocidad

Nota, Fuente: Elaboración del autor, en esta investigación con uso de RStudio

Análisis de Correlación en Jugadores Elite Mediante Gráficos en Pares

Para explorar las correlaciones entre variables y características de jugadores elite, se emplearon gráficos en pares, que incluyen diagramas de dispersión con elipses de tendencia y histogramas diagonales para evaluar distribuciones. Estos gráficos revelan las interdependencias y distribuciones de las variables analizadas**. Se destacan correlaciones significativas como la observada entre el salto vertical y el salto longitudinal** (coeficiente de 0.72) y **entre las pruebas de agilidad de tres conos** y de ida y vuelta (**coeficiente de 0.45**), subrayando las competencias donde los jugadores elite sobresalen. Estos hallazgos permiten un entendimiento más profundo de las habilidades que diferencian a los atletas de alto rendimiento. Vease figura 16.

Comandos usados en leguanje R (en RStudio):

```
pairs.panels(data1_trained[c("PesoKg", "X40.Yard", "Bench.Press", "Vert.Leap..in.", "Broad.Jump..in.", "Shuttle", "X3Cone")])
```

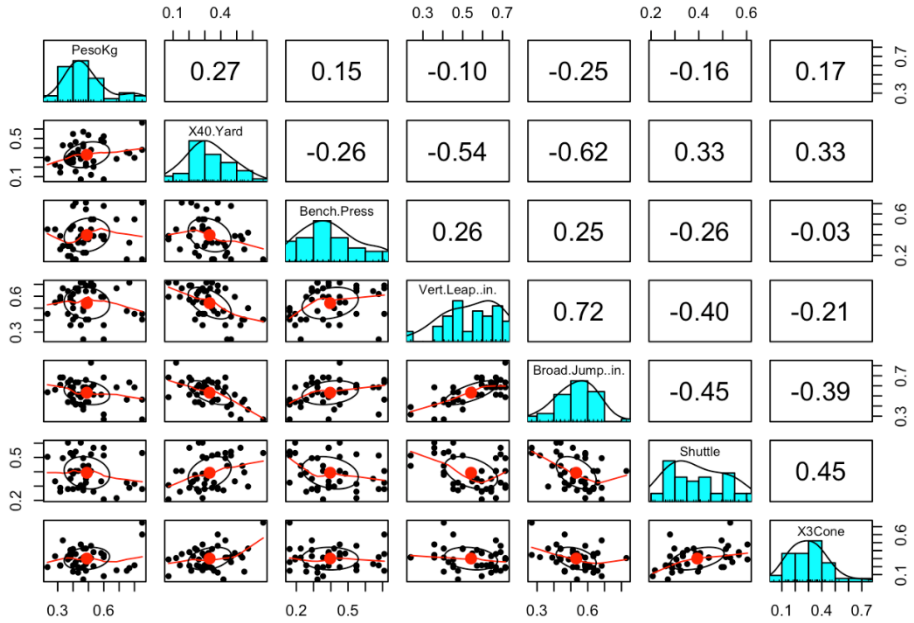

Figura 16. *Mapa de Pares de Regresión Lineal, Histogramas y Factor de Correlación*
Nota. Fuente: Elaboración del autor con herramienta RStudio y datos del estudio

K Nearest Neighbors

En un esfuerzo por mejorar la selección de los jugadores para futuros eventos del combinado, se ha adoptado un enfoque de aprendizaje supervisado utilizando la técnica de clasificación K-Nearest Neighbors (K-NN). Esta metodología, considerada parte del aprendizaje "perezoso", se centra en identificar y agrupar candidatos con perfiles similares a los de jugadores elite previamente clasificados. Para implementar este modelo, se prepararon y normalizaron los datos para asegurar que todos los registros estuvieran completos y sin valores faltantes.

Se configuró el modelo utilizando un conjunto de datos que incluye registros de 42 jugadores clasificados como elite. Este conjunto se dividió en dos segmentos: uno de entrenamiento con 30 jugadores y otro de prueba con 12 jugadores clasificados y varios no clasificados. El modelo se entrenó exclusivamente con el segmento de entrenamiento para identificar patrones y características de jugadores elite.

Breve extracto de las actividades realizadas:

```
##
## Yes_Ranked  NO_Ranked
##         42        765
```

Posteriormente se extrae el campo clasificador y se almacena temporalmente

```
mydata_train_n_labels <- mydata_trained_n[1:30,13]
mydata_test_n_labels <- mydata_trained_n[31:60,13]
```

De los 42 jugadores de elite se construyen dos archivos, uno para entrenar al modelo, el otro para comprobar el modelo, de tal forma el archivo entrenador (train) se le asignan 30 elementos con Yes_Ranked, es decir jugadores elite o los mejores

```
train <- mydata_trained_n_short[1:30,]
```

La tabla de prueba incluye a 12 jugadores exitosos y el resto no está rankeado

```
test <- mydata_trained_n_short[31:60,]
```

Ejecución del modelo k-nn, lazy y supervisada

El modelo es entrenado con 30 jugadores elite. Al terminar el entrenamiento se despliega el contenido y se aprecia que es correcto predice que hay 30 jugadores elite en la lista

```
myPrediction <- knn(train, test, mydata_train_n_labels, k = 1, prob=TRUE)
attributes(.Last.value)

table(myPrediction)
```

```
## myPrediction
## Yes_Ranked  NO_Ranked
##         30         0
```

Una vez completado el entrenamiento, el modelo fue evaluado utilizando el segmento de prueba. Los resultados mostraron que el modelo fue capaz de identificar correctamente a todos los jugadores clasificados como elite en el conjunto de prueba, demostrando una precisión del 100% en la detección de estos perfiles. Sin embargo, en cuanto a la confianza en sus predicciones, el modelo indicó un 40% de certeza en sus predicciones positivas y un 60% en las negativas, reflejando cierta incertidumbre en la clasificación final.

```
CrossTable(x= mydata_test_n_labels, y =myPrediction, prop.chisq = FALSE)
```

```
##   Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
## Total Observations in Table:  30
##                     | myPrediction
## mydata_test_n_labels | Yes_Ranked |  Row Total |
## ---------------------|------------|------------|
##          Yes_Ranked |         12 |         12 |
##                     |      0.400 |            |
## ---------------------|------------|------------|
##           NO_Ranked |         18 |         18 |
##                     |      0.600 |            |
##        Column Total |         30 |         30 |
```

Árbol de Decisión

Se utilizó un modelo de árbol de decisión para predecir la posición de apoyadores internos (ILB) o externos (OLB) en la NFL, basado en datos del Combinado Scout desde 1982 hasta 2022. El modelo, entrenado con datos de 30 jugadores exitosos, mostró que el 62% de los jugadores con los mejores tiempos en la prueba de 40 yardas se clasificaron como OLB, mientras que el 60% de los ILB demostraron mayor fuerza. La prueba de tres conos, que evalúa agilidad y cambio de dirección, fue una característica clave en ambas posiciones. Este modelo proporciona una base estadística para la selección de posiciones defensivas basadas en capacidades físicas.

Breve extracto de las actividades realizadas:

Creación del modelo

```
arbol <- rpart(
  formula = POS ~ X40.Yard + Vert.Leap..in.+ Bench.Press + X3Cone + PesoKg +
    Broad.Jump..in. ,
  data = mydata_trained_n,
  method = "class")
```

Gráficando el árbol con la probabilidad de ser OLB o ILB

```
fancyRpartPlot(arbol)
```

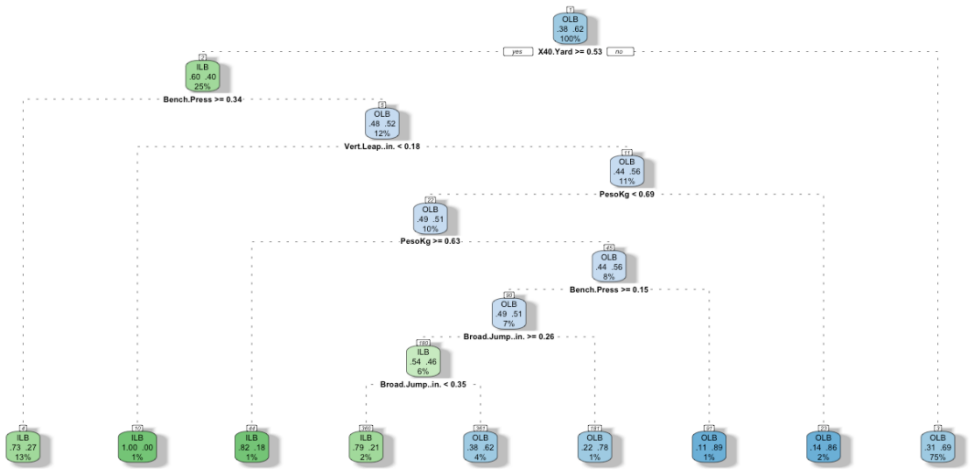

Figura 17
Árbol de Decisión Para Estimar La Probabilidad de Ser ILB u OLB
Nota, Fuente de elaboración del autor, con la herramienta de RStudio.

Naive Bayes

Dado un evento independiente que probabilidad hay de que suceda un segundo evento, con este principio se ha entrenado el modelo de Bayes, se le ha entrenado con un archivo que cuenta con 30 jugadores de elite, así tambien se ha segmentado un archivo de prueba híbrido, con 12 jugadores elite y 18 que no lo son.

Se entrena y se ejecuta el modelo de Naive Bayes, el cual identifica los nombres de los jugadores exitosos al paso de la ejecución como se aprecia a continuación.

```
m <- naiveBayes(train2, mydata_train_n_labels, laplace = 0)
```

Evaluación del modelo de Naive Bayes

El modelo identifica bien a los 12 jugadores identificados como los apoyadores de la “elite” ; ahora se podría probar contra otros jugadores para probar su resultado probabilistico de seleccionar jugadores que serán exitosos como profesionales.

```
m_test_Prediction <- predict(m, test2)
CrossTable(m_test_Prediction, mydata_test_n_labels, prop.chisq = FALSE,
           prop.c = FALSE, prop.r = FALSE, dnn = c("predicted", "actual"))
```

```
##   Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  30
##
##
##                  | mydata_test_n_labels
## m_test_Prediction | Yes_Ranked |  NO_Ranked |  Row Total |
## ----------------------  |-------------------|-------------------|------------|
##        Yes_Ranked -|         12 -------|         18 --------|         30 |
##              ----------   - |    0.400 -------|     0.600------- |            |
## -------------------------|-------------------|-------------------|------------|
##      Column Total --|         12 --------|         18 --------|         30 |
## -------------------------|-------------------|-------------------|------------|
##
```

El modelo identifica bien a los 12 jugadores pertenecientes a la “elite”. El modelo está listo para probarse en próximos resultados de las pruebas de los “Combinados Scouts NFL”, así entregará la probabilidad de ser jugadores exitosos como profesionales.

CAPÍTULO 4

CONCLUSIONES

En este capítulo del informe de investigación, se presentan las conclusiones generadas por el análisis de los resultados elaborados. Con el objeto de organizar el cuerpo de conclusiones, se agrupan atendiendo al objetivo general y los objetivos específicos a saber.

En cuanto al objetivo general en este estudio, se aplicó la Ciencia de Datos, para verificar la relevancia de las pruebas en el Combinado Scout NFL y se describió la probabilidad de ser exitoso en la posición del “Linebackers” (Apoyadores) con información del periodo de tiempo 1987 al 2022 y utilizando, la minería de datos y el machine learning, con la herramienta de R, este objetivo general se alcanzó con una propuesta de dos modelos estadísticos, KNN y Naive Bayes, realizados en machine learning y la herramienta de R en RStudio, que fueron entrenados para estimar la relación dado el evento previo del “Combinado Scout de la NFL” se de una probabilidad de un segundo evento el cual es el jugador seleccionado sea exitoso, describiendo así la probabilidad de ser exitoso como Apoyador, previamente se calculó con regresión lineal que los jugadores elite, en su paso por las pruebas del combinado, demostraron características como ser más rápidos, más fuertes y más ágiles, permitiendo entrenar a los modelos estadísticos con estas características, de tal forma que el objetivo general se cumplió,

En cuanto a los objetivos específicos

En el capítulo uno, se describieron los antecedentes de la investigación explicandose el evento del Combinado Scout NFL, las pruebas que se realizan y se describió el alcance de este estudio, el problema en detalle, la justificación de llevar a cabo el estudio y se describió la metodología de esta investigación como descriptiva, mixta y transversal, por lo cual se cumplió el objetivo completamente.

A lo largo del apartado referente al marco teórico conceptual, se disertó la Ciencia de los Datos y sirvió de soporte a esta investigación, por lo cual el objetivo se cumplió

Se verificó la relevancia de las pruebas en el Combinado Scout NFL y se describió la probabilidad de ser exitoso en la posición del “Linebackers” (Apoyadores) con información recolectada del periodo 1987 al 2022 usando, minería de datos & machine learning, con la herramienta de R. Se aplicaron al unisono dos conceptos hermanos, la minería de datos y el machine learning, con ellos se sintetizaron los pasos que fueron guia de las acciones en el estudio, con machine learning en la herramienta R, se creo un algoritmo, y se ejecutaron los pasos de la minería, permitiendo identificar las estructuras de los datos, los tipos de variables y permitiendo trabajar exitosamente la limpieza, la transformación, la agregación y la eliminación de los datos. Se identificó a través de la regresión lineal, las variables que distinguen a los jugadores considerados la elite (los mejores) en el campo de juego, permitiendo con ello entrenar a dos modelos estadísticos el KNN y el Naive Bayes, fueron diseñados para estimar la relación dado el evento previo del “Combinado Scout de la NFL” se de una probabilidad de un segundo evento el cual es el jugador seleccionado sea exitoso, describiendo así la probabilidad de ser exitoso como Apoyador.

Adicionalmente se desarrolló un tercer modelo estadístico con el Árbol de Decisión y se le entrenó para identificar con los resultados de las pruebas del combinado en que posición (OLB / ILB) se podría tener mayor probabilidad de éxito al ser seleccionados, se probó el modelo exclusivamente con los jugadores de la elite y entregó un calculó con el 62% de la elite juegan como apoyadores externos OLB y el 38% son ILB, acertando al 100%.

Ahora bien, será importante, con los resultados de las próximas pruebas del combinado, probar los modelos, durante el entrenamiento y la evaluación acertaron el 100% de las veces, pero claramente estimaban que estaban 40% seguros de haber encontrado a los mejores jugadores y 60% confiados de haber identificado a los apoyadores no destacados, tanto el KNN, como el Naive Bayes predicen con el mismo grado de confianza.

Recomendación

Existe actualmente datos de la NFL, capturados por dispositivos IoT, que incluyen datos con el movimiento y el desplazamiento de los jugadores en el campo de juego, con esta big data, se pueden obtener patrones de comportamiento, que permitirían construir una prueba nueva y relevante al Combinado Scout NFL”.

Referencias Bibliográficas

Bibliografía

- Ahmed, M., Davis, V., Gamliel, S., Irizarry, R., Mastrodomenico, R., McClellan, S., . . . Westerhof, K. (2021). 50 principios de la Ciencia de los Datos (30 seconds Data Science) (Vols. ISBN 978-84-18459-51-1). (D. Breuer, T. Kitch, & N. Price-Cabrera, Edits.) Vallvidrera, Barcelona: BLUME.
- Ben-Ishay, S. (7 de June de 2020). Is the NFL Combine Related to Player Performance? Recuperado el octubre de 2022, de GitHub: https://github.com/SolB77/Is-the-NFL-Combine-Relevant-to-Player-Performance/blob/master/Is-the-NFL-Combine-Relevant-to-Player-Performance-.pdf
- Bernal, C. (2010). Metodología de la investigación. (O. Fernández, Ed.) Colombia: Pearson Educación.
- Blais, D. (14 de enero de 2018). Profesor de estadísticas de la Universidad de Colombia. (S. T. Analytics, Entrevistador)
- Bowman, D. (14 de enero de 2018). Dr. Dubois Bowman, profesor de botánica en la Universidad de Columbia. Statistical Thinking for Data Science and Analytics. (M. R. Arriaga, Traductor)
- Carroll, B., & Neft, D. (1999). Total Football II: The Official Encyclopedia of the National Football League (Vols. ISBN 978-0062701749). EEUU: William Morrow; Revised, Updated edition (4 Agosto 1999).
- Casan, S. W. (2022). Analytics of the NFL Combine (Vol. 1). (S. W. Casan, Ed.) Coppel, Texas, USA.
- Chapman, P., Clinton, J., Kerber, R., Thomas, K., Reinartz, T., Colin, S., & Wirth, R. (2000). Metodología CRISP-DM. Recuperado el noviembre de 2022, de Guía paso a paso de Minería de Datos: https://www.dataprix.com/files/Metodologia_CRISP_DM.pdf
- Decisión, C. c. (2021). Clasificación con Árboles de Decisión ¡EN 15 MINUTOS! Recuperado el noviembre de 2022, de https://youtu.be/kqaLlte6P6o
- Grades, N. P. (2022). ProFootBallFocus. Recuperado el 2022 de octubre, de NFL Players Grades: https://www.pff.com/nfl/grades/position/qb
- Hansen, M. (14 de enero de 2018). Dr. Mark Hansen, profesor de periodismo y director de Instituto of Media Innovation. (S. T. Analytics, Entrevistador, & M. R. Arriaga, Traductor)
- Herbert, J. (2019). Minería de Datos. USA: Herbert Jones.
- Herbert, J. (2020). “Data Science What the Best Data Scientists Know About Data Analytics, Data Mining, Statistics, Machine Learning, and Big Data – That You Don't” (Vol. 1). USA: Herbert, Jones.
- Hernández, S. R., Fernández, C. C., & Baptista, L. M. (2010). Metodología de la investigación (Quinta edición) (Vol. 5ta edición). México D.F., R. Hernández Sampieri, C. F. C. y P. B. L. (2006). Metodología de la investigación. En M. Rocha (Ed.), Metodología de la investigación (6ta ed.): McGraw-Hill Educación.
- Información, P. S. (2021). Introducción al aprendizaje automático. Introducción al aprendizaje automático. UFSM, Brasil.
- Kelleher, J. D., & Tierney, B. (2018). Ciencia de Datos (Vols. ISBN 978-956-14-2758-7). Santiago de Chile, Chile: Ediciones Universidad Católica de Chile | MIT.

- Lantz, B. (2019). Machine Learning with R (Vol. Tercera edición). (V. Naik, Ed.) Livery Place, Birmingham, UK: Packt Publishing Ltd.
- Lantz, B. (s.f.). Graphics Machine Learning with R. Chapter 01: Introduction Machine Learning. EEUU.
- Lewis, M. (2004). Moneyball: The Art of Winning an Unfair Game. Nueva York.
- LFPDPPP.pdf. (2010). EY FEDERAL DE PROTECCIÓN DE DATOS PERSONALES EN POSESIÓN DE LOS PARTICULARES. Recuperado el noviembre de 2022, de Diputados.gob.mx: https://www.diputados.gob.mx/LeyesBiblio/pdf/LFPDPPP.pdf
- Linebackers. (2022). Linebackers. Recuperado el noviembre de 2022, de Wikipedia: https://en.wikipedia.org/wiki/Linebacker
- Mckeown, K. (14 de enero de 2018). Dra. Mckeown Kathleen, directora del Data Science Institute Henry and Gertrude Rothschild y profesora de Ciencias de la Computación de la Universidad de Columbia. (S. T. Analytics, Entrevistador, & M. R. Arriaga, Traductor)
- Medium, M. (s.f.). K-Nearest Neighbors. K-Nearest Neighbors. Medium.
- Microsoft. (2017). Microsoft.com. Recuperado el noviembre de 2022, de Principios IA responsable de Microsoft en la práctica: https://www.microsoft.com/es-mx/ai/responsible-ai?activetab=pivot1%3aprimaryr6
- Münch Galindo, L. (2009). Métodos y técnicas de investigación. México: Trillas, 3er edición.
- NFL. (2022). NFL Big Data Bowl. Recuperado el octubre de 2022, de NFL Football Operations: https://operations.nfl.com/gameday/analytics/big-data-bowl/
- NFL. (2022). NFL Fantasy. Obtenido de Fantasy NFL: https://www.nfl.com/stats/player-stats/
- NFL. (nov de 2022). Stats. Obtenido de NFL Stats: https://www.nfl.com/stats/player-stats/
- Nwanganga, F., & Chapple, M. (2020). Practical Machine Learning in R (Vols. ISBN 978-1-119-59151-1). (E. Aguiar, & B. Seth, Edits.) EEUU: Wiley.
- Operations, N. F. (2022). The History of the Draft. (NFL) Recuperado el octubre de 2022, de NFL Football Operations: https://operations.nfl.com/journey-to-the-nfl/the-nfl-draft/the-history-of-the-draft/
- PFF. (21 de noviembre de 2022). ProFootball Focus. Recuperado el noviembre de 2022, de PFF Premium Stats: https://premium.pff.com/nfl/positions/2022/REGPO/defense?position=LB
- Sheet, B. R. (2021). Base R Cheat Sheet. Obtenido de GitHub: https://iqss.github.io/dss-workshops/R/Rintro/base-r-cheat-sheet.pdf
- Shett, D. V. (2021). Data Visualization with ggplot2::Cheat Shett. Obtenido de GitHub: https://github.com/rstudio/cheatsheets/blob/main/data-visualization-2.1.pdf
- Smith, B. (2022). Ciencia de los datos. (B. Smith, Ed.) Coppell, Texas, EEUU.
- Spiegelhalter, D. (2021). The Art of Statistics, how to learn from Data (Vols. ISBN 978-1-5416-7570-4). NY, New York, EEUU: Perseus Books, LLC.
- Today, N. D. (2011). NFL Defensive Quarterbacks : The 10 Best Middle Linebackers in Football Today. Recuperado el nov de 2022, de The Bleacher Report: https://bleacherreport.com/articles/826932-nfl-defensive-quarterbacks-the-10-best-middle-linebackers-in-football-today

- Wedell, S. (Jun de 2010). The NFL´s Top 50 Linebackers of Modern Era. Recuperado el noviembre de 2022, de Bleacherreport: https://bleacherreport.com/articles/409994-top-50-linebackers-of-the-modern-era
- Wiggins, C. (14 de enero de 2018). Dr. Chris Wiggins, asociado en matemáticas de la Universidad de Columbia. (S. T. Analytics, Entrevistador, & M. R. Arriaga, Traductor)
- Zimmer, B. (2012). The World. Recuperado en noviembre de 2022, de How Sam, Mike, and Will Became Football Positions: https://www.bostonglobe.com/ideas/2012/09/08/how-sam-mike-and-will-became-football-positions/URHq2XoAdviKJYZLQfbKNK/story.html

APENDICE A

Lista de los mejores jugadores, considerados por NFL y PFF

La lista de los mejores jugadores se conforma por aquellos que están dentro del rango temporal de estudio, desde 1987 hasta 2022, seleccionados por la lista de los 100 mejores jugadores de todos los tiempos de la NFL. Además, se incluyen los linebackers activos en la NFL que han demostrado un rendimiento destacado, acorde a PFF, como se muestra en la figura 18.

Figura 18

Lista de los Mejores Jugadores

Year	Name	College	POS	Height (in)	Weight (lbs)	40 Yard	Bench Press	ert Leap (i	oad Jump	Shuttle	3Cone	RANK
2012	Bobby Wagner	Utah State	OLB	72.38	241	4.46	24	39.5	132	4.28	7.1	1
2018	Tremaine Edmur	Virginia Tec	OLB	76.5	253	4.54	19		117			2
2017	Matt Milano	Boston Coll	OLB	72.25	223	4.67	24	35	126	4.38		4
2012	Lavonte David	Nebraska	OLB	72.63	233	4.57	19	36.5	119	4.22	7.28	5
2018	Fred Warner	Brigham Yo	OLB	75.38	236	4.64	21	38.5	119	4.28	6.9	5
2012	Demario Davis	Arkansas S	OLB	74	235	4.49	32	38.5	124	4.28	7.19	7
2019	EJ Speed	Tarleton Sta	OLB	76	224	4.6	24	34	120	4.39	6.9	8
2018	JaWhaun Bentley	Purdue	ILB	73.63	246	4.75	31	29.5	111	4.4	7.12	10
2016	DeVondre Campb	Minnesota	OLB	75.63	232	4.58	16	34	116	4.5	7.07	11
2021	Nick Bolton	Missouri	ILB	71.13	237	4.6	24	32	115	4.5	7.4	13
2017	Duke Riley	Louisiana S	OLB	72.25	232	4.58	18	34.5	122		6.89	15
2021	Ernest Jones	South Caro	ILB	73.5	230	4.72	19	38.5	126	4.38	7.49	17
2018	Leighton Vander	Boise State	OLB	76.25	256	4.65	20	39.5	124	4.15	6.88	18
2019	Sione Takitaki	Brigham Yo	OLB	73.13	233	4.63	24	37	125	4.28	7.21	19
2016	Cory Littleton	Washington	OLB	75.13	238	4.73	17	29.5	114	4.32	7.11	21
2021	Jeremiah Owusu	Notre Dame	OLB	73.5	221			36.5	124	4.15		24
2015	Denzel Perryman	Miami (FL)	ILB	70.75	236	4.78	27	32	113			25
2019	Jahlani Tavai	Hawaii	ILB	74.38	246	4.86		33.5	110	4.41		25
2015	Shaq Thompson	Washington	OLB	72.13	228	4.64		33.5	117	4.08	6.99	27
2019	Kaden Elliss	Idaho	OLB	74.25	238	4.71	20	34.5	120	4.13	6.63	28
2021	Pete Werner	Ohio State	OLB	74.88	238	4.62	20	39.5	122	4.38	6.9	29
2018	Jerome Baker	Ohio State	OLB	73.13	229	4.53	22	36.5	126	4.15	6.93	30
2015	Jordan Hicks	Texas	OLB	73.38	236	4.68	20	38	124	4.15	6.78	31
2019	Cole Holcomb	North Carol	ILB	73.25	231	4.51	22		132	4.14	6.77	32
2018	Josey Jewell	Iowa	ILB	73	234	4.82	18	33	117	4.27	6.8	33
2015	Kwon Alexander	Louisiana S	OLB	72.75	227	4.55	24	36	121	4.2	7.14	36
2018	Ben Niemann	Iowa	OLB	74.88	235	4.75	15	33.5	115	4.43	7.01	37
2019	Drue Tranquill	Notre Dame	OLB	74	234	4.57	31	37.5	122	4.14	6.94	38
2014	C.J. Mosley	Alabama	ILB	74	234	4.65	15	35	116	4.4	7.3	41
2018	Rashaan Evans	Alabama	OLB	73.88	232			30	116	4.36	6.95	41
2022	Malcolm Rodrigu	Oklahoma S	ILB	71	232	4.52		39.5	120			43
2016	Jaylon Smith	Notre Dame	OLB	74	223							44
2017	Dylan Cole	Missouri St	OLB	72.5	239	4.54	32	39	125	4.19	6.82	46
2014	Christian Kirksey	Iowa	OLB	73.75	233	4.72	16	32	122	4.42	7.11	52
2016	Myles Jack	UCLA	OLB	73	245		19	40	124			52
1997	Derrick Barnes	Oregon	OLB	72.9	261	4.92	15	33	109	4.42	7.88	54
2014	Anthony Barr	UCLA	OLB	76.88	255	4.66	15	34.5	117	4.19	6.82	56
2015	Damien Wilson	Minnesota	ILB	72	245	4.77	22	37	119	4.2	7.21	58
2022	Quay Walker	Georgia	OLB	75.75	241	4.52		32	122			59
2018	Roquan Smith	Georgia	ILB	72.88	236	4.51		33.5	117			60
2018	Foyesade Oluoku	Yale	OLB	73.88	229	4.48	18	37	123	4.12	6.94	61
2016	Elandon Roberts	Houston	ILB	71.38	234	4.6	25	36	120	4.26	7.2	62
2011	Josh Bynes	Auburn	ILB	73.63	240	4.81	21	33	116	4.32	7.11	63
2015	Eric Kendricks	UCLA	ILB	72.25	232	4.61	19	38	124	4.14	7.14	64
2018	Zaire Franklin	Syracuse	OLB	72.13	239	4.62	30	38	122	4.22	6.97	65
2022	Devin Lloyd	Utah	ILB	74.75	237	4.66	25	35	126			67
2019	Mack Wilson	Alabama	ILB	73.13	240	4.71		32	117	4.5	7.2	72
2019	Devin White	Louisiana S	ILB	72.13	237	4.42	22	39.5	118	4.17	7.07	74
2022	Troy Andersen	Montana St	OLB	75.5	243	4.42		36	128			76
2022	Christian Harris	Alabama	ILB	72.5	226	4.44		34.5	132			81

Fuente: Elaboración del autor, con datos de NFL y PPF

……
Research Article
TAU eJournal of Multidisciplinary Research
Trabajo de investigación desarrollado en el marco del Doctoral Program:
Doctor of Science in Business Intelligence (2022-2024)
TECANA AMERICAN UNIVERSITY, of the USA.
Recibido el: 7 de julio 2024
Aprobado el: 9 de julio 2024
VOL: 20
Nro.: 1
……