**TECANA AMERICAN UNIVERSITY**
**DOCTORATE OF SCIENCE IN BUSINESS INTELLIGENCE**

**INFORME No. I**

# LA CIENCIA DE DATOS APLICADA.
# Caso: La relevancia de las pruebas en el "Combinado Scout NFL" en la posición de "Linebackers" (Apoyadores) 1987 al 2022.

Autor: Mauricio R. Arriaga Guajardo

Por el presente juro y doy fe que soy el único autor del presente informe y que su contenido es fruto de mi trabajo, experiencia e investigación académica

Monterrey, N.L., México; diciembre 2022

**TECANA AMERICAN UNIVERSITY**

DOCTORATE OF SCIENCE IN BUSINESS INTELLIGENCE,

**Informe No. I**
**La ciencia de datos aplicada.**
**Caso: La relevancia de las pruebas en el "Combinado scout NFL" en la posición de "Linebackers" (Apoyadores) 1987 al 2022.**

Autor: Mauricio R. Arriaga Guajardo
Diciembre 2022

## RESUMEN

El presente informe tiene como objetivo general, aplicar la Ciencia de los Datos, para verificar la relevancia de las pruebas en el Combinado Scout NFL y 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. por consiguiente 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?; los fundamentos bibliográficos contemplan autores como, (Kelleher & Tierney, 2018) (Lantz, 2019) (Nwanganga & Chapple, 2020) (Casan, 2022) (Herbert, 2020) (Chapman, y otros, 2000), entre otros; ahora bien, el estudio abarca, la comprensión de las características necesarias en este tipo de jugador; la explicación y los resultados de las pruebas del "Combinado Scout NFL"; se estudia a través de la regresión lineal a los jugadores más exitosos (que han participado en el combinado) contra el resto de jugadores, parte de la información que se revela, incluyen datos como, los jugadores de la elite en promedio son más ligeros, más rápidos y son más fuertes que el promedio. Se descubren algunos jugadores de la elite que no tienen resultados en sus combinados, entre ellos está Ray Lewis; se comparten los modelos estadísticos en machine learning y los algoritmos para su reproducción, fueron realizadas con herramienta de R, en ambiente de RStudio, las pruebas realizadas incluyen observaciones de correlación en mapeo por pares; adicionalmente se desarrollaron dos modelos estadísticos para predecir jugadores exitosos una con KNN, y la segunda con Naive Bayes, así también se creó un tercer modelo con Árboles de decisión, el cual

provee la probabilidad estadística de jugar como linebacker interno o externo la investigación empleó, una metodología de tipo descriptiva (mixta) aplicada de manera transversal.

En conclusión, se genera conocimientos reutilizables, que aportan a subsecuentes estudios y se verifica que sí es relevante aplicar la prueba de combinados, ya que se observa una relación uniforme de las características medidas en los 300 jugadores que participan, haciendo que la liga sea competitiva y tenga estándares muy altos de cumplir, sin embargo se sugiere, utilizar los modelos elaborados con knn, árboles, así como Naive Bayes que ya fueron entrenados y usarlos con futuros resultados de los jugadores que participen en el combinado para verificar que sea posible seleccionar mejores jugadores.

Palabras Clave: “Combinado `NFL´ ”, “Data Science”, “K-NN”, “Machine learning”, “RStudio”.

# Índice General

Tabla de contenido

## Índice de Figuras



## Índice de Tablas

# INTRODUCCIÓN

**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 (Kelleher & Tierney, 2018, pág. 28); así que en esta investigación se ha aplicado convenientemente la Ciencia de los Datos, (sólo que enfocada en Football Americano) y se ha profundizado en el área de conocimiento del machine learning y sus herramientas, particularmente aplicando modelos estadísticos y algoritmos en "R", para buscar respuestas a **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?; ahora bien, hace sentido explicar que dada la definición y concepto generalmente aceptado de la "Ciencia de los Datos" es posible responder a preguntas como "qué pasó", "por qué pasó", "qué pasará" y "qué se puede hacer con los resultados; ya que se define como, el estudio de los datos con el fin de extraer la información significativa para las empresas, es un enfoque multidisciplinario que, combina los principios y las prácticas del campo de las matemáticas, la estadística, la inteligencia artificial y/o el machine learning y la ingeniería de computación para analizar grandes cantidades de datos; este análisis permite que los científicos de datos planteen y respondan a preguntas como las ya mencionadas" (Herbert, 2020); para tal efecto el documento se ha estructurado, iniciando con una breve introducción y especificando el objetivo general, así como los objetivos especificos y se han integrado cuatro capítulos; en el "Capítulo 1", incluye, los antecedentes que iniciaron esta investigación, en el cual se explican los orígenes de las pruebas del "Combine scout NFL" y se detalla la problemática identificada; se sustenta la justificación de realizar el trabajo, se acota el alcance; se delimita el problema, formulando la pregunta de investigación; también se menciona ***la metodología requerida en esta investigación, siendo descriptiva (mixta) de manera transversal,*** adicionalmente incluye las técnicas y los instrumentos requeridos para la recopilación de los datos; en el "Capítulo 2", detalla la teoría fundamental para sustentar la investigación, definiendo el marco teórico requerido, para la disertación de la Ciencia de los Datos, incluyendo, el proceso de la minería de datos, el aprendizaje automático (machine learning) y la aplicación de estadística avanzada en la herramienta RStudio, que permita generar conocimientos repetibles, para futuras investigaciones; el Capítulo 3, describe los resultados y la

principal contribución de esta disertación, con los algoritmos estadísticos empleados y los modelos en machine learning “R”, que persiguen la búsqueda de la respuesta a la pregunta de investigación, explica el desarrollo del estudio y despliega los resultados.

Finalmente, el Capítulo 4 presenta la conclusión del estudio de la investigación y las recomendaciones para futuros trabajos relacionados.

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?

**Objetivo General**

El presente informe tiene como objetivo general: Aplicar la Ciencia de Datos, para verificar la relevancia de las pruebas en el Combinado Scout NFL y 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.

**Objetivos específicos**

- Describir los antecedentes de la investigación, el alcance, el problema, la justificación y la metodología
- Disertar La Ciencia de los Datos, que de soporte a la investigación
- 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.

# CAPÍTULO 1

## PLANTAMIENTO 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

### 1. Antecedentes que dan inicio a la investigación

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 (Operations, 2022); sobre el evento detalla el Dr. Casan (2022) que los jugadores participantes, son sometidos a pruebas tanto mentales como físicas por un grupo de caza talentos (en inglés "scouts") así como por los gerentes generales y los entrenadores de la NFL; los atletas participan sólo por invitación y en dicho evento se obtiene data de las características físicas, sus exámenes médicos, las entrevistas que se les realizan, las mediciones de estatura, la velocidad de los jugadores en las 40 yardas, el salto de altura, el salto de longitud, la fuerza, la agilidad de correr ida y regreso, entre otras pruebas de tipo físicas y mentales (p. 1); sin embargo con la proliferación de este evento, por los medios de comunicación (con canales como "Fox Sports", "NFL channel", "ESPN", entre otros) un sin fin de fanáticos se han involucrado y siguen las transmisiones y las estadísticas del "Combine Scout NFL" y el magno evento del día de "DRAFT de la NFL" (el DRAFT se refiere al día que cada equipo pasa por turnos y selecciona al jugador que considera con más alta probabilidad de aportar triunfos al equipo); a consecuencia, los analistas deportivos y los fanáticos, han comenzado a cuestionar cada vez con mayor intensidad, la utilidad de las pruebas del "Combinado", se han publicado diversos artículos en el que dudan de la relevancia y trazabilidad de las pruebas, contra el desempeño real que tienen los jugadores como profesionales

### 1.1 Descripción del problema

Existen varios autores que comentan que las mediciones recolectadas por las pruebas de combinados de la NFL, no son necesariamente relevantes en el resultado del desempeño en el juego y la carrera del jugador en la NFL (Casan, 2022); así también el ingeniero y atleta Sol (Ben-Ishay, 2020) en su estudio publicado, refleja la poca o nula correlación del éxito en ciertos jugadores y no encuentra datos contundentes de que las pruebas hayan colaborado a la selección adecuada de los mismo; por tanto es prudente saber, si existe una relación real, medible y rastreable entre las pruebas realizadas en el "NFL Scouting Combine" y el buen desempeño de los jugadores en la posición de Linebacker, que son seleccionados en cada evento; ya que de otra manera se estarían gastando los recursos económicos, las emociones y muchos esfuerzos de manera completamente innecesarios.

### 1.2 Pregunta de investigación

**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?;

### 1.3 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.

### 1.4 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".

**L**a 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.

### 1.5 Metodología empleada

Explican Hernández, Fernández & Baptista (2010) que dado el propósito fundamental de este trabajo se circunscribe dentro de un tipo de investigación descriptiva, entonces, la investigación se puede realizar bajo dos enfoques: el cualitativo y cuantitativo; puntualiza Münch (2009) que la investigación de tipo “descriptivo” incluye características como, la frecuencia, la aparición y el desarrollo (p. 22).

### 1.6 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

### 1.7 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); 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).<br>Documentos: Datos estructurados y no estructurados disponibles en formato CSV o 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 acota al ciento por ciento de los jugadores, en la posición de Linebackers, (apoyadores) participantes registrados en el "Combine Scout NFL" desde 1987 hasta 2022; adicionalmente se usa la lista de los 100 mejores apoyadores de la NFL. |
| Técnica | Técnica de revisión documental digital por internet (secundarios) |
| Herramientas | Machine Learning con "R"; Excel y RStudio |

**Fuente:** Elaboración del autor, con datos de investigación

### 1.8. 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 expresa el marco teórico conceptual de la Ciencia de los Datos que acorde a Keller & Tierney (2018) contempla, el conjunto de los principios, la definición del problema, los algoritmos y los procesos para extraer patrones no obvios y útiles de grandes conjuntos de datos y así también está relacionada con el uso de **aprendizaje automático (**Machine learning) y la "Minería de Datos" (p. 13).

## 2 La Ciencia

La ciencia acorde a Münch (2009) "puede ser conceptualizada en su sentido más amplio, como un conjunto sistemático de conocimientos con los cuales, al establecer principio y leyes universales, el hombre explica, describe y transforma el mundo que lo rodea" (p. 13); agrega Bernal (2010) que el significado de la ciencia, esta fundado en un sentido histórico, en el espacio en el que la sociedad lo interpreta y usa, sin embargo hay evidencia consensada, en que es racional y se logra construir de manera sistemática (p.286).

## 2.1 Los Datos

Acorde a Kelleher & Tierney (2018) a los sucesos, a las personas y a los objetos, se les puede representar (de manera sintetizada) como los "Datos", siendo estos una representación abstracta de la realidad, en los cuales se les puede adjudicar, por su propia naturaleza atributos, características y variables, y como resultado denotan una abstracción particular y única de la realidad, esto permite identificarles y agruparles en conjuntos de datos con similitudes (p. 35); al extraer estos datos en bruto Herbert (2019) profundiza que al pasar a través de las actividades asociadas a la ciencia de los datos y la minería (como la transformación, la agregación, la

limpieza, entre otras) debe resultar en información valiosa y útil, véase la pirámide, en la figura 1, propuesta por Han, Kanber y Pei 2011(pp. 9,10).

**Figura 1**

*La Pirámide Refleja en la Parte Inferior, los Datos Crudos, en la Cima el Conocimiento*


**Fuente:** Tomada de la Figura 2 de Kelleher & Tierney (2018, pág. 45)

Acorde a Kelleher (2018) los tipos de los datos se pueden agrupar en, los *numéricos, los nominales, y los ordinales*, cada uno de ellos provee distintas capacidades, los numéricos permiten hacer cálculos matemáticos, son comprendidos con números enteros y/o reales; mientras que los datos nominales son aquellos elementos, que se distinguen por un sistema de nombre simple, son datos sin valor numérico, un ejemplo sería, la profesión (ingenieros, arquitectos, administradores); adicionalmente logran describir las características y los atributos, siendo que no tienen una jerarquía y sólo sirven para clasificar; mientras que los datos ordinarios se colocan en algún tipo de orden por su posición en la escala, por ejemplo pueden indicar superioridad, sin embargo, no pueden hacer operaciones aritméticas, ya que sólo muestran una secuencia (primero, segundo, tercero).(pp. 36,38); adicionalmente se pueden considerar que existen siete categorías de datos, los datos estructurados, los datos no estructurados, datos de lenguaje natural, datos generados por máquina, datos basados en gráficos, datos de audio, video e imagen y la transmisión de datos; para reconocerlos fácilmente, **los datos estructurados** pueden ser almacenados y o desplegados en una tabla y cada instancia tiene los mismos atributos, ya sean fechas, nombres, cantidades, son iguales en su tamaño y definición (normalmente definidos como, números enteros, números con signo (negativos), científicos que son números de coma

flotante (decimales), cadenas alfanuméricas (y unicode) de estados o boléanos); mientras que *los datos no estructurados*, cada dato tiene su propia estructura única, como los videos, los correos electrónicos, el audio, las imágenes, los documentos, las páginas web(Smith, 2022, págs. 8-11).

### 2.2 La Ciencia de los Datos

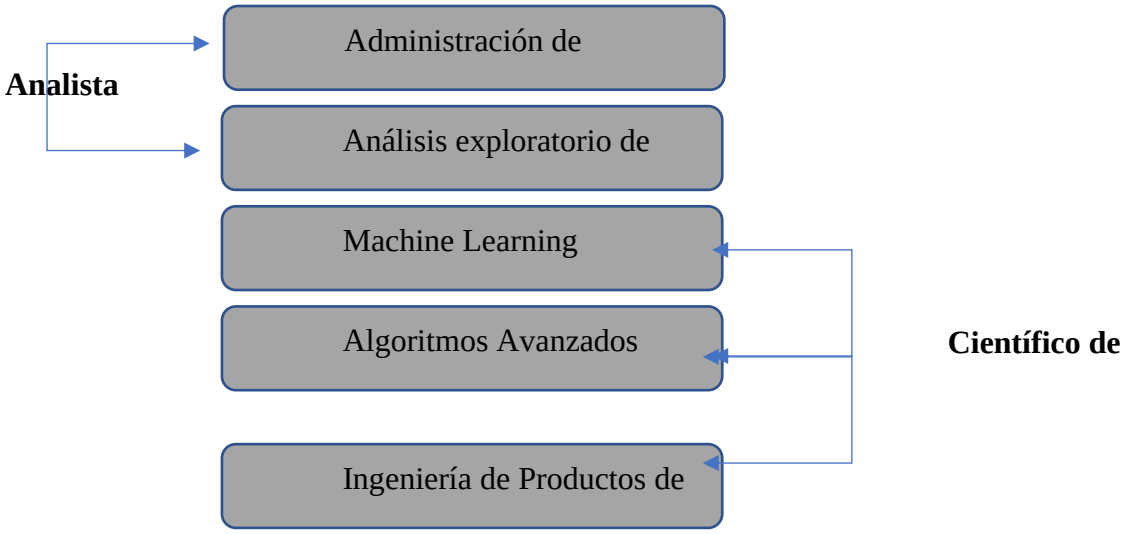
La Universidad de Columbia en su departamento especializado en la Ciencia de los Datos, compuesto por profesores de muy distintas disciplinas, han sido entrevistados, durante el seminario llamado ”Pensamiento estadístico para ciencia de datos y análisis” durante el año 2018, en el cual definieron la Ciencia de los Datos de las siguientes maneras, la Dra. Kathleen, menciona que, la gente que genera tecnología, generan datos y deben de colaborar con personas que se dedican a resolver problemas (Mckeown, 2018); sintetiza el Dr. Blais (2018) se refiere a construir herramientas con los datos obtenido, para resolver problemas; mientras que para el Dr. Dubois Bowman, profesor de botánica, explica que consiste de una línea de producción que comienza, con la extracción de los datos, sin estar refinados o limpios, decodificarlos, para posteriormente integrarles y/o agruparles entre ellos, facilitando el proceso a través de la minería de datos, que lleve al poder analizar e interpretar como información (Bowman, 2018); cierra el seminario la directora de los programas estratégicos, del departamento de Ciencia de los Datos, de la Universidad de Columbia, y predice que es un momento de cambios de paradigmas, antes de ahora, las computadoras sólo ejecutaban lo solicitado, y ahora con la tecnología de “Machine Learning (ML) parte de la Inteligencia Artificial (AI) a través de las máquinas habrá una revolución de adelantos en todas las ramas del conocimiento, las perspectivas son múltiples y las expectativas grandes de lo que veremos en el futuro inmediato; sustentado del seminario cursado por el autor, en la Universidad de Columbia “Statistical Thinking for Data Science and Analytics” de la Universidad de Columbia , 14 de enero 2018; dentro de los elementos mencionados previamente explica Herbert (2020) que el aprendizaje profundo (deep learning), así como aprendizaje automático (machine learning) son aplicaciones de la Inteligencia Artificial (IA) (p. 143) y son los analistas de datos y los científicos de datos, quienes trabajan la información, con perspectivas alineadas pero tareas y campos de cobertura distintos(p. 9).

### *2.2.1 Los Científico de Datos*

Los científicos de datos explica Herbert (2020) utilizan diferentes perspectivas y distintos ángulos, para entender la historia que cuentan los datos, se apoyan en el uso de herramientas especializadas, y la creación de modelos inteligentes a los cuales les pueden suministrar datos constantemente y como resultado se obtienen mejoras y se pueden hacer los ajustes y predicciones pertinentes, un ejemplo en lo que trabaja un científico es el automóvil autónomo de Google, entre más lo usan más aprende y se hace más eficiente el modelo (p. 11); en la figura 2, véase a continuación, se muestra que mientras el analista de datos explica la historia de los datos; el científico de datos, va más allá y se apoya en algoritmos avanzados (estadística, probabilidad y matemática).

**Figura 2**

*Campos de Acción del Analista de Datos y del Científico de Datos*




**Fuente:** Tomada de Herbert (2020, pág. 10) quien recuperó de (Chapman, y otros, 2000)

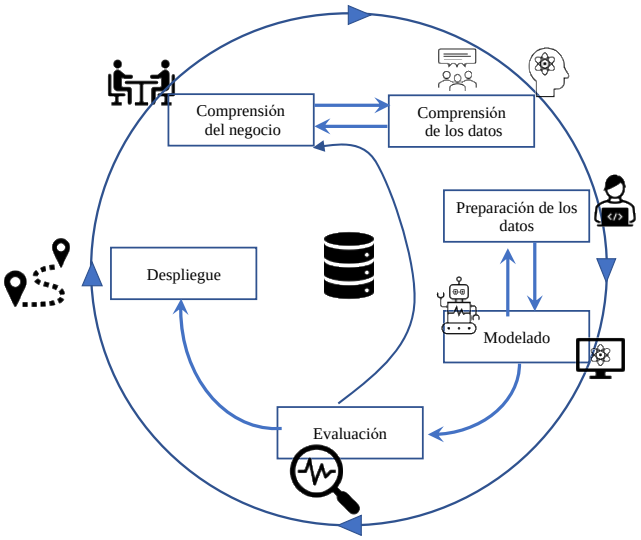
## 2.3 La Minería de Datos

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 la Minería de Datos, También Conocido Como del Ciclo de Vida CRISP-DM




**Fuente:** Tomada de (Kelleher & Tierney, 2018, pág. 47) quien lo basaron en (Chapman, y otros, 2000)

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

Las dos primeras fases, permiten al científico de datos investigar ¿qué es lo que el negocio hace? y ¿qué es lo que desean obtener como resultado?, así también es el momento correcto, para explorar qué tipos de datos están accesibles (por diferentes fuentes) y ser capaz de entender su descripción, con ello decidir si los datos son o no de utilidad (Kelleher & Tierney, 2018, pág. 46).

### *2.3.2 Fase III Preparación de los datos*

La fase de **preparación de datos** cubre las actividades que construyen el conjunto de los datos finales, son estos datos los que serán ingresados a las herramientas de modelado, como R, y son los datos en brutos iniciales (Chapman, y otros, 2000); especifica Herbert (2019) que la preparación de los **datos**, inicia desde la **identificación de la fuente** que genera los datos, ya sea un teléfono inteligente, una base de datos o cualquiera que sea el origen, identificadas éstas, se

deberán obtener los datos con las **herramientas de recolección** (que se detallan más adelante en el punto 2.7, así como en el anexo B) y se requerirá llevar a cabo la tarea de **limpieza de datos** (esto es debido a los errores e inconsistencias de los datos recopilados) (pp. 42-52); las actividades referentes a la preparación de los datos, explica Kelleher (2018), que muy factiblemente serán repetidas en múltiples ocasiones (y no necesariamente con un orden similar) las actividades contemplan la selección de tablas, de los registros, y de los atributos, **también la transformación, la agregación, la concatenación, la eliminación**, la fase concluirá al verificar la calidad e integridad de los datos (se debe considerar que los datos requerirán un almacenaje conveniente en costo y velocidad de acceso); con la verificación y validación de los datos concluida, se podrá ingresar a la siguiente fase "Modelado" (P. 48);

### *2.3.3 Fase IV el Modelado*

Ahmed y otros (2021) explican que el **modelo y/o la modelización**, son "los procedimientos o problemas del mundo real en **términos matemáticos**, pueden ser simples o muy complejos, se usan a menudo para hacer predicciones y pronósticos"; por otro lado, definen **el algoritmo**, el cual es un conjunto de instrucciones o cálculos, diseñados para que sean ejecutados por un ordenador, escribir algoritmos recibe el nombre de codificación o **programación informática**, el resultado de un algoritmo puede ir desde una suma de dos números hasta el movimiento de un automóvil autónomo. (p. 15); el cómo funciona la minería de datos, describe Herbert, (2019) que esta disciplina está constituida por técnicas y métodos que aportan y facilitan el análisis, se divide en tres partes principales, **el modelo descriptivo**, el cual busca el encontrar similitudes en la descripción y poder agrupar los datos históricos, ayuda a entender el porqué de las cosas, algunas de las técnicas que usa son, la agrupación, aprendizaje de reglas de asociación (detección de relaciones de registros), análisis de componentes principales y la agrupación de afinidades; el segundo es el **modelo predictivo** ayuda a estimar resultados desconocidos, y permite clasificar eventos en el futuro, las técnicas que lo conforman son, las redes neuronales (IA, ayuda a detectar patrones), la regresión (fuerte relación entre variable dependiente de la independiente), los árboles de decisiones, las máquinas de vectores; por último **el modelo prescriptivo**, el cual analiza las variables internas y externas, así como las restricciones, logrando aportar recomendaciones que permitan tomar decisiones y cursos de

acciones, las técnicas que le componen, optimización de marketing, el análisis predictivo más reglas (pp. 17-19).

### *2.3.4 Fase V Evaluación*

La guía que comparte Chapman y otros (2000) menciona que, para generar la prueba del diseño, “antes de que en realidad se construya el modelo, se debe generar un procedimiento, es decir que se debe tener listo el mecanismo para **probar la calidad y validez del modelo**; por ejemplo, típicamente se separa el conjunto de datos en una serie (conocida y probada) en dos conjuntos de pruebas, unos será usado para entrenar el modelo (con lo resultados de los cuales se debe crear el modelo de manera automática) el segundo conjunto de series (con los datos también conocidos), se usará para suministrarse al modelo entrenado y arroje la predictibilidad y precisión del modelo; así se puede conocer que es lo que debe ajustarse para mejorar el modelo (esto se hace reiteradamente” (pp. 14 - 16).

### *2.3.5 Fase VI Despliege*

El llevar a cabo el despliegue del modelo a producción, implica examinar la mejor estrategia para que el grupo de usuarios (con problemas similares) tomen ventaja y se beneficien de los modelos realizados;

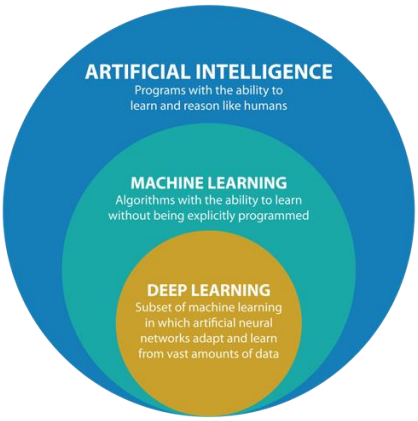
Finalmente es importante considerar, que los procesos son iterativos, existe un círculo externo a todas las fases, lo cual implícitamente resalta que todo el proceso es iterativo, puede ser por mejoras, obsolescencias, actualizaciones o cualquier otra situación. (Kelleher & Tierney, 2018)

## 2.4 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, Mínería de Datos, Machine Learning, Deep learning*




Fuente: Tomada de (Información, 2021) recuperado de https://www.ufsm.br/pet/sistemas-de-informacao/2021/05/11/introducao-a-machine-learning/

### *2.4.1 Machine learning llevado a la práctica*

Considera Lantz (2019) que para aplicar el ML en tareas del mundo real, se utiliza el proceso de **cinco pasos**, primero, ***recopilación de datos***: el paso de recopilación de datos implica recopilar el material de aprendizaje que utilizará un algoritmo para generar conocimiento procesable, en la mayoría de los casos, los datos deberán combinarse en una sola fuente, como archivo, hoja de cálculo o base de datos; segundo la ***exploración y preparación de datos***, la calidad de cualquier proyecto de aprendizaje automático se basa en gran medida en la calidad de sus datos de entrada, se requiere arreglar o limpiar los llamados datos "desordenados", eliminar datos innecesarios y recodificar los datos para ajustarse a las entradas esperadas del alumno; el tercer paso es el ***entrenamiento de modelos***, para cuando los datos se hayan preparado para el análisis, es probable que se tenga idea de lo que es capaz de aprender de los datos, la tarea específica de aprendizaje automático elegida informará la selección de un algoritmo apropiado, y el algoritmo representará los datos en forma de modelo; el cuarto paso, es la ***evaluación del modelo***, cada modelo de aprendizaje automático da como resultado una solución sesgada al problema de aprendizaje, lo que significa que es importante evaluar qué tan bien aprendió el algoritmo de su experiencia. Según el tipo de modelo utilizado, es posible que se pueda evaluar la precisión del modelo mediante un conjunto de datos de prueba o que se necesite desarrollar

medidas de rendimiento específicas para la aplicación prevista; ahora bien durante el quinto paso se debe ***mejorar del modelo***, si se necesita un mejor rendimiento, se hace necesario utilizar estrategias más avanzadas para aumentar el rendimiento del modelo (p.18).

### *2.4.2 Técnicas de aprendizaje Supervisado y No supervisado*

El ***modelo predictivo*** señala Lantz (2019) que permite predecir valores, apoyandose en un grupo de datos (conocidos); los modelos predictivos, no exclusivamente tienen que prever eventos futuros, por ejemplo, se podría usar para predecir eventos en el pasados, también se utilizan en contextos en tiempo real, como al controlar los semáforos en operación; dentro de los procesos de entrenamiento, existen los ***modelo predictivo, se les conoce como aprendizaje supervisado;*** La supervisión no refiere a la participación humana, refiere a aprender los patrones basados en ejemplos de datos etiquetados e identificados del pasado; la tarea más comunmente utilizada por machine learning, ***es la clasificación***, como ejemplo identificar spam en el correo electrónico; por su lado el aprendizaje **no supervisados,** buscan descubrir patrones sin la ayuda de datos etiquetados (p.21); en contraste, el ***no supervisado***, explican Nwanganga & Chapple (2020), queen el modelo descriptivo, ninguna característica individual es más importante que otra, debido a que no hay un objetivo para aprender, el proceso de entrenar un ***modelo descriptivo se llama aprendizaje no supervisado, p***or ejemplo, se puede utilizar para detectar patrones de comportamiento fraudulento, detectar defectos genéticos o identificar puntos. La tarea de modelado descriptivo de dividir un conjunto de datos en grupos homogéneos se ***denomina agrupación, a*** veces esto se usa para el análisis de segmentación, que identifica grupos de personas con un comportamiento o información demográfica similar (p. 12)

#### **2.4.2.1 Técnicas de Clasificación**

Existen tres tipos principales de conocimiento que podemos aprender de los datos; el primero, son las ***técnicas de clasificación***, entrenan modelos que permiten predecir la pertenencia **a una categoría (no númerica)** ejemplo educación (ninguna, primaria, secundaria, preparatoria); la segunda, incluye las ***técnicas de regresión*** que permiten predecir un resultado **numérico**; la tercera, define las ***técnicas de aprendizaje de similitud (clustering o de Agrupación)*** ayudan a descubrir las formas en que las observaciones en nuestro conjunto de

datos se parecen y difieren entre sí (Nwanganga & Chapple, 2020, pág. 14); por su lado Lants (2019) acota que para trabajar un proyecto en machine learning, se requiere identificar cual de estas cuatro técnicas representan los datos, la ***técnica de clasificación, los patrones númericos (regresiones), la detección de patrones (también llamada clustering (agrupamiento)) y la clasificación de errores*** está divididas de acuerdo con su propósito.(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)

### 2.5 Machine learning con "R"

Acorde a Nwanganga & Chapple (2020) el lenguaje de programación R comenzó en 1992 como un esfuerzo por crear un lenguaje de propósito especial para uso en aplicaciones estadísticas, usado por los científicos de datos y los analistas de negocios de todo el mundo; es un lenguaje libre y de código abierto, desarrollado por una comunidad de comprometidos desarrolladores; casi cualquier nueva técnica de aprendizaje automático (creada hoy en día) está disponible rápidamente para los usuarios de R en un paquete distribuible, que se ofrece como

código fuente abierto en “Comprehensive R Archive Network” (CRAN); adicionalmente han empoderado y hecho más amigable al lenguaje de programación “R” a través de la herramienta con un ambiente integrado de desarrollo (IDE) llamado “RStudio” el cual permite hacer multiples tareas de manera de ambiente gráfico (p. 26).

### *2.5.1 Estructura de Datos en “R”*

Explicado por Lantz (2019) “R” cuenta con numerosos tipos de estructuras para manejar los datos, se utilizan para el análisis de datos estadísticos, algunas estructuras de R que se utilizan con mayor frecuencia en el aprendizaje automático son, l**os vectores, los factores, las listas, los arreglos, las matrices y los marcos de datos(data frame);** la estructura de datos fundamental de R es el ***vector***, que almacena un conjunto ordenado de valores llamados elementos, todos sus elementos deben ser del mismo tipo; por ejemplo, un vector no puede contener tanto números como text0 (p. 30); agregan Nwanganga & Chapple (2020) que el factor, es un caso particular de los vectores, es únicamente usado para representar categorías o variables ordinales; a su vez las listas son estructuras, similares a los vectores, en ellas se guardan en orden un grupo de elementos, lo importante es que los elementos pueden ser de diferente tipos (a diferencia del vector que sólo es un tipo de elemento); por su lado los arreglos, son similares a una hoja de cálculo que tiene dos dimenciones, renglones y columnsa; en contraste las matrices, permiten sólo un tipo de elemento, normalmente númerico y es comunmente usada para operaciones matemáticas; por último los marcos de datos, es la estructura con mayor importancia,es análoga a la hoja de cálculo y a la base de datos, todas contienen renglones y columnas, pero el data frame puede ser una lista de vectores o factores, convina tanto a los vectores como a las listas.(pp. 46-50).

### *2.5.2 Explorando los Datos en “R”*

Explicado por Lantz (2019) que “R” y la paquetería disponible, permite llevar a la práctica todo el ciclo necesario para la gestión de los datos, desde la carga de la información en diferentes formatos, la limpieza de los datos, la transformación, inclyendo el análisis estadístico y el desplegado de datos, por tal motivo se ha adicionado en el anexo “A”, algunos de los comandos más comunes y necesarios, que son usados para los propósitos mencionados (p. 44)

#### *2.5.2.1 “R” estadísticas básicas*

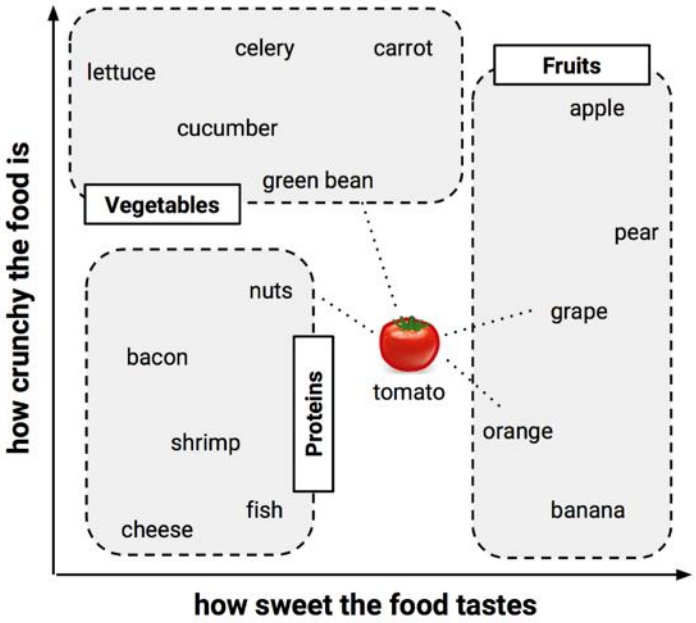
Profundizan Nwanganga & Chapple, (2020) que la **“estadísticas descriptivas”** son útiles en la exploración y comprensión de datos; implica la utilización de estadísticas para describir el **comportamiento de los datos y su distribución**, la denominada **moda** nos dice qué valor ocurre más para esa característica, **la frecuencia**( detecta cuantas veces se repite el dato) y junto con la moda se usan típicamente para describir datos categóricos. Para datos continuos, las medidas como la **media y la mediana** a menudo se usan para describir las propiedades de los datos; En la estadística, la correlación es el grado de relación que existe entre dos variables, es decir, se dice que dos variables están correlacionadas cuando el aumento o disminución de una provoca un cambio en la otra (p. 63); **los cuartiles** son útilizados comunmente en el análisis más básico de la distribución de los datos define Lantz (2019) que el cuartil mínimo, representa el inicio de la serie de datos, sucesivamente, el 1er cuartil refleja el 25%, el segundo cuartil (50%) y es el mismo número de la serie, que refleja la media, así que el tercer cuartil representa Q3 (75%) y evidentement el Q4 representa el final de la serie (pp. 48-50).

#### *2.5.2.2 K-vecino más cercanos (K-Nearest Neighbors)*

Para desarrollar este modelo explica Lantz (2019) sólo hace falta suministrar los datos del conjunto de entrenamiento y elegir cuántos vecinos se considerarán en su vecindario, la única limitante es, que requiere que los datos estén almacenados y no es entrega un modelo al arrojar los resultados, algunas fortalezas, simple y efectivo, no asume la distribución de los datos, es rápido de se entrenado, ahora bien algunas debilidades son, no produce un modelo (por ende no se puede saber cómo relacionó los datos), es necesario seleccionar un parámetro k, la fase de clasificación es lenta, en los casos que hay funcionalidades nominales y datos incompletos, se requerírá procesos adicioanles para resolver (pp. 66,67). Se muestra en figura 5 un ejemplo de clasificación

**Figura 5**

*K-Nearest Neightbor*




Fuente:Tomadas de (Lantz, Gráficos Machine Learning with R)

#### *2.5.2.3 Naive Bayes o Redes Bayesianas*

Acorde a Nwanganga & Chapple (2020) utilizan una tabla de probabilidades para estimar que tan probable sea que una instancia pertenezca a una clase en particular, el enfoque de Bayes se basa en la premisa de que la probabilidad de eventos anteriores puede ser una buena estimación de la probabilidad de eventos futuros, por ejemplo, al pronosticar la probabilidad de lluvia para hoy, informaríamos sobre la proporción de días anteriores con las mismas condiciones climáticas que hoy, en los que llovió, entonces, llovió 4 de cada 10 de esos días, entonces estimamos un 40 por ciento de probabilidad de lluvia hoy. Este enfoque es útil en varios dominios y áreas problemáticas, otro ejemplo del uso de Bayes es el filtrado de spam para etiquetar correos electrónicos no vistos en función de cómo se etiquetaron correos electrónicos similares anteriores (p. 250). La siguiente figura 6 muestra la representación lógica de la probabilidad, de que el evento B suceda, dado que que hay un evento A.

**Figura 6**

*La Probabilidad de que el Evento B Suceda, Dado que Hay un Evento A, Desarrollada por Naive Bayes*

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

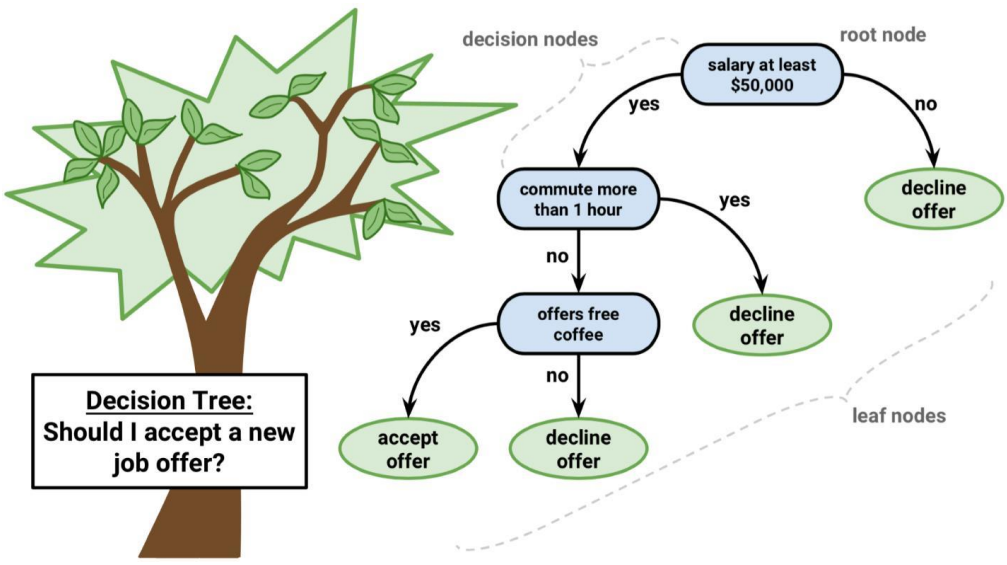
**Fuente:** Tomada de (Lantz, 2019, pág. 94)

#### *2.5.2.4 Árboles de Decisión*

Definen Kelleher & Tierney (2018) que es "un tipo de modelo de predicción que codifica las reglas [si-entonces] en una etructura de árbol, es decir predice el resultado de la probabilidad que sucede el evento dado en un camino dado" (p. 159). En la figura 7 se muestra un ejemplo de árbol de decisión.

**Figura 7**

*Árbol de Decisión*



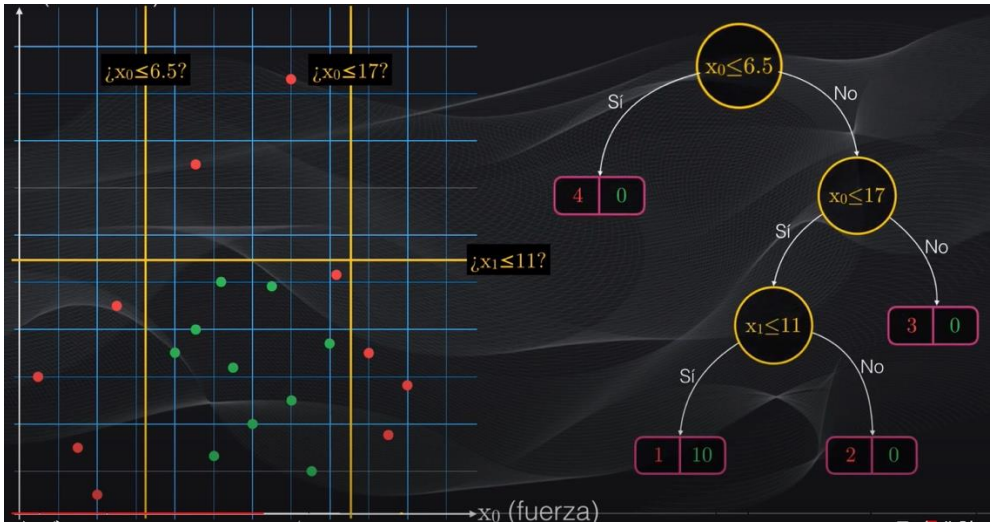

Fuente: (Lantz, Gráficos Machine Learning with R) https://static.packt-cdn.com/downloads/9781788295864_ColorImages.pdf

### *2.5.2.5 Árboles de Regresión CART (Classification and Regression Trees)*

Los árboles de decision son tal vez el algoritmo más sencillo, pero a la vez es uno de los más poderosos del machine learning y son usados cuando se tienen grupos de datos, relativamente complejos, sirven para resolver **problemas de clasificación**, especificamente el **algoritmo CART**, el modelo aprende a calcular una frontera de decisión que pueda calcular el dato, a una u otra categoría, un ejemplo es la clasificación de correo spam, una gran ventaja es que permite saber los resultados de la predicción ( es decir que modelo aplicó al tomar la decisión), el más usado se llama CART ( árboles de clasificación y sucesión), la figura 8 muestra las fracciones binarias y la expresión gráfica del arbol de decisiones y clasificacion y cálculos requeridos.

**Figura 8**

*Árboles de Decisión*




Fuente: (Decisión, 2021)

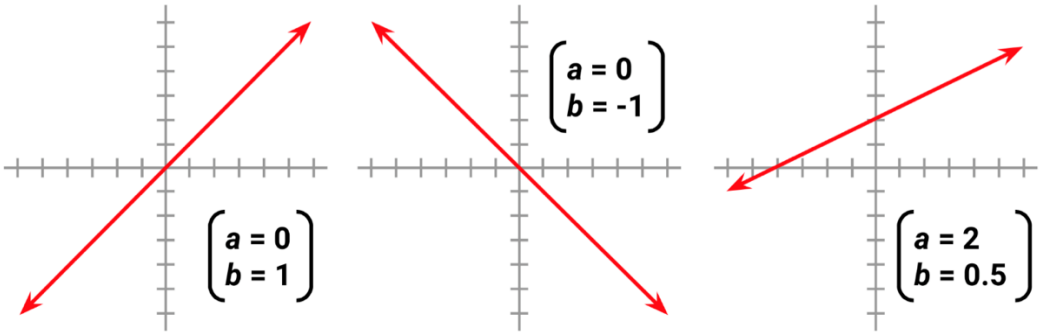
### *2.5.2.6 Regresión lineal*

Acorde a Nwanganga & Chapple (2020) las técnicas de regresión son una categoría de algoritmos de aprendizaje automático que buscan predecir una respuesta numérica cuantificando el tamaño y la fuerza de la relación entre valores numéricos; agrega que la correlación es un término estadístico utilizado para describir y cuantificar la relación entre dos variables.

Proporciona un único valor numérico de la relación entre las variables, que se conoce como coeficiente de correlación (pp. 103,106); sintetiza la explicación de regresión lineal Lant (2019) es un modelo de regresión lineal simple define la relación entre una variable dependiente y una única variable predictora independiente utilizando una línea definida por una ecuación de la siguiente forma: y=(alfa)+Bx esta ecuación es prácticamente idéntica a la forma pendiente-intersección, la intersección, a (alfa), describe dónde la línea cruza el eje y, mientras que la pendiente, ß (beta), describe el cambio en y dado un aumento de x. (p.172); vease la figura 9 para esquematizar lo explicado.

**Figura 9**

*Regresión Líneal*




Fuente: (Lantz, Gráficos Machine Learning with R)

### 2.6 Ética y privacidad

En referencia a la ética que se debe tener al usar los datos , explica el Dr. Spiegelhalter (2021) que es importante considerar, que existe un fuerte potencial de hacer mal uso de los datos, particularmente cuando se obtienen de las cuentas en redes sociales, mientras que la información gubernamental tiene mayores restricciones y es de mayor veracidad, afirma que la ética de los datos aún es una disciplina en desarrollo; es importante que los algoritmos utilizados , que puedan tener impacto en las personas, sean honestos, reproducible por la ciencia y que provean confiabilidad al comunicarse(p. 371); adiciona Lantz (2019) que el aprendizaje máquina ayuda a entender de manera que haga sentido el mundo, sin embargo como toda herramienta es posible usarla para el bien o el mal, se ha observado que se ha utilizado en observar a los humanos como

si fueran conejillos de india, para entenderlos y ofrecerles productos de mercadotecnia sin ninguna restricción, es por eso que sugiere que debe usarse a través del arte de la ética, menciona también que Google a integrado un principio en la recolección de datos, que pide no se malévolo con ellos, algunos países como los Europeos integraron principios del manejo de los datos en el "Genaral Data Protection Regulation" (GDPR) documentación que dicta que y como deben ser protegidos los datos(p. 7); en México se aplica "la Ley Federal de Protección de Datos en propiedad de los particulares", la cual obliga a las personas que manejan datos de indole personal personales, a proteger y mantener la privacidad, asegurando que sólo son usados con el objeto indicado en previa aprobación de recolección del dato, por el usuario final, el no acatar el manejo correcto de los datos, con lleva a multas y hasta a la cárcel (LFPDPPP.pdf, 2010); mientras que, la empresa de Microsoft que al conocer el problema inherente que con lleva desarrollar en Inteligencia Artificial, machine learning y el uso de los datos, conformó un organizmo el cual ha desarrollado varios principios de legalidad, equidad, confiabilidad & seguridad, inclusión, transparencia y responsabilidad (Microsoft, 2017); existen dos enfoques para preservar la privacidad, explica Kelleher (2018) sobre la privacidad diferencial y el aprendizaje federado, la privacidad diferencial consiste en un acercamiento matemático al problema de aprender información en referencia a la población, que al mismo tiempo no aprende nada de las personas (así que no es comprometida su privacidad), con esa misma intención de proteger al individo el marco federado lleva los algoritmos y modelos hasta la base de datos y sistemas que resguardan la información, así que se ejecuta de manera distribjida, manteniendo la seguridad y privacidad de los datos (p. 138); el marco legal para regular el uso de los datos y la protección, acota claramente lo que es considerado los pilares centrales (en la mayoria de las juridicciones) temas que protegen de la discriminación por edad, etnia, raza, sexo, nacionalidad, orientación sexual, religión y política (Kelleher & Tierney, 2018, pág. 140)

### 2.7 Herramientas de ciencia de datos hacia la ciencia de datos

Ilustra Herbert (2020) que la ciencia de los datos, requiere del apoyo de herramientas especializada en diversos temas de acción, para simplificar el trabajo que realizan los científicos de datos en el proceso de limpieza, modelado, transformación, análisis de datos y presentación (pp. 56-65).

### *2.7.1 Herramientas de datos denominadas "Abiertas"*

Cita Helbert (2020) algunas de las herramientas de código abierto y por ende sin costo asociado, son; OpenRefine, Orange, Knime, R & RStudio, RapidMinder, Pentaho, Weka, Node XL, Gelphi, Talend, como referencias, se ha agregado una tabla explicativa sintetizada en el anexo B (págs. 56-63)

### *2.7.2 Herramientas de Visualización de Datos"*

Acorde a Herbert (2020) enlista las siguientes herramientas dentro del grupo comúnmente conocido como Data Viz o bien visualizador de datos, cada uno tiene sus características, **Datawrapper**, permite construir graficas interactivas, el usuario puede leer datos de tipo CSV, Excel o PDF; por su lado Tableau Público, democratiza la visualización de datos, sencillo de usar y poderoso, facilita el análisis de datos; el programa Infogram, contiene 35 gráficas interactivas y más de 500 mapas que facilitan la visualización de la información; sin embargo también Google Fusion Tables, poderoso visualizador de datos cuando se tiene grandes volúmenes de información que desplegar; el Solver, regularmente utilizado para las finanzas, para manejar presupuestos y análisis de data; se incluye el OpenText "Este es un motor de clasificación especializado aplicado en la identificación y evaluación de expresiones y patrones en un contenido textual. El análisis se lleva a cabo a nivel de documento, oración y tema"; por su lado el Trackur, trabaja con el análisis de sentimiento alimentado por las redes sociales(pp. 59,60).

### *2.7.3 Herramientas para la extracción de Datos"*

El análisis de este tipo de herramientas presentado por Herbert (2020) incluye las siguientes Content Grabber, La herramienta está construida con la función de admitir la extracción de contenido de cualquier sitio web y guardarlo en un formato estructurado. Esto consiste en informes CSV, XML y Excel; por su lado la empresa de IBM Cognos Analytics, soporta la visualización, diseñado con una interfaz basada en web para admitir la visualización de datos en el producto de BI. Tiene módulos para el gobierno de datos, el análisis de autoservicio y la gestión. Esta herramienta también admite la integración de datos de diferentes fuentes para crear informes y visualizaciones"; la herramienta de Sage Live, basada en la nube,

permite crear facturas para pagos usando dispositivos móviles; interesante la evolución de Apache Spark, la herramienta está diseñada, para el manejo de analítica a tiempo real (pp. 60, 61)

### 2.8 "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).

#### *2.8.1 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 anexo E.

#### *2.8.2 La posición de Linebacker (apoyador)*

El apoyador (linebacker) es el orquestador de la defensiva, debe evitar que el equipo opuesto avance yardas con el balón, cubrir el pase y enviar las señales a la defensiva, adicionalmente, debe ser un líder para el equipo; existen comúnmente tres posiciones diferentes y se les denomina por diferente nombre, el "Mike" suele ser el apoyador medio, el "Sam" es el apoyador del lado fuerte y el "Will" es el apoyador del lado débil. La S(Strong en inglés) en Sam ayuda a los jugadores a recordar fuerte, y la W(Weak en inglés) en Will ayuda a los jugadores a recordar débil (Carroll & Neft, 1999) véase también resultado de la investigación en anexo "C".

# CAPÍTULO 3

# DESARROLLO

Basado en la congruencia metodológica antes definida en este informe, se procede a 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 utilizando la minería de datos & machine learning, con la herramienta de R y considerando las premisas de la ciencia de los datos, se considera a Lantz (2019) el cual aplica el aprendizaje automático, para tareas del mundo real, con el proceso de **cinco pasos** (p.18); congruentemente, el proceso de la "Minería de datos", gestiona y colabora a través de las **seis fases Chapman** y otros (2000); así **que la síntesis de ambos**, son los pasos que guían este desarrollo, inicia con **la comprensión del negocio**, **la recopilación de datos y a través de Machine Learning y la herramienta de R, se lleva a cabo la exploración, la comprensión, la reparación** de datos, una vez listo estos pasos, se procede a ***construir los modelos estadísticos*** a utilizar, los cuales se componen de, el **KNN**, **la Regresión Lineal,** el modelo de **Árbol de Decisión** y el modelo **de Naive Bayes, se entrena** a cada uno de los modelos mencionados se ***ejecutan y se les evalúa usando machine learning.***

## 3.0 La comprensión del negocio

Explicado lo anterior, **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 (véase anexo D, 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 (ver explicación de las pruebas en anexo E.

### 3.1 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. Por cuestión de simplicidad, la preparación del ambiente R en

RStudio, se comparte en el anexo H y el trabajo realizado en referencia a la minería de datos menos relevante como la recopilación de los datos (mixtos y de obtención transversal) vía internet y ***la exploración, la comprensión, la preparación de datos y la transformación,*** que permite este estudio se puede encontrar y usar para su reproducción paso a paso en el anexo I; aquí solamente se muestran los resultados más relevantes.

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.

**Figura 10**

*Ray Lewis y tres super estrellas sin resultados del Combinado*

| 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 |

Fuente: Elaboración del autor con la herramienta de RStudio

Una variable atractiva 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 (véase anexo I en estadísticas) 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) 4.744, 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.

El análisis en la minería también revela evidencia de los jugadores elite, son más ligeros en kilogramos que el resto de los participantes, así también son más veloces. En los datos provistos por la NFL, existe una variable llamada Wonderlic la cual identifica a los que son considerados como las mejores selecciones en esta posición de apoyador, sin embargo, la métrica de Wonder sólo ha sido exitosa (acorde a este estudio) en 3 ocaciones de 24 jugadores ahora elite, es decir que la predicción realizada por esta variable es sólo 0.125 confiable.

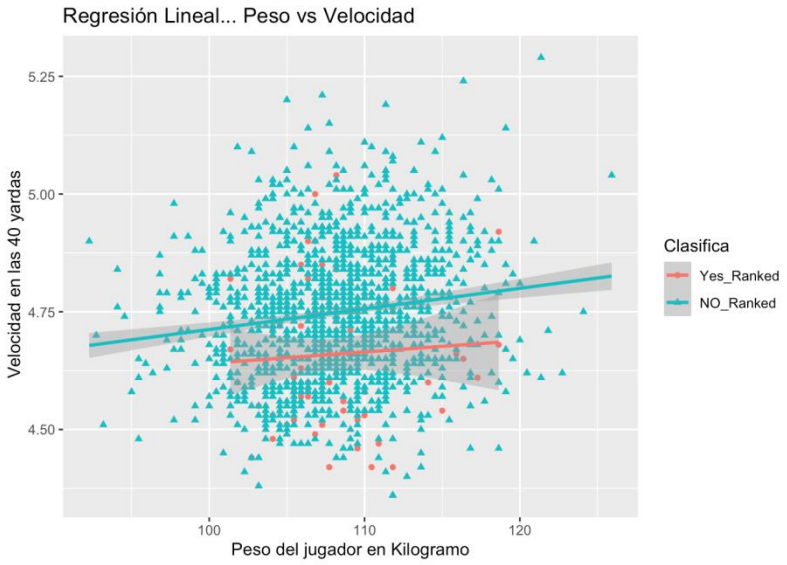
## 3.2 Machine Learning

### *3.2.1 Regresión Lineal*

Para comprender las relaciones entre una variable dependiente y una o varias variables independientes se ha aplicado el análisis con regresión lineal, en la figura11 se presenta uno de varias regresiones realizadas, se sugiere al lector si desea revisar el procedimiento completo y todas las comparaciones detallas ver el anexo J, se aprecia que dada la variable del peso, entre mayor peso de los jugadores, mayores los tiempos en las 40 yardas, es decir son más lentos, adicionalmente es evidente que los jugadores elite son más veloces y más ligeros, como lo muestra la línea roja(tenue).

**Figura 11**

*Regresión Lineal Comparativo, Peso del Jugador vs Velocidad*

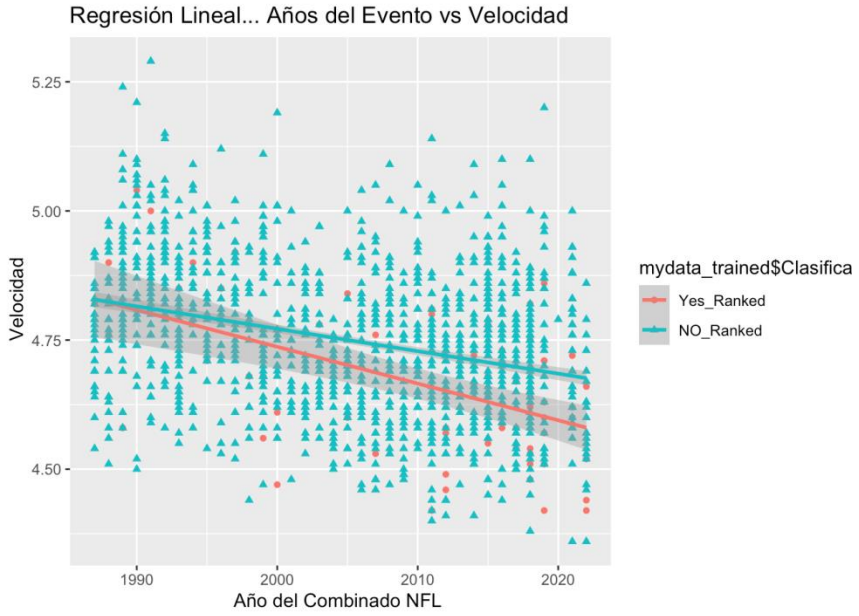



Fuente: recuperada de los trabajor realizados en RStudio

Dentro de las regresiones, dada la variable del año en que participan en el combinado, tanto la fuerza como la velocidad ha mejorado radicalmente, abajo en la figura 12, se muestra solamente la mejora en velocidad cada año y prevalece el comportamiento de los jugadores elite,, obteniendo mejores resultados .

**Figura 12**

*Regresión Lineal Comparativo, Año del Combinado vs Velocidad*

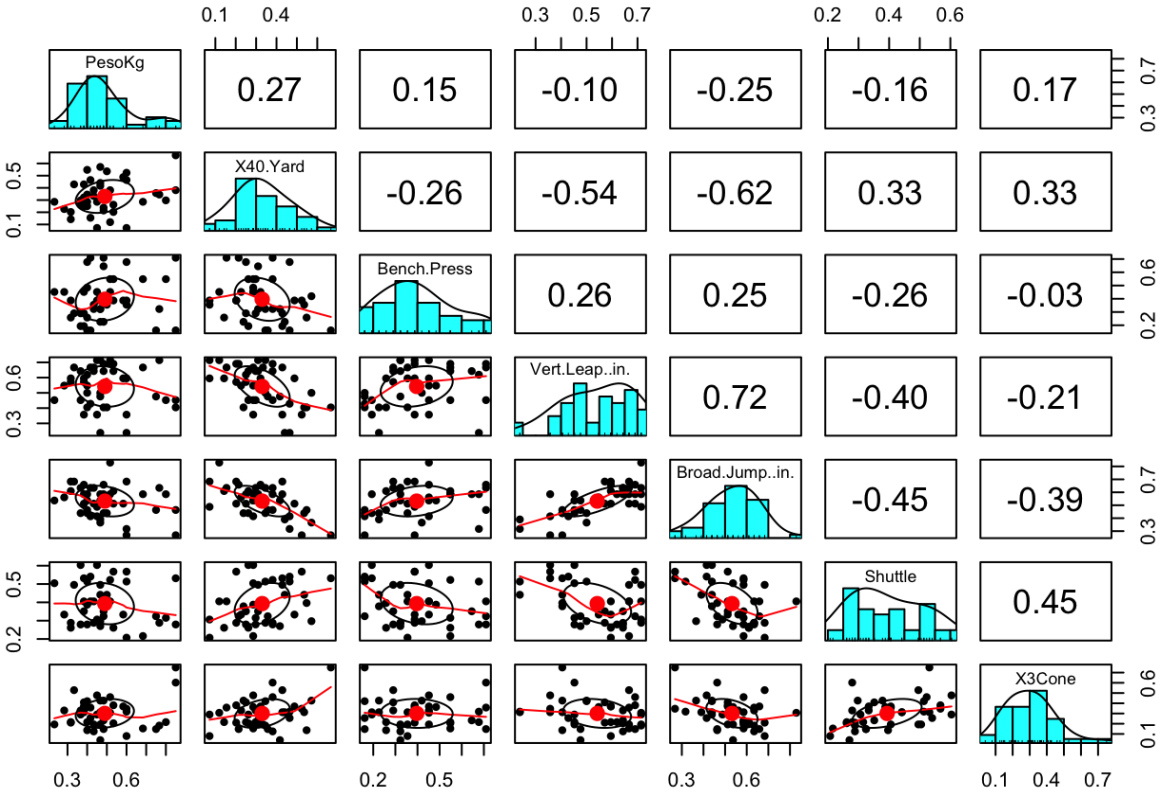



### *3.2.2 Pares de regresión lineal*

En la búsqueda para comprender las características y correlación entre las variable dependiente e independientes con los jugadores de la “elite”. Se realiza una gráficas en pares; mostrando en las gráficas de dispersión, los ovalos que reflejan la tendencia de la regresión, así diagonalmente un histograma para saber la distribución de los jugadores en cantidades y valores, por último el número refleja la correlación entre los pares comparados. La correlación de variables con mayor fuerza se encuentra en la prueba de salto vertical vs salto a lo largo, con un coeficiente de correlación de 0.72; así también en las pruebas de los 3 conos (X·Cone) vs las pruebas de ida y regreso (Shuttle) con .45 en otras palabras los jugadores elite destacan en este par de pruebas. Vease figura 13 a continuación.

**Figura 13**

*Mapa de Pares de Regresión Lineal, Histogramas y Factor de Correlación*




Fuente de elaboración con herramienta RStudio y datos del estudio

### *3.3.3 K Nearest Neighbors*

Para generar un modelo que permita decidir seleccionar entre jugadores nuevos participantes en futuras pruebas de combinados, se ha decidido aplicar la técnica de aprendizaja "supervisado", considerada como "lazy", es la técnica de clasificación K-Nearest Neighbors que permite agrupar a elementos similares, como lo pueden ser los jugadores elite con sus prospectos más cercanos, para tal efecto se han preparado los datos, transformado y dado las discrepancia en los datos númericos se decide transformarlos a normalizados ( si se desea revisar paso a paso el procedimiento véase el anexo K) ; se crea la tabla de entrada llamada (mydata_trained), todos los campos deben estar completos, no faltantes o (NA), se segmenta en mydata_trained para (entrenamiento y pruebas)

Se logran obtener 42 registros entrenados con los jugadores elite que deseamos buscar en futuros combinados, como lo muestra el resultado desplegado a continuación.

```
##
## 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,]
```

#### 3.3.3.1 Ejecución del modelo k-nn, lazy y supervisad

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
```

#### 3.3.3.2 Comprobación del modelo k-nn, lazy y supervisado

Comparando el modelo que ha sido entrenado vs test: El modelo ha encontrado a los 12 mejores jugadores correctamente (Yes_Ranked son embargo expresa que sólo esta 40% seguro de la redicción) también identifica el resto como NO rankeados, con el 60% de confianza Cabe mencionar que el modelo acertó al 100%..

Nota, el modelo de entrenamiento le entregó las muestras de los jugadores elite y está listo para ser probado con datos que surjan de próximas pruebas de combinados.

```
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 |
## ---------------------|------------|------------|
##
```

### *3.3.4 Árbol de Decisión*

Por último se construye el modelo estadístico usando el "Árbol de Decisión" con el cual se desea saber la probabilidad que un futuro apoyador sea LB Interno ILB o externo OLB, basado en los resultados de sus pruebas. Para entrenar este modelo ahora se han considerado a todos los jugadores apoyadores en el "Combinado Scout NFL" desde 1982 hasta 2022, y estudiado con el uso de las siguientes variables: X40.Yard + Vert.Leap..in.+ Bench.Press + X3Cone + PesoKg + Broad.Jump..in.

#### 3.3.4.1 Creación del modelo

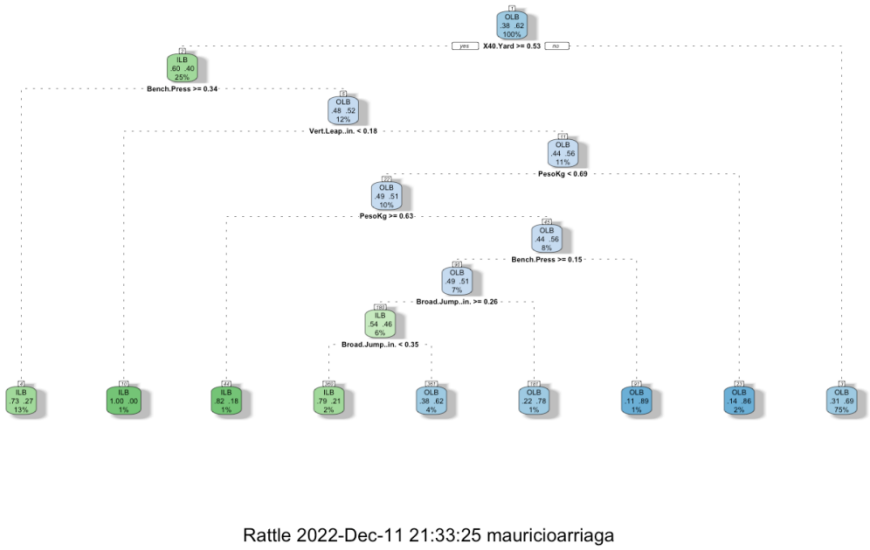
```
arbol <- rpart(
  formula = POS ~ X40.Yard + Vert.Leap..in.+ Bench.Press + X3Cone + PesoKg +
    Broad.Jump..in. ,
  data = mydata_trained_n,
  method = "class")
```

3.3.4.2 Graficando el árbol con la probabilidad de ser OLB o ILB

```
fancyRpartPlot(arbol)
```

**Figura 14**

*Árbol de Decisión Para Estimar La Probabilidad de Ser ILB u OLB*




Fuente de elaboración del autor, con la herramienta de RStudio.

Se observa que el 62% de los participantes más rápidos, en la prueba de velocidad (40 yardas) son apoyadores externos (OLB), lo cual coincide en el desempeño esperado al cubrir pases y evitar las carreras por fuera de los tackles, en las llamadas “sweeps”, “rápidas” y pases al “hook” y al “flat”, tambien se observa en el siguiente nivel de la izquierda, que el apoyador ILB en un 60% de las veces es más fuerte, lo cual le ayudará a contra-bloquear a los linieros ofensivos, así como podrá romper bloqueos del fullback, bloqueos dobles y de trayectoria de trampa. Es remarcable también que de aquellos jugadores en el tercer nivel, que siendo 38% más lentos que el OLB, un 40% de ellos, no son muy fuertes, dado ésta característica su tendencia es ser OLB, si desea ver el algoritmo y modelo paso a paso, vea el anexo L.

### *3.3.5 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, el proceso a detalle lo puede ver en el anexo M , así también 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)
m
```

#### 3.3.5.1 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".

## Bibliografía


Ahmed, M., Davis, V., Gamliel, S., Irizarry, R., Mastrodoménico, 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 edición (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 Institute of Media Innovation. (S. T. Analytycs, 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.). Gráficos 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 Cience 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 Neightboard. K-Nearest Neightboard. 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)

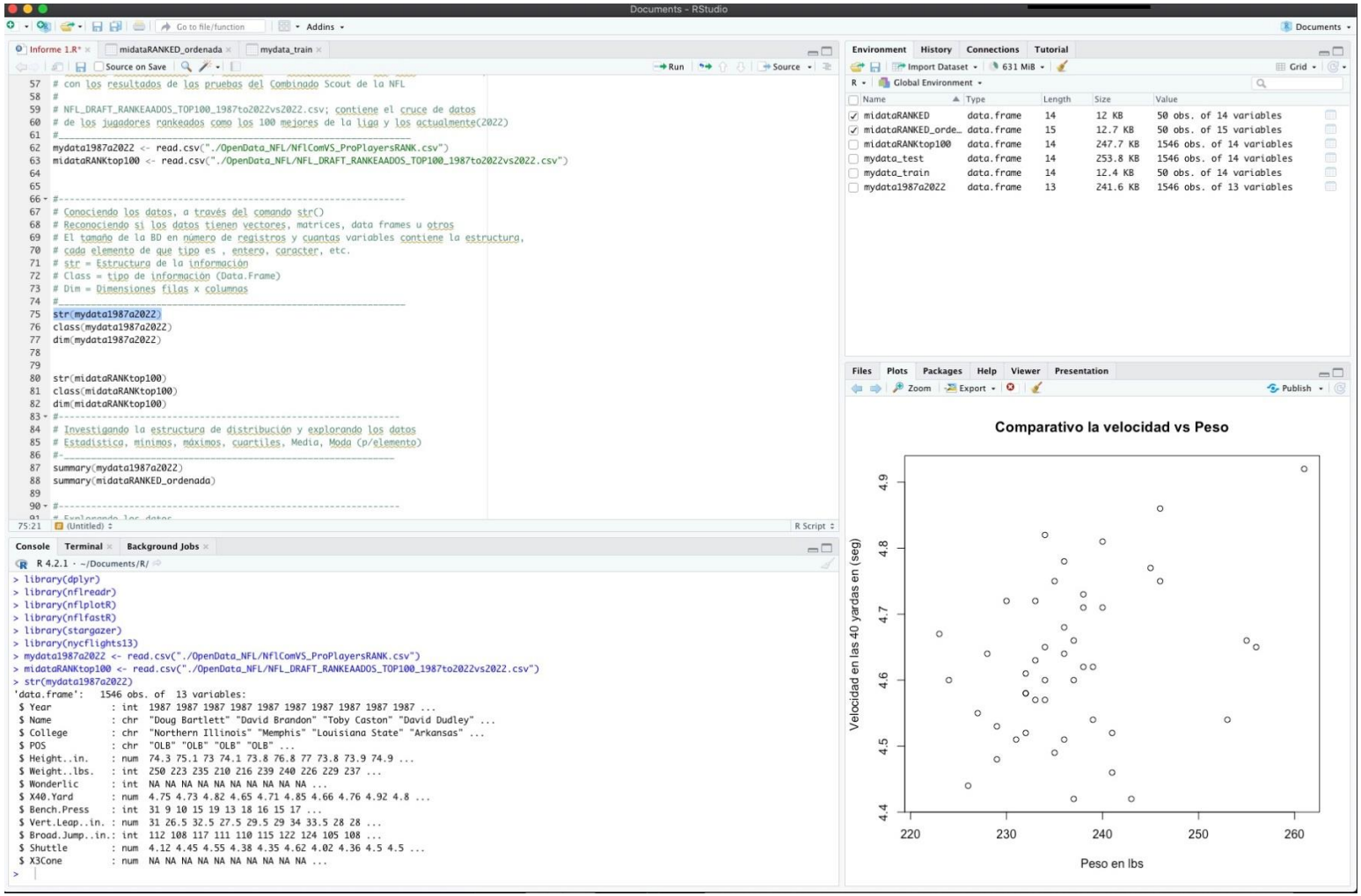
Zimmer, B. (2012). The World. Recuperado el noviembre de 2022, de How Sam, Mike and Will Became Football Positions: https://www.bostonglobe.com/ideas/2012/09/08/how-sam-mike-and-will-became-football-positions/URHq2XoAdviKJYZLQfbKNK/story.html

## ANEXO A

**El ambiente de RStudio, con el ambiente de trabajo de este informe**

**Figura 155**

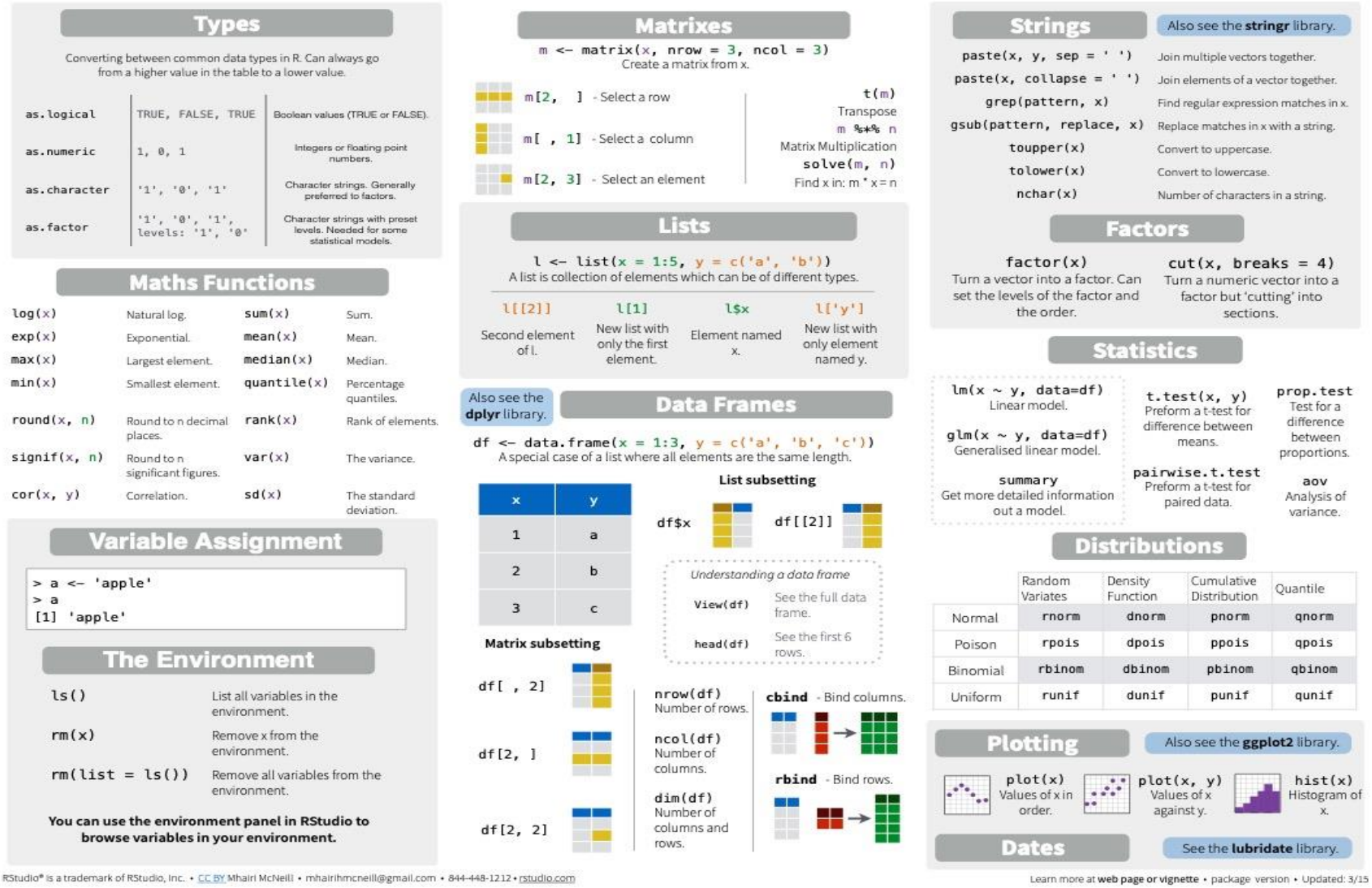
*Ambiente Gráfico de RStudio*




Fuente: Imagen de Rstudio, con datos de elaboración del autor

**Figura 16**

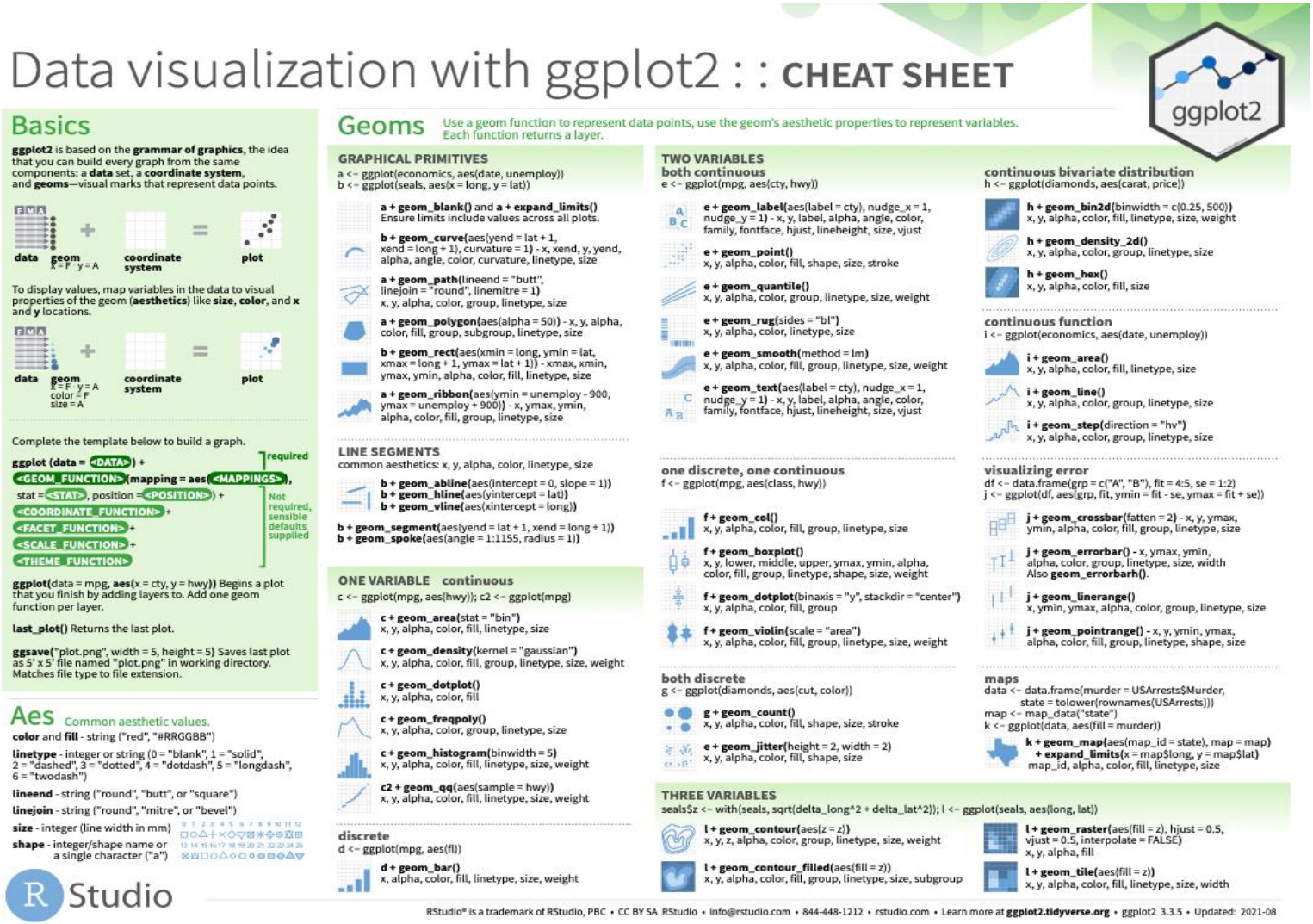
*Comandos en R (continúa)*




Fuente: Recuperada de (Sheet, 2021)

**Figura 17**

*Data Visualización ggplot2*




**Fuente:** (Shett, 2021) (https://github.com/rstudio/cheatsheets/blob/main/data-visualization-2.1.pdf)

## ANEXO B

### Herramientas para el manejo de la Ciencia de Los Datos.

**Tabla 4**

*Herramientas Abiertas(sin costos) para la Ciencia de los Datos*

| Nombre | Aplicación |
|---|---|
| OpenRefine | Es una gran herramienta para trabajar con datos que están desorganizados; esta herramienta permite que un científico de datos limpie y cambie el formato de los datos |
| Orange | Una herramienta de visualización y análisis de datos de código abierto diseñada. Tiene un flujo de trabajo interactivo simple y una caja de herramientas avanzada para ayudar a una persona a crear un flujo de trabajo interactivo que pueda analizar y visualizar datos. |
| Knime | Es otra solución de código abierto para usar en el análisis de datos. Esta herramienta permite a una persona explorar y descubrir información oculta en los datos. La herramienta tiene más de 1000 módulos y cientos de ejemplos que se pueden ejecutar para aprender a usarla. Además, existe una gama avanzada de herramientas integradas y algoritmos complejos. |
| R & RStudio | Se considera el estándar entre los lenguajes de programación estadísticos, es un software de código abierto que cualquiera puede instalar y usar en computación estadística y gráficos; compatible con plataformas Windows, MacOS y UNIX; dado que R es gratuito, cualquiera puede instalarlo, usarlo, actualizarlo, modificarlo, clonarlo y revenderlo; es también un lenguaje de alto rendimiento que ayudará a los usuarios a manejar un extenso paquete de datos y crear una gran herramienta para ayudar a administrar Big Data. |
| RapidMiner | "Al igual que KNIME, RapidMiner se ocupa de la programación visual y es el mejor cuando se trata de modelado, análisis y manipulación de datos. RapidMiner mejora la productividad de los equipos de Data Science. Tiene una plataforma de código abierto para admitir Machine Learning, implementación de modelos y preparación de datos. " |
| Pentaho | Pentaho se ocupa de los problemas que afectan la capacidad de una organización para aceptar el valor de diferentes datos. La plataforma simplificará la preparación y combinación de datos, así como una colección de herramientas utilizadas en el análisis, visualización, generación de informes, exploración y predicción. Pentaho está diseñado para garantizar que cada miembro de un equipo pueda transformar los datos en valor |

| Nombre | Aplicación |
|---|---|
| OpenRefine | Es una gran herramienta para trabajar con datos que están desorganizados; esta herramienta permite que un científico de datos limpie y cambie el formato de los datos |
| Orange | Una herramienta de visualización y análisis de datos de código abierto diseñada. Tiene un flujo de trabajo interactivo simple y una caja de herramientas avanzada para ayudar a una persona a crear un flujo de trabajo interactivo que pueda analizar y visualizar datos. |
| Knime | Es otra solución de código abierto para usar en el análisis de datos. Esta herramienta permite a una persona explorar y descubrir información oculta en los datos. La herramienta tiene más de 1000 módulos y cientos de ejemplos que se pueden ejecutar para aprender a usarla. Además, existe una gama avanzada de herramientas integradas y algoritmos complejos. |

**Fuente:** Elaboración del autor, con datos de (Herbert, 2020, págs. 56-63)

**Tabla 5**

*Herramientas Abiertas(Sin Costos) para la Ciencia de los Datos (Continuación)*

| Nombre | Aplicación |
|---|---|
| Weka | Otro software de código abierto diseñado con la capacidad de manejar algoritmos de aprendizaje automático para usar en tareas de minería de datos. Puede utilizar directamente el algoritmo para procesar un conjunto de datos. También es el mejor para usar en el desarrollo de un nuevo esquema de aprendizaje automático porque se implementa completamente en la programación JAVA.<br>Dado que la interfaz gráfica de usuario de Weka es simple y fácil de usar, facilita una transición fácil al campo de la ciencia de datos. |
| NodeXL | Una herramienta de análisis y visualización de datos que muestra las relaciones en un conjunto de datos; dado que es un software de código abierto, es de uso gratuito para analizar y crear visualizaciones a partir de datos; cuenta con diferentes módulos como importadores de datos de redes sociales y automatización. |
| Gelphi | Escrito en Java, visualizador y es una herramienta de análisis de red |
| Talend | "Es el proveedor líder de software de integración de código abierto para la mayoría de las empresas basadas en datos. Talend permite |

que los clientes de cualquier lugar se conecten fácilmente.

**Fuente:** Elaboración del autor, con datos de (Herbert, 2020, págs. 56-63)

## ANEXO C

### Tipos de Linebackers o Apoyadores

***Linebacker Central***

“El apoyador medio o interior (MLB o ILB), a veces llamado "Mike" o "Mac", (Zimmer, 2012) a menudo se le conoce como el "mariscal de campo de la defensa” (Today, 2011) los entrenadores le instruyen con la formación y actividades que deben realizar, él debe comunicar a sus compañeros en el campo; también debe ser un experto en todas las posiciones defensivas, se le exige que haga penetraciones sobre el QB, cubra a jugadores personalmente y/o zonas de pase establecidas en patrones que debe estudiar cuidadosamente y por supuesto mantenga impenetrable el medio del campo en sus huecos de responsabilidad, asistiendo a sus compañeros en todo momento; en base a su responsabilidad y ubicación al inicio de la jugada se le puede denominar "Mike", mientras que el jugador más rápido, más orientado a la protección de pases y cobertura de rutas se llama "Will" o “Wanda”; "Mikes" por lo general se alinea hacia el lado fuerte (es decir el lado con más jugadores ofensivos) o en el lado en el que es más probable que se desarrolle la ofensiva, mientras que "Wills" “ (Carroll & Neft, 1999).

***Linebacker fuerte ( se coloca al lado de más jugadores ofensivos)***

“El apoyador del lado fuerte (SLB) o apoyador interno (ILB internal Linebacker) a menudo recibe el apodo de "Sam"; dado que el lado fuerte del equipo ofensivo es el lado en el que se alinea el ala cerrada, o el lado que contenga la mayor cantidad de jugadores; comúnmente es el apoyador de mayor fuerza física; al menos posee la capacidad de resistir, contra bloquear a la línea ofensiva o un fullback (corredor de poder) que defiende al pasador o abre espacios a los corredores; el apoyador también debe tener rapidez y agilidad en situaciones de pase y requiera cubrir hombre a hombre a corredores o ala cerradas; su agilidad mental para hacer lecturas y reconocer patrones de jugadas”  (Linebackers, 2022)

***Linbacker rápido( juega del lado débil , es decir donde hay menos jugadores ofensivos)***

“El apoyador del lado débil (WLB) también llamado "Will" o OLB (outside Linebacker en inglés) en la defensa 4-3, a veces llamado apoyador trasero, o "Buck", así como otros nombres

como Jack o Bandit, (Today, 2011), debe ser el más rápido de los tres. porque a menudo es el llamado a la cobertura de pase. Por lo general, también persigue la jugada desde atrás, por lo que la capacidad de maniobrar a través del tráfico es una necesidad para Will. Will generalmente se alinea fuera de la línea de golpeo a la misma profundidad que Mike. Está posición tiene gran ventaja, sobre los otros linebackers, ya que se coloca atrás de sus compañeros, protegido normalmente por el tackle defensivo y el ala defensiva (existen varios escenarios) lo que le da velocidad para asistir en la carrera o en el pase.” (Linebackers, 2022) https://en.wikipedia.org/wiki/Linebacker

## ANEXO D

### Lista de los mejores jugadores, considerados por NFL y PFF

La lista de los mejores jugadores, se obtiene de la lista de los 100 mejores de todos los tiempos de la NFL, pero que estén en el rango de tiempo del estudio 1987 al 2022; adicionalmente son incluidos en la lista aquiellos paoyadores que están activos en la NFL y son lo suficiente mente buenos, para estar considerados en el rank de PFF, vease siguiente figura 20.

**Figura 188**

*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 | Washingtor | 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 | Washingtor | 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 PFF

## ANEXO E

### Descripción de las pruebas en el Combinado Scout NFL

#### *La carrera de 40 yardas*

La carrera de 40 yardas es inequívocamente el evento más popular del Combinado. La carrera de 40 yardas está destinada a medir la velocidad lineal, la aceleración, la explosión desde un inicio estático e incluso la forma de correr o la forma de andar de un atleta. Los atletas son cronometrados a intervalos de 10, 20 y 40 yardas (Casan, 2022).

#### *Levantamiento de pesas, con el pectoral (acostado en un banco) (Bench press)*

El bench press, consiste en repeticiones máximas con 225 libras (recostado en una banca) este ejercicio mide tanto la fuerza de la parte superior del cuerpo, como la condición fñisica de recuperación, es un idicador de la preparación que ha deicado el jugador en la universidad (Casan, 2022)

#### *El Salto vertical*

El salto vertical es un escaparate ideal para la explosión de la parte inferior del cuerpo, la potencia y la producción de fuerza, durante este ejercicio, un atleta se parará con los pies planos, medirá su alcance al saltar y realizará el salto sin dar ningún paso. (Ben-Ishay, 2020)

#### *Salto de longitud*

Al igual que el salto vertical, el salto de longitud también pondrá a prueba la potencia y la explosión de la parte inferior del cuerpo de un atleta. El salto de longitud también mide la fuerza en la dirección horizontal, el equilibrio y la coordinación a través del salto y el aterrizaje. El atleta comenzará desde una posición de pie y luego explotará lo más que pueda en dirección horizontal.

***Ejercicio de explosividad a 3 conos***

El ejercicio de los tres conos, pone a prueba la capacidad de un atleta para cambiar de dirección mientras corre a alta velocidad y realiza giros de 90°, los tres conos están orientados en forma de L; el prospecto comenzará desde la línea de salida, correrá 5 yardas hasta el primer cono y regresará a la línea de salida, dará la vuelta y correrá de regreso al primer cono, dará un giro de 90°, correrá 5 yardas hasta el segundo cono que gira alrededor , y corre de regreso a la línea de partida mientras gira alrededor del cono central.

***Carrera de ida y regreso***

La carrera de ida y vuelta (o ida y vuelta corta) es una carrera de ida y vuelta clásica en la que los atletas comienzan el ejercicio, corren 5 yardas en una dirección, invierten la dirección, corren 10 yardas en la dirección opuesta, invierten la dirección nuevamente y terminan el ejercicio corriendo 5 más yardas hasta la línea de meta; éste tipo de prueba pone a prueba la rapidez lateral del atleta, la aceleración, la capacidad de cambio de dirección, la capacidad de detenerse, el equilibrio, la agilidad y la explosión en áreas cortas.

## ANEXO F

***El evento del combinado scout 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); sobre el evento detalla el Dr. Casan (2022) que los jugadores participantes, son sometidos a pruebas tanto mentales como físicas por un grupo de caza talentos (en inglés "scouts") así como por los gerentes generales y los entrenadores de la NFL; los atletas participan sólo por invitación y en dicho evento se recolectan datos relacionados con las características físicas, sus exámenes médicos, las entrevistas que se les realizan, las medidas de estatura, la velocidad de los jugadores en las 40 yardas, el salto de altura, el salto de longitud, la fuerza, la agilidad de correr ida y regreso, entre otras pruebas de tipo físicas y mentales (p. 1) detalla el Dr. Casan (2022) que los jugadores participantes, son sometidos a pruebas tanto mentales como físicas por un grupo de caza talentos (en inglés "scouts") así como por los gerentes generales y los entrenadores de la NFL; los atletas participan sólo por invitación y en dicho evento se recolectan datos relacionados con las características físicas, sus exámenes médicos, las entrevistas que se les realizan, las medidas de estatura, la velocidad de los jugadores en las 40 yardas, el salto de altura, el salto de longitud, la fuerza, la agilidad de correr ida y regreso, entre otras pruebas de tipo físicas y mentales (p. 1).

## ANEXO G

### Catálogo de Datos de PFF

| | | |
|---|---|---|
| **#G**: Number of games in which the player appeared | **#**: Jersey Number | **POS**: Season position |
| **PRSH**: Pass Rush Snaps | **TOT**: Total Snaps | **RDEF**: Snaps in a run defense role |
| **RDEF**: PFF Grade for Run Defense | **COV**: Coverage Snaps | **DEF**: PFF Grade for Defense |
| **COV**: PFF Grade for Defensive Coverage against Receivers | **TACK**: PFF Grade for Tackling | **PRSH**: PFF Grade for Pass Rush |
| **HIT**: Hits - when the passer is hit by the defender | **TOT**: Total pressures of the passer of any kind (generated by the defense) | **SK**: Sacks |
| **TKL**: Tackles | **HUR**: Hurries - when the passer is hurried by the defender | **BAT**: Batted Passes - the deflected at the line of scrimmage |
| **MIS%**: Missed Tackle Rate | **AST**: Assisted Tackles | **MIS**: Missed Tackles |
| **TGT**: Receiving Targets | **STOP**: Defensive Stops - tackles that constitute a "failure" for the offense | **FFM**: Forced Fumbles |
| **YDS**: Receiving Yards | **REC**: Receptions | **REC%**: Percentage of targets caught |
| **LNG**: Longest | **Y/REC**: Yards per Reception | **YAC**: Yards After Catch |
| **PBU**: Pass Breakups | **TD**: Receiving TD | **INT**: Receiving Interceptions |
| **DL**: Number of player's snaps lined up on the Line | **NFL**: NFL Passer Rating Against | **PEN**: Total (Declined+Offset): Total and (declined or offsetting) penalties |
| **Slot**: Number of player's snaps lined up at Slot Corner | **Box**: Number of player's snaps lined up in the Box | **FS**: Number of player's snaps lined up at Free Safety |
| **BGP**: Number of player's snaps lined up on DL as a DT | **Cnr**: Number of player's snaps lined up at Corner | **AGP**: Number of player's snaps lined up on DL as a NT |
| | **OVT**: Number of player's snaps lined up on DL over an OT | **OUT**: Number of player's snaps lined up on DL outside the OT |

Fuente: (PFF, 2022)

## ANEXO H

### Preparación del ambiente de R en RStudio

#### Algoritmo de instalación con los paquetes requeridos

Dentro de esta documentación se puede identificar todas las actividades referente a la ciencia de los datos, se comienza con la minería de datos a través de la herramienta de machine learning R, la cual requiere llevar a cabo, la carga de paquetería especializada que le permite la manipulación, la revisión, la agregación, la eliminación y la transformación de los datos, hasta convertirlos en modelos estadísticos aplicables.

Los comandos (expresados en cajas de color gris) y resultados obtenidos en la consola, se pueden identificar con un renglón inicial marcado como ##.

A continuación se documenta la instalación de la paquetería.

```
options(repos = list(CRAN="http://cran.rstudio.com/"))
install.packages(c("nycflights13", "gapminder", "Lahman"))
```

```
##
## The downloaded binary packages are in
##  /var/folders/hr/yfn8jh751kdg8lzz2l72q0740000gn/T//RtmpaeRmAk/downloaded_p
ackages
```

```
install.packages("tidyverse")
```

```
##
## The downloaded binary packages are in
##  /var/folders/hr/yfn8jh751kdg8lzz2l72q0740000gn/T//RtmpaeRmAk/downloaded_p
ackages
```

```
install.packages("ggrepel", type = "binary")
```

```
##
## The downloaded binary packages are in
##  /var/folders/hr/yfn8jh751kdg8lzz2l72q0740000gn/T//RtmpaeRmAk/downloaded_p
ackages
```

```
install.packages("ggplot2")
```

```
##
## The downloaded binary packages are in
##  /var/folders/hr/yfn8jh751kdg8lzz2l72q0740000gn/T//RtmpaeRmAk/downloaded_p
ackages
```

```
install.packages("plotly")
```

```
##
## The downloaded binary packages are in
##  /var/folders/hr/yfn8jh751kdg8lzz2l72q0740000gn/T//RtmpaeRmAk/downloaded_p
ackages
```

```
install.packages("devtools")
```

```
##
## The downloaded binary packages are in
##  /var/folders/hr/yfn8jh751kdg8lzz2l72q0740000gn/T//RtmpaeRmAk/downloaded_p
ackages
```

```
install.packages("stargazer")
```

```
##
## The downloaded binary packages are in
##  /var/folders/hr/yfn8jh751kdg8lzz2l72q0740000gn/T//RtmpaeRmAk/downloaded_p
ackages
```

```
install.packages('class')
```

```
##
## The downloaded binary packages are in
##  /var/folders/hr/yfn8jh751kdg8lzz2l72q0740000gn/T//RtmpaeRmAk/downloaded_p
ackages
```

```
install.packages("fastDummies")
```

```
##
## The downloaded binary packages are in
##  /var/folders/hr/yfn8jh751kdg8lzz2l72q0740000gn/T//RtmpaeRmAk/downloaded_p
ackages
```

```
install.packages("RColorBrewer")
```

```
##
## The downloaded binary packages are in
##  /var/folders/hr/yfn8jh751kdg8lzz2l72q0740000gn/T//RtmpaeRmAk/downloaded_p
ackages
```

```
install.packages("gapminder")
```

```
##
## The downloaded binary packages are in
##  /var/folders/hr/yfn8jh751kdg8lzz2l72q0740000gn/T//RtmpaeRmAk/downloaded_p
ackages
```

```
install.packages("gmodels")
```

```
##
## The downloaded binary packages are in
```

```
##   /var/folders/hr/yfn8jh751kdg8lzz2l72q0740000gn/T//RtmpaeRmAk/downloaded_p
ackages
```

```
install.packages("psych")
```

```
##
## The downloaded binary packages are in
##   /var/folders/hr/yfn8jh751kdg8lzz2l72q0740000gn/T//RtmpaeRmAk/downloaded_p
ackages
```

```
install.packages("e1071")
```

```
##
## The downloaded binary packages are in
##   /var/folders/hr/yfn8jh751kdg8lzz2l72q0740000gn/T//RtmpaeRmAk/downloaded_p
ackages
```

```
install.packages("rpart")
```

```
##
## The downloaded binary packages are in
##   /var/folders/hr/yfn8jh751kdg8lzz2l72q0740000gn/T//RtmpaeRmAk/downloaded_p
ackages
```

```
install.packages("rattle")
```

```
##
## The downloaded binary packages are in
##   /var/folders/hr/yfn8jh751kdg8lzz2l72q0740000gn/T//RtmpaeRmAk/downloaded_p
ackages
```

```
install.packages("rpart.plot")
```

```
##
## The downloaded binary packages are in
##   /var/folders/hr/yfn8jh751kdg8lzz2l72q0740000gn/T//RtmpaeRmAk/downloaded_p
ackages
```

**Cargar las librerías**

La instalación de todo los paquetes en R, requiere que se le invoque para poder utilizar las bondades que ofrece a través de sus comandos, un ejemplo es paquete ”dplyr" provee herramientas de apoyo como el pipe( %<%), así como más de 6 de las funciones, algunas de ellas son : select() ; filter() ; group_by() ; summarize() ; mutate(), otro ejemplo es el de “stargazer” elaborado por Hlavac, Marek (2022). stargazer: Well-Formatted Regression and

Summary Statistics Tables. R package version 5.2.3. https://CRAN.R-project.org/package=stargazer.

Se procede a la carga de las librerias y se despliegan los resultados y comentarios.

```
library(shiny)
library(rmarkdown)
devtools::session_info()
```

```
## ─ Session info ─────────────────────────────────────────────────────────
─────
##  setting  value
##  version  R version 4.2.1 (2022-06-23)
##  os       macOS Catalina 10.15.7
##  system   x86_64, darwin17.0
##  ui       X11
##  language (EN)
##  collate  en_US.UTF-8
##  ctype    en_US.UTF-8
##  tz       America/Monterrey
##  date     2022-12-11
##  pandoc   2.19.2 @ /Applications/RStudio.app/Contents/MacOS/quarto/bin/too
ls/ (via rmarkdown)
##
## ─ Packages ─────────────────────────────────────────────────────────────
─────
##  package     * version date (UTC) lib source
##  cachem        1.0.6   2021-08-19 [1] CRAN (R 4.2.0)
##  callr         3.7.2   2022-08-22 [1] CRAN (R 4.2.0)
##  cli           3.4.1   2022-09-23 [1] CRAN (R 4.2.0)
##  crayon        1.5.2   2022-09-29 [1] CRAN (R 4.2.0)
##  devtools      2.4.5   2022-10-11 [1] CRAN (R 4.2.0)
##  digest        0.6.29  2021-12-01 [1] CRAN (R 4.2.0)
##  ellipsis      0.3.2   2021-04-29 [1] CRAN (R 4.2.0)
##  evaluate      0.16    2022-08-09 [1] CRAN (R 4.2.0)
##  fastmap       1.1.0   2021-01-25 [1] CRAN (R 4.2.0)
##  fs            1.5.2   2021-12-08 [1] CRAN (R 4.2.0)
##  glue          1.6.2   2022-02-24 [1] CRAN (R 4.2.0)
##  highr         0.9     2021-04-16 [1] CRAN (R 4.2.0)
##  htmltools     0.5.3   2022-07-18 [1] CRAN (R 4.2.0)
##  htmlwidgets   1.5.4   2021-09-08 [1] CRAN (R 4.2.0)
##  httpuv        1.6.6   2022-09-08 [1] CRAN (R 4.2.0)
##  knitr         1.40    2022-08-24 [1] CRAN (R 4.2.0)
##  later         1.3.0   2021-08-18 [1] CRAN (R 4.2.0)
##  lifecycle     1.0.3   2022-10-07 [1] CRAN (R 4.2.0)
##  magrittr      2.0.3   2022-03-30 [1] CRAN (R 4.2.0)
##  memoise       2.0.1   2021-11-26 [1] CRAN (R 4.2.0)
##  mime          0.12    2021-09-28 [1] CRAN (R 4.2.0)
##  miniUI        0.1.1.1 2018-05-18 [1] CRAN (R 4.2.0)
##  pkgbuild      1.3.1   2021-12-20 [1] CRAN (R 4.2.0)
```

```
##  pkgload       1.3.0   2022-06-27 [1] CRAN (R 4.2.0)
##  prettyunits   1.1.1   2020-01-24 [1] CRAN (R 4.2.0)
##  processx      3.7.0   2022-07-07 [1] CRAN (R 4.2.0)
##  profvis       0.3.7   2020-11-02 [1] CRAN (R 4.2.0)
##  promises      1.2.0.1 2021-02-11 [1] CRAN (R 4.2.0)
##  ps            1.7.1   2022-06-18 [1] CRAN (R 4.2.0)
##  purrr         0.3.4   2020-04-17 [1] CRAN (R 4.2.0)
##  R6            2.5.1   2021-08-19 [1] CRAN (R 4.2.0)
##  Rcpp          1.0.9   2022-07-08 [1] CRAN (R 4.2.0)
##  remotes       2.4.2   2021-11-30 [1] CRAN (R 4.2.0)
##  rlang         1.0.6   2022-09-24 [1] CRAN (R 4.2.0)
##  rmarkdown   * 2.18    2022-11-09 [1] CRAN (R 4.2.0)
##  rstudioapi    0.14    2022-08-22 [1] CRAN (R 4.2.0)
##  sessioninfo   1.2.2   2021-12-06 [1] CRAN (R 4.2.0)
##  shiny       * 1.7.2   2022-07-19 [1] CRAN (R 4.2.0)
##  stringi       1.7.8   2022-07-11 [1] CRAN (R 4.2.0)
##  stringr       1.4.1   2022-08-20 [1] CRAN (R 4.2.0)
##  urlchecker    1.0.1   2021-11-30 [1] CRAN (R 4.2.0)
##  usethis       2.1.6   2022-05-25 [1] CRAN (R 4.2.0)
##  xfun          0.33    2022-09-12 [1] CRAN (R 4.2.0)
##  xtable        1.8-4   2019-04-21 [1] CRAN (R 4.2.0)
##  yaml          2.3.5   2022-02-21 [1] CRAN (R 4.2.0)
##
##  [1] /Library/Frameworks/R.framework/Versions/4.2/Resources/library
##
## ──────────────────────────────────────────────────────────────────────
────
```

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────

tidyverse 1.3.2 ──
## ✓ ggplot2 3.4.0      ✓ purrr   0.3.4
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.2.1      ✓ stringr 1.4.1
## ✓ readr   2.1.3      ✓ forcats 0.5.2
## ── Conflicts ────────────────────────────────────── tidyverse_conflict
s() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(ggrepel)
library(ggplot2)
library(stargazer)
library(nycflights13)
library(plotly)
library(class)
```

```
library(fastDummies)
library(RColorBrewer)
library(gapminder)
library(crosstable)library(gmodels)
library(psych)
library(e1071)
library(rpart)
library(rattle)
library(rpart.plot)
library(gmodels)
```

**Configuración del ambiente de trabajo**

En RStudio se especifica el subdirectorio de trabajo, a través del comando “setwd”; así también, con el comando “options()” se define el uso de la notación científica.

```
setwd("~/Documents/R")
options(scipen = 9999)
```

# ANEXO I

## Minería de Datos usando Machine Learning en R y RStudio

### Carga de los datos crudos

***La recopilación de datos***, es tomada de la "NFL" (NFL, Stats, 2022), "Fantasy Football" (NFL, NFL Fantasy, 2022), NFL Draft (NFL, Stats, 2022); Bleacherreport (Wedell, 2010); ProFootballFocus (PFF, 2022), con el comando "read.csv" se suben a memoria los archivos crudos, NflCombinado1987a2022.csv, contiene los resultados del combinado NFL, entre los años 1987 a 2020. El archivo "NFL_DRAFT_conRANK(all).csv", es igual al archivo previo, sin embargo se le han agregado los datos con información, que permite identificar los mejores jugadores considerados la elite, siendo esta distinción otorgada por la NFL en las lista de los 100 mejores jugadores, y también por contar con la calificación embestida por la PFF (Pro Football Focus); es facilmente identificable con un número mayor a 1 dentro de la variable RANK.

```
mydata1987a2022 <- read.csv("./OpenData_NFL/NflCombinado1987a2022.csv")
midataRANKtop100 <- read.csv("./OpenData_NFL/NFL_DRAFT_conRANK(all).csv")
```

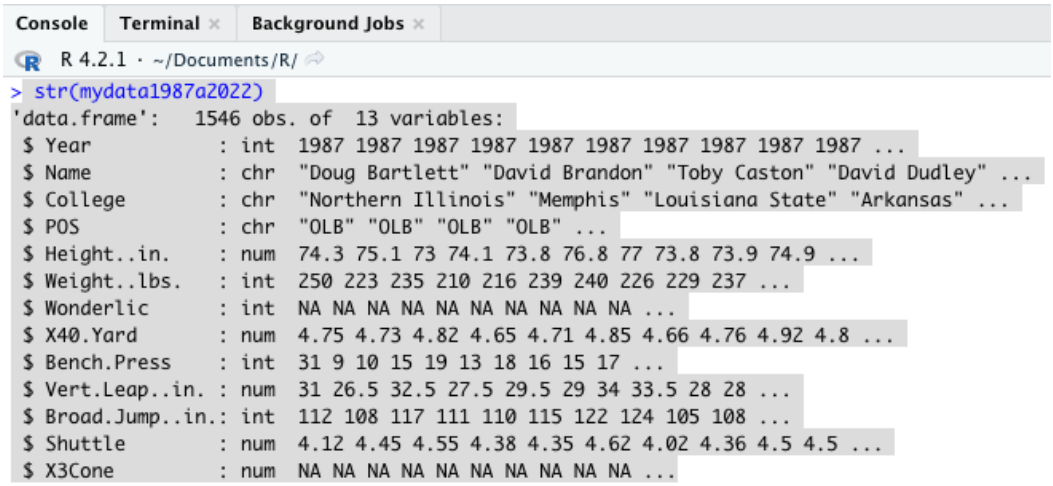
## mydata1987a2022 <- read.csv("./OpenData_NFL/NflCombinado1987a2022.csv")

##midataRANKtop100 <- read.csv("./OpenData_NFL/NFL_DRAFT_conRANK(all).csv")

Se explora la estructura de la información contenida; para simplificar la manipulación y análisis, se le asigna a un objeto que es identificado como, mydata1987a2022; del cual, se ha obtenido una vista con el número de 1546 registros y 13 variables, desplegado en un formato denominado, data.frame, en forma de matriz, así también se conocen los tipos de datos de cada variable y se logra ver el contenido de los primeros registros, en los cuales identificamos, los años de las pruebas realizadas, los nombres de los jugadores, la Universidad de procedencia, la posición OLB e ILB, la estatura en pulgadas, el peso en libras, el tiempo en las 40 yardas, el número de repeticiones con 255lbs, las medidas del salto vertical y del salto horizontal, su tiempo en las carreras de ida y regreso, así como los tiempos en el ejercicio de los tres conos; ver figura 11 con el resultado expresado de la estructura del data.frame, a través destr(mydata1987a2022) (Comando str de R)

**Figura 19**

*Estructura de la BD con Resultados de las Pruebas del Combinado 1987 a 2022*



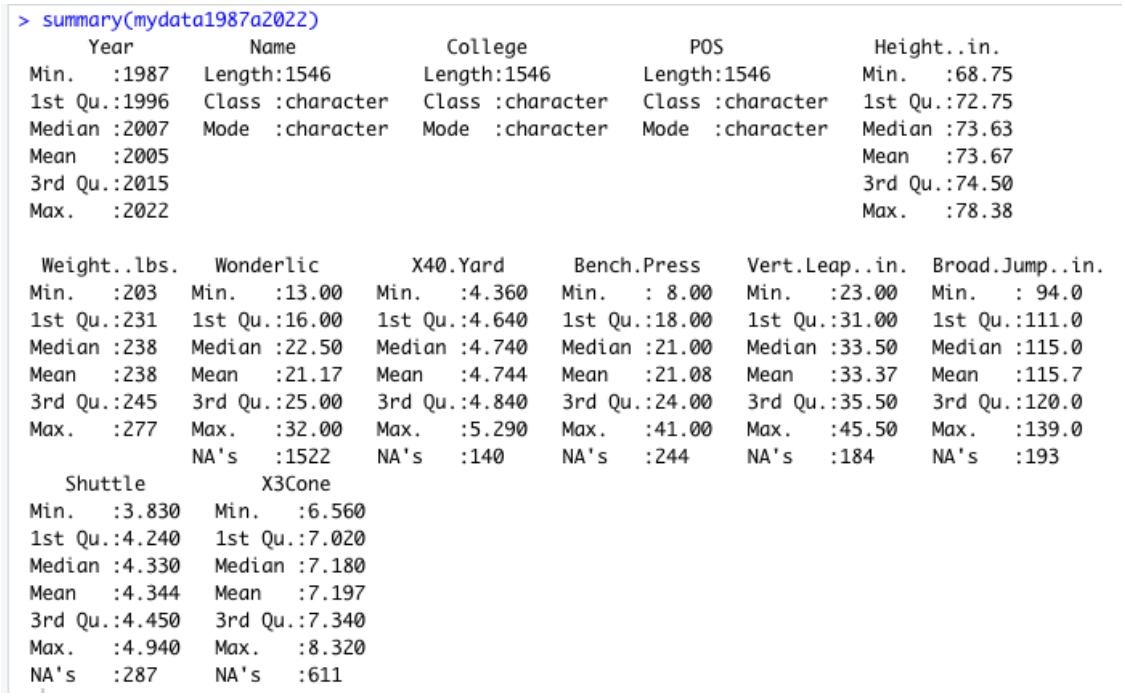

Fuente: Recuperado de la herramienta de Rstudio (reflejado en línea de commando)

Al revisar la BD (NFLCombinado_y_RANK_TOP100.csv), que ha sido creada partiendo de la primera BD, ya explicada, se encuentra que contiene un campo adicional el “RANK”, en el se observa que puede tener un número para identificar a los mejores 100 jugadores, de la temporada 2022 provisto por la PFF (PFF, 2022) adicionalmente a la variable “RANK” se le ha agregado un número “1” en aquellos jugadores identificados por la NFL como los 100 mejores de todos los tiempos; cabe mencionar que pudiera agregarse o quitarse jugadores, según opiniones distintas, es por eso que se han explicado los criterios seguidos.

Explorando con mayor profundidad, se aplica estadística, se observa que los años de recolección incluyen efectivamente de 1987 al 2022, que las estaturas de los jugadores se encuentran desde la mínima de 68.75 pulgadas (1.74 mts) hasta un máximo de 78.38 pulgada (1.91mts), que en la fuerza para cargar pesas en repeticiones de 255lb varían las repeticiones desde desde 8 hasta 41, sin embargo se ve que existen 244 datos incompletos en ese rubro, como otra cantidad considerable en las otras variables, vease figura 12 a continuación.

**Figura 20**

*Mínimos, Máximos, Cuartiles, Media, Moda de los Resultados del Combinado*



```
> summary(mydata1987a2022)
      Year          Name             College             POS             Height..in.
 Min.   :1987   Length:1546        Length:1546        Length:1546        Min.   :68.75
 1st Qu.:1996   Class :character   Class :character   Class :character   1st Qu.:72.75
 Median :2007   Mode  :character   Mode  :character   Mode  :character   Median :73.63
 Mean   :2005                                                           Mean   :73.67
 3rd Qu.:2015                                                           3rd Qu.:74.50
 Max.   :2022                                                           Max.   :78.38

  Weight..lbs.    Wonderlic        X40.Yard      Bench.Press    Vert.Leap..in.  Broad.Jump..in.
 Min.   :203   Min.   :13.00   Min.   :4.360   Min.   : 8.00   Min.   :23.00   Min.   : 94.0
 1st Qu.:231   1st Qu.:16.00   1st Qu.:4.640   1st Qu.:18.00   1st Qu.:31.00   1st Qu.:111.0
 Median :238   Median :22.50   Median :4.740   Median :21.00   Median :33.50   Median :115.0
 Mean   :238   Mean   :21.17   Mean   :4.744   Mean   :21.08   Mean   :33.37   Mean   :115.7
 3rd Qu.:245   3rd Qu.:25.00   3rd Qu.:4.840   3rd Qu.:24.00   3rd Qu.:35.50   3rd Qu.:120.0
 Max.   :277   Max.   :32.00   Max.   :5.290   Max.   :41.00   Max.   :45.50   Max.   :139.0
               NA's   :1522    NA's   :140     NA's   :244     NA's   :184     NA's   :193
    Shuttle          X3Cone
 Min.   :3.830   Min.   :6.560
 1st Qu.:4.240   1st Qu.:7.020
 Median :4.330   Median :7.180
 Mean   :4.344   Mean   :7.197
 3rd Qu.:4.450   3rd Qu.:7.340
 Max.   :4.940   Max.   :8.320
 NA's   :287     NA's   :611
```

Fuente: Elaboración del autor, con la herramienta de Rstudios y muestra los resultados obtenidos

***Conociendo "la estructura" de los datos, con el comando str()***

Se identifican las estructuras de datos: [data frames] que incluye, [vectores], [matrices]. Se visualizan 13 variables de tipo : [Integer], [Character], [númericas]. Los comandos usados en R: str = Estructura de la información Class = tipo de información (Data.Frame) Dim = Dimensiones filas x columnas Count = n (número de renglones)

```
str(mydata1987a2022)
```

```
## 'data.frame':    1546 obs. of  13 variables:
##  $ Year           : int  1987 1987 1987 1987 1987 1987 1987 1987 1987 1987
...
##  $ Name           : chr  "Doug Bartlett" "David Brandon" "Toby Caston" "Da
vid Dudley" ...
##  $ College        : chr  "Northern Illinois" "Memphis" "Louisiana State" "
Arkansas" ...
##  $ POS            : chr  "OLB" "OLB" "OLB" "OLB" ...
##  $ Height..in.    : num  74.3 75.1 73 74.1 73.8 76.8 77 73.8 73.9 74.9 ...
##  $ Weight..lbs.   : int  250 223 235 210 216 239 240 226 229 237 ...
##  $ Wonderlic      : int  NA NA NA NA NA NA NA NA NA NA ...
```

```
##  $ X40.Yard       : num  4.75 4.73 4.82 4.65 4.71 4.85 4.66 4.76 4.92 4.8
...
##  $ Bench.Press    : int  31 9 10 15 19 13 18 16 15 17 ...
##  $ Vert.Leap..in. : num  31 26.5 32.5 27.5 29.5 29 34 33.5 28 28 ...
##  $ Broad.Jump..in.: int  112 108 117 111 110 115 122 124 105 108 ...
##  $ Shuttle        : num  4.12 4.45 4.55 4.38 4.35 4.62 4.02 4.36 4.5 4.5 .
..
##  $ X3Cone         : num  NA NA NA NA NA NA NA NA NA NA ...
```

```
class(mydata1987a2022)
```

```
## [1] "data.frame"
```

```
dim(mydata1987a2022)
```

```
## [1] 1546   13
```

```
count(mydata1987a2022)
```

```
##      n
## 1 1546
```

```
str(midataRANKtop100)
```

```
## 'data.frame':    1546 obs. of  14 variables:
##  $ Year           : int  1987 1987 1987 1987 1987 1987 1987 1987 1987 1987
...
##  $ Name           : chr  "Doug Bartlett" "David Brandon" "Toby Caston" "Da
vid Dudley" ...
##  $ College        : chr  "Northern Illinois" "Memphis" "Louisiana State" "
Arkansas" ...
##  $ POS            : chr  "OLB" "OLB" "OLB" "OLB" ...
##  $ Height..in.    : num  74.3 75.1 73 74.1 73.8 76.8 77 73.8 73.9 74.9 ...
##  $ Weight..lbs.   : int  250 223 235 210 216 239 240 226 229 237 ...
##  $ Wonderlic      : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ X40.Yard       : num  4.75 4.73 4.82 4.65 4.71 4.85 4.66 4.76 4.92 4.8
...
##  $ Bench.Press    : int  31 9 10 15 19 13 18 16 15 17 ...
##  $ Vert.Leap..in. : num  31 26.5 32.5 27.5 29.5 29 34 33.5 28 28 ...
##  $ Broad.Jump..in.: int  112 108 117 111 110 115 122 124 105 108 ...
##  $ Shuttle        : num  4.12 4.45 4.55 4.38 4.35 4.62 4.02 4.36 4.5 4.5 .
..
##  $ X3Cone         : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ RANK           : int  NA NA NA NA NA NA NA NA NA NA ...
```

```
class(midataRANKtop100)
```

```
## [1] "data.frame"
```

```
dim(midataRANKtop100)
```

```
## [1] 1546   14
```

**Transformación de los datos, agregación & conversión**

Para simplificar la observación se decide transformar el peso de los jugadores de libras a kilogramos, para ello se adiciona la variable: “PesoKg” a la cual se le aplica la fórmula de conversión de libras a kilogramos (dividiendo libras entre 2.2) Con el comando glimpse() se observa de la variable 1 a la 14 y/o 15, junto ocn los primeros registros

```
mydata1987a2022$PesoKg <- mydata1987a2022$Weight..lbs./2.2
glimpse(mydata1987a2022[1:14])
```

```
## Rows: 1,546
## Columns: 14
```

```
## $ Year       <int> 1987, 1987, 1987, 1987, 1987, 1987, 1987, 1987, 1987, …
## $ Name       <chr> "Doug Bartlett", "David Brandon", "Toby Caston", "Davi…
## $ College    <chr> "Northern Illinois", "Memphis", "Louisiana State", "Ar…
## $ POS        <chr> "OLB", "OLB", "OLB", "OLB", "OLB", "OLB", "OLB", "OLB"…
## $ Height..in.<dbl> 74.3, 75.1, 73.0, 74.1, 73.8, 76.8, 77.0, 73.8, 73.9, …
## $ Weight.lbs <int> 250, 223, 235, 210, 216, 239, 240, 226, 229, 237, 240,…
## $ Wonderlic  <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…
## $ X40.Yard   <dbl> 4.75, 4.73, 4.82, 4.65, 4.71, 4.85, 4.66, 4.76, 4.92, …
## $ Bench.Press<int> 31, 9, 10, 15, 19, 13, 18, 16, 15, 17, 22, 21, 12, 21,…
## $ Vert.L  <dbl> 31.0, 26.5, 32.5, 27.5, 29.5, 29.0, 34.0, 33.5, 28.0, …
## $ Broad.J <int> 112, 108, 117, 111, 110, 115, 122, 124, 105, 108, 116,…
## $ Shuttle <dbl> 4.12, 4.45, 4.55, 4.38, 4.35, 4.62, 4.02, 4.36, 4.50, …
## $ X3Cone  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…
## $ PesoKg  <dbl> 113.63636, 101.36364, 106.81818, 95.45455, 98.18182, 1…
```

```
midataRANKtop100$PesoKg <-midataRANKtop100$Weight..lbs./2.2
glimpse(midataRANKtop100[1:15])
```

```
## Rows: 1,546
## Columns: 15
## $ Year            <int> 1987, 1987, 1987, 1987, 1987, 1987, 1987, 1987, 1987, …
## $ Name            <chr> "Doug Bartlett", "David Brandon", "Toby Caston", "Davi…
## $ College         <chr> "Northern Illinois", "Memphis", "Louisiana State", "Ar…
## $ POS             <chr> "OLB", "OLB", "OLB", "OLB", "OLB", "OLB", "OLB", "OLB"…
## $ Height..in.     <dbl> 74.3, 75.1, 73.0, 74.1, 73.8, 76.8, 77.0, 73.8, 73.9, …
## $ Weight..lbs.    <int> 250, 223, 235, 210, 216, 239, 240, 226, 229, 237, 240,…
## $ Wonderlic       <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…
## $ X40.Yard        <dbl> 4.75, 4.73, 4.82, 4.65, 4.71, 4.85, 4.66, 4.76, 4.92, …
## $ Bench.Press     <int> 31, 9, 10, 15, 19, 13, 18, 16, 15, 17, 22, 21, 12, 21,…
## $ Vert.Leap..in.  <dbl> 31.0, 26.5, 32.5, 27.5, 29.5, 29.0, 34.0, 33.5, 28.0, …
## $ Broad.Jump..in. <int> 112, 108, 117, 111, 110, 115, 122, 124, 105, 108, 116,…
## $ Shuttle         <dbl> 4.12, 4.45, 4.55, 4.38, 4.35, 4.62, 4.02, 4.36, 4.50, …
## $ X3Cone          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…
## $ RANK            <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…
## $ PesoKg          <dbl> 113.63636, 101.36364, 106.81818, 95.45455, 98.18182, 1…
```

**Transforma los datos, ordenar & filtrar**

Se crea una lista, que contiene a los jugadores elite y se le ORDENA alfabeticamente de la A a la Z; adicionalmente, se crea una lista de los jugadores sin rank (es decir no elite)

```
myRANKPlayers <- midataRANKtop100[order(midataRANKtop100$RANK),] %>%
          filter(!is.na(RANK))

my_NO_RANKPlayers <- midataRANKtop100 %>% filter(is.na(RANK))

str(myRANKPlayers)
```

```
## 'data.frame':    76 obs. of  15 variables:
##  $ Year           : int  1987 1988 1988 1989 1990 1991 1993 1994 1995 1996 ...
##  $ Name           : chr  "Hardy Nickerson" "Chris Spielman" "Bill Romanowski"
##  $ College        : chr  "California" "Ohio State" "Boston College" "Alabama"
##  $ POS            : chr  "ILB" "ILB" "OLB" "OLB" ...
##  $ Height..in.    : num  73.5 72 75.5 74.5 76.1 75.3 74.5 75.3 72.3 72.4 ...
##  $ Weight..lbs.   : int  223 234 231 234 238 235 236 235 229 235 ...
##  $ Wonderlic      : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ X40.Yard       : num  4.82 4.9 4.76 4.58 5.04 5 4.85 4.9 4.71 NA ...
##  $ Bench.Press    : int  19 23 NA NA 18 9 NA 18 NA NA ...
##  $ Vert.Leap..in. : num  28 32.5 31.5 NA 31 27.5 33 36 NA NA ...
##  $ Broad.Jump..in.: int  108 109 111 NA 115 112 117 116 NA NA ...
##  $ Shuttle        : num  4.15 4.13 4.25 NA 4.17 4.47 4.15 4.4 NA NA ...
##  $ X3Cone         : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ RANK           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ PesoKg         : num  101 106 105 106 108 ...
```

```
str(my_NO_RANKPlayers)
```

```
## 'data.frame':    1470 obs. of  15 variables:
##  $ Year           : int  1987 1987 1987 1987 1987 1987 1987 1987 1987 1987
##  $ Name           : chr  "Doug Bartlett" "David Brandon" "Toby Caston"
##  $ College        : chr  "Northern Illinois" "Memphis" "Louisiana State"
##  $ POS            : chr  "OLB" "OLB" "OLB" "OLB" ...
##  $ Height..in.    : num  74.3 75.1 73 74.1 73.8 76.8 77 73.8 73.9 74.9 ...
##  $ Weight..lbs.   : int  250 223 235 210 216 239 240 226 229 237 ...
##  $ Wonderlic      : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ X40.Yard       : num  4.75 4.73 4.82 4.65 4.71 4.85 4.66 4.76 4.92 4.8
##  $ Bench.Press    : int  31 9 10 15 19 13 18 16 15 17 ...
##  $ Vert.Leap..in. : num  31 26.5 32.5 27.5 29.5 29 34 33.5 28 28 ...
##  $ Broad.Jump..in.: int  112 108 117 111 110 115 122 124 105 108 ...
##  $ Shuttle        : num  4.12 4.45 4.55 4.38 4.35 4.62 4.02 4.36 4.5 4.5
```

## Estadística

***Explorando y analizando la estructura y la distribución estadística.***

Se aprecian los, [mínimos], [máximos], [cuartiles], [Media], [Moda] de cada una de las las 13 variables.

```
summary(mydata1987a2022)
```

```
##       Year          Name              College              POS
##  Min.   :1987   Length:1546        Length:1546        Length:1546
##  1st Qu.:1996   Class :character   Class :character   Class :character
##  Median :2007   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2005
##  3rd Qu.:2015
##  Max.   :2022
##
##   Height..in.     Weight..lbs.   Wonderlic        X40.Yard      Bench.Pres
s
##  Min.   :68.75   Min.   :203   Min.   :13.00   Min.   :4.360   Min. : 8.00
##  1st Qu.:72.75   1st Qu.:231   1st Qu.:16.00   1st Qu.:4.640   1stQu. :18.00
##  Median :73.63   Median :238   Median :22.50   Median :4.740   Median :21.00
##  Mean   :73.67   Mean   :238   Mean   :21.17   Mean   :4.744   Mean   :21.08
##  3rd Qu.:74.50   3rd Qu.:245   3rd Qu.:25.00   3rd Qu.:4.840   3rd Qu.:24.00
##  Max.   :78.38   Max.   :277   Max.   :32.00   Max.   :5.290   Max.   :41.00
##                                NA's   :1522    NA's   :140     NA's   :244
##  Vert.Leap..in.  Broad.Jump..in.    Shuttle         X3Cone
##  Min.   :23.00   Min.   : 94.0   Min.   :3.830   Min.   :6.560
##  1st Qu.:31.00   1st Qu.:111.0   1st Qu.:4.240   1st Qu.:7.020
##  Median :33.50   Median :115.0   Median :4.330   Median :7.180
##  Mean   :33.37   Mean   :115.7   Mean   :4.344   Mean   :7.197
##  3rd Qu.:35.50   3rd Qu.:120.0   3rd Qu.:4.450   3rd Qu.:7.340
##  Max.   :45.50   Max.   :139.0   Max.   :4.940   Max.   :8.320
##  NA's   :184     NA's   :193     NA's   :287     NA's   :611
```

```
summary(midataRANKtop100)
```

```
##       Year          Name              College              POS
##  Min.   :1987   Length:1546        Length:1546        Length:1546
##  1st Qu.:1996   Class :character   Class :character   Class :character
##  Median :2007   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2005
##  3rd Qu.:2015
##  Max.   :2022
##
```

```
##   Height..in.     Weight..lbs.   Wonderlic        X40.Yard      Bench.Press
##  Min.   :68.75   Min.   :203   Min.   :13.00   Min.   :4.360   Min.   : 8.00
##  1st Qu.:72.75   1st Qu.:231   1st Qu.:16.00   1st Qu.:4.640   1st Qu.:18.00
##  Median :73.63   Median :238   Median :22.50   Median :4.740   Median :21.00
##  Mean   :73.67   Mean   :238   Mean   :21.17   Mean   :4.744   Mean   :21.08
##  3rd Qu.:74.50   3rd Qu.:245   3rd Qu.:25.00   3rd Qu.:4.840   3rd Qu.:24.00
##  Max.   :78.38   Max.   :277   Max.   :32.00   Max.   :5.290   Max.   :41.00
##                                NA's   :1522    NA's   :140     NA's   :244
##  Vert.Leap..in.  Broad.Jump..in.    Shuttle         X3Cone
##  Min.   :23.00   Min.   : 94.0   Min.   :3.830   Min.   :6.560
##  1st Qu.:31.00   1st Qu.:111.0   1st Qu.:4.240   1st Qu.:7.020
##  Median :33.50   Median :115.0   Median :4.330   Median :7.180
##  Mean   :33.37   Mean   :115.7   Mean   :4.344   Mean   :7.197
##  3rd Qu.:35.50   3rd Qu.:120.0   3rd Qu.:4.450   3rd Qu.:7.340
##  Max.   :45.50   Max.   :139.0   Max.   :4.940   Max.   :8.320
##  NA's   :184     NA's   :193     NA's   :287     NA's   :611
##       RANK
##  Min.   : 1.00
##  1st Qu.: 1.00
##  Median :17.50
##  Mean   :24.58
##  3rd Qu.:43.25
##  Max.   :81.00
##  NA's   :1470
```

**Mínimos, Máximos, Media, Moda y Cuartiles.**

Se “crea” un “matriz” con el peso del jugador vs la velocidad, se observa, los comportamientos extremos, es decir el más pesado y el más rápido, así también se observa el comportamiento de la media y más importante la moda

```
summary(myRANKPlayers[c("PesoKg", "X40.Yard")])
```

```
##      PesoKg         X40.Yard
##  Min.   :100.5   Min.   :4.42
##  1st Qu.:105.8   1st Qu.:4.56
##  Median :107.3   Median :4.64
##  Mean   :107.9   Mean   :4.66
##  3rd Qu.:109.7   3rd Qu.:4.76
##  Max.   :118.6   Max.   :5.04
##                  NA's   :7
```

```
summary(my_NO_RANKPlayers[c("PesoKg", "X40.Yard")])
```

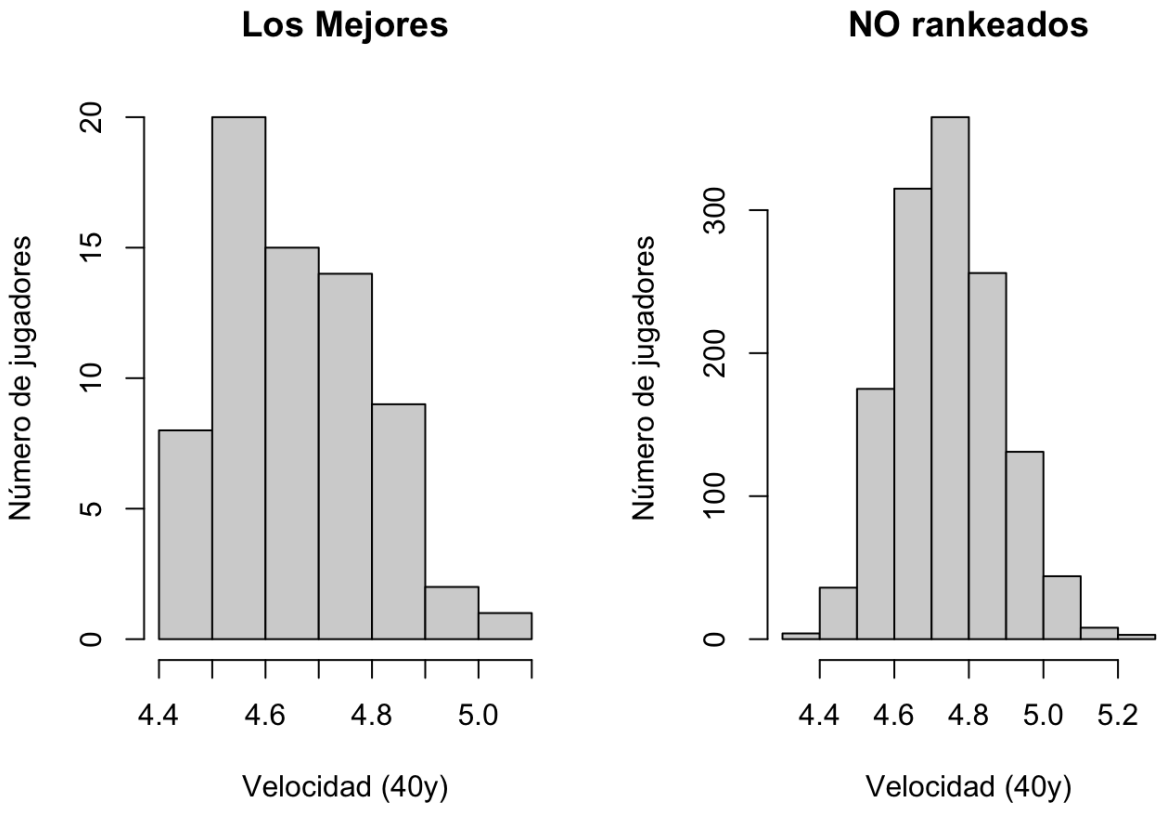
```
##      PesoKg          X40.Yard
##  Min.   : 92.27   Min.   :4.360
##  1st Qu.:105.00   1st Qu.:4.650
##  Median :108.18   Median :4.750
##  Mean   :108.20   Mean   :4.748
```

```
##  3rd Qu.:111.36   3rd Qu.:4.840
##  Max.   :125.91   Max.   :5.290
##                   NA's   :133
```

**Gráfico de histograma, la distribución de los jugadores por velocidad**

Se hace un comparativo con el histograma, para conocer la distribución de los jugadores acorde a su velocidad en las 40 yardas

```
par(mfrow=c(1,2))
hist(myRANKPlayers$X40.Yard,
     main = "Los Mejores",
     xlab = "Velocidad (40y)",
     ylab = "Número de jugadores")
hist(my_NO_RANKPlayers$X40.Yard,
     main = "NO rankeados",
     xlab = "Velocidad (40y)",
     ylab = "Número de jugadores")
```



**Comparando con el mercado**

Se analizan dos puntos de vista contrastantes para comprender lo que implica la velocidad de estos jugadores apoyadores, por un lado el famoso comentarista de la NFL y Podcast Rich Eisen, por correr cada año (para recolectar ayuda, sus tiempos están fuera de la media inclusive de los jugadores más lentos, lo cual lo deja fuera de nuestra selección; ahora bien el otro lado de la moneda, el hombre más rápido de las olimpiadas y mundiales Usain Bolt, quien participara con la NFL para hacer la prueba de las 40 yardas, obteniendo un tiempo de 4.22 segundos, empatando a uno de los receptores abiertos más rápidos John Ross que lograra mismo tiempo en las pruebas del 2017. Ambas personas (atleta y receptor) podrían fácilmente vencer y dejar muy a tras a cualquiera de los apoyadores.

**Figura 21**
*Rich Eisen y sus tiempos en las 40 yardas*


Fuente: NFL (NFL, 2022); https://www.nfl.com/videos/run-rich-run-year-by-year-results-of-rich-eisen-s-40-yard-dash

**Transformando factor() "calificador" tipo [factor]**

El campo factor ayudará hacer las preparaciones para el análisis de modelos estadísticos más adelante.

```
mydata_trained <- midataRANKtop100
mydata_trained$Clasifica = ifelse(is.na(mydata_trained$RANK), "N", "Y")
mydata_trained$Clasifica <- factor (mydata_trained$Clasifica,
                                    levels = c("Y", "N"),
```

```
                                    labels = c("Yes_Ranked", "NO_Ranked"))
```

Se revisa el porcentaje de elementos calificados como elite () se observa una cifra muy baja de la muestra par entrenar de sólo un 4.9% de los participantes en el combinado scout han estacado como profesionales

```
round(prop.table(table(mydata_trained$Clasifica)) *100, digit =1)
```

```
##
## Yes_Ranked  NO_Ranked
##        4.9       95.1
```

**Ordenar y mostrar factor()**

```
mydata_trained <- mydata_trained[order(mydata_trained$RANK),]

table(mydata_trained$Clasifica)
```

```
##
## Yes_Ranked  NO_Ranked
##         76       1470
```

```
str(mydata_trained)
```

```
## 'data.frame':    1546 obs. of  16 variables:
##  $ Year           : int  1987 1988 1988 1989 1990 1991 1993 1994 1995 1996
##  $ Name           : chr  "Hardy Nickerson" "Chris Spielman" "Bill Romanows
ki" "Derrick Thomas" ...
##  $ College        : chr  "California" "Ohio State" "Boston College"...
##  $ POS            : chr  "ILB" "ILB" "OLB" "OLB" ...
##  $ Height..in.    : num  73.5 72 75.5 74.5 76.1 75.3 74.5 75.3 72.3 72.4
##  $ Weight..lbs.   : int  223 234 231 234 238 235 236 235 229 235 ...
##  $ Wonderlic      : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ X40.Yard       : num  4.82 4.9 4.76 4.58 5.04 5 4.85 4.9 4.71 NA ...
##  $ Bench.Press    : int  19 23 NA NA 18 9 NA 18 NA NA ...
##  $ Vert.Leap..in. : num  28 32.5 31.5 NA 31 27.5 33 36 NA NA ...
##  $ Broad.Jump..in.: int  108 109 111 NA 115 112 117 116 NA NA ...
##  $ Shuttle        : num  4.15 4.13 4.25 NA 4.17 4.47 4.15 4.4 NA NA ...
##  $ X3Cone         : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ RANK           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ PesoKg         : num  101 106 105 106 108 ...
##  $ Clasifica      : Factor w/ 2 levels "Yes_Ranked","NO_Ranked": 1 1 1 1 1
```

```
summary(mydata_trained)
```

```
##       Year          Name             College             POS
##  Min.   :1987   Length:1546        Length:1546        Length:1546
##  1st Qu.:1996   Class :character   Class :character   Class :character
##  Median :2007   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2005
##  3rd Qu.:2015
```

```
##  Max.   :2022
##
##   Height..in.     Weight..lbs.   Wonderlic        X40.Yard      Bench.Press
##  Min.   :68.75   Min.   :203   Min.   :13.00   Min.   :4.360   Min.   : 8.00
##  1st Qu.:72.75   1st Qu.:231   1st Qu.:16.00   1st Qu.:4.640   1st Qu.:18.00
##  Median :73.63   Median :238   Median :22.50   Median :4.740   Median :21.00
##  Mean   :73.67   Mean   :238   Mean   :21.17   Mean   :4.744   Mean   :21.08
##  3rd Qu.:74.50   3rd Qu.:245   3rd Qu.:25.00   3rd Qu.:4.840   3rd Qu.:24.00
##  Max.   :78.38   Max.   :277   Max.   :32.00   Max.   :5.290   Max.   :41.00
##                                NA's   :1522    NA's   :140     NA's   :244
##  Vert.Leap..in.  Broad.Jump..in.    Shuttle          X3Cone
##  Min.   :23.00   Min.   : 94.0   Min.   :3.830   Min.   :6.560
##  1st Qu.:31.00   1st Qu.:111.0   1st Qu.:4.240   1st Qu.:7.020
##  Median :33.50   Median :115.0   Median :4.330   Median :7.180
##  Mean   :33.37   Mean   :115.7   Mean   :4.344   Mean   :7.197
##  3rd Qu.:35.50   3rd Qu.:120.0   3rd Qu.:4.450   3rd Qu.:7.340
##  Max.   :45.50   Max.   :139.0   Max.   :4.940   Max.   :8.320
##  NA's   :184     NA's   :193     NA's   :287     NA's   :611
##       RANK           PesoKg           Clasifica
##  Min.   : 1.00   Min.   : 92.27   Yes_Ranked:  76
##  1st Qu.: 1.00   1st Qu.:105.00   NO_Ranked :1470
##  Median :17.50   Median :108.18
##  Mean   :24.58   Mean   :108.18
##  3rd Qu.:43.25   3rd Qu.:111.36
##  Max.   :81.00   Max.   :125.91
##  NA's   :1470
```

```
summary(myRANKPlayers)
```

```
##       Year          Name             College             POS
##  Min.   :1987   Length:76          Length:76          Length:76
##  1st Qu.:2002   Class :character   Class :character   Class :character
##  Median :2015   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2011
##  3rd Qu.:2018
##  Max.   :2022
##
##   Height..in.     Weight..lbs.     Wonderlic        X40.Yard     Bench.Press
##  Min.   :70.75   Min.   :221.0   Min.   :14.00   Min.   :4.42   Min.   : 9.00
##  1st Qu.:72.47   1st Qu.:232.8   1st Qu.:15.50   1st Qu.:4.56   1st Qu.:19.00
##  Median :73.50   Median :236.0   Median :17.00   Median :4.64   Median :22.00
##  Mean   :73.65   Mean   :237.4   Mean   :19.67   Mean   :4.66   Mean   :21.81
##  3rd Qu.:74.78   3rd Qu.:241.2   3rd Qu.:22.50   3rd Qu.:4.76   3rd Qu.:24.00
##  Max.   :76.88   Max.   :261.0   Max.   :28.00   Max.   :5.04   Max.   :32.00
##                                  NA's   :73      NA's   :7      NA's   :17
##  Vert.Leap..in.  Broad.Jump..in.    Shuttle          X3Cone
##  Min.   :27.50   Min.   :104     Min.   :4.060   Min.   :6.630
##  1st Qu.:33.00   1st Qu.:116     1st Qu.:4.150   1st Qu.:6.900
##  Median :35.00   Median :119     Median :4.250   Median :7.100
##  Mean   :34.99   Mean   :119     Mean   :4.266   Mean   :7.092
##  3rd Qu.:38.00   3rd Qu.:123     3rd Qu.:4.395   3rd Qu.:7.210
##  Max.   :40.00   Max.   :132     Max.   :4.500   Max.   :7.880
```

```
##  NA's   :8       NA's   :5       NA's   :17      NA's   :27
##       RANK           PesoKg
##  Min.   : 1.00   Min.   :100.5
##  1st Qu.: 1.00   1st Qu.:105.8
##  Median :17.50   Median :107.3
##  Mean   :24.58   Mean   :107.9
##  3rd Qu.:43.25   3rd Qu.:109.7
##  Max.   :81.00   Max.   :118.6
##
```

## ANEXO J

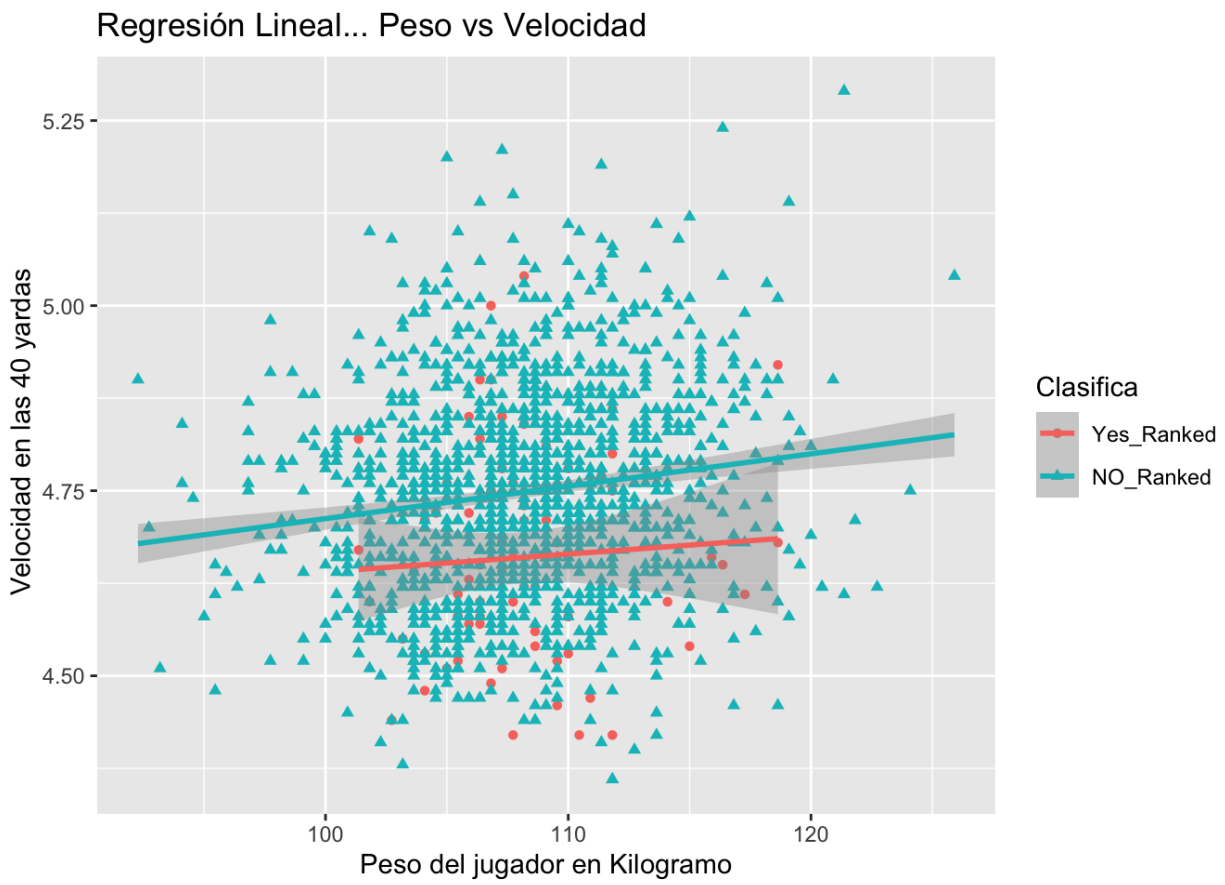
## Regresión Lineal

***Explorando los datos con "regresión lineal***

La regresión lineal simple consiste en estudiar los cambios de la variable no aleatoria sobre otra variable aleatoria, la relación funcional entre las variables se establece por la expresión lineal, la cual es representada por una línea recta (Alicia Vila, 2018). "La regresión permite y se utiliza para establecer predicciones, estimaciones y pronósticos con la historia de los datos; a la variable que se va a predecir se le llama variable dependiente y a la variable o variables que se usan para predecir el valor de la variable dependiente se les llama variables independientes o predictoras".(Mendenhall, Beaver, & Beaver, 2010).

Se analiza a través de la regresión líneal como afecta el PESO en Kg, en la variable velocidad de los jugadores sobre las 40 yardas recorridas:

```
p <- ggplot(mydata_trained, aes( x = PesoKg, y = X40.Yard, colour = Clasific
a))
p <- p + geom_point(aes(shape = Clasifica ))
p <- p + xlab("Peso del jugador en Kilogramo") +
         ylab(" Velocidad en las 40 yardas") +
         ggtitle("Regresión Lineal... Peso vs Velocidad")
p <- p + geom_smooth(method = "lm")
p
```




En la figura de regresión lineal del resultado, se aprecia la tendencia marcada, a mayor peso de los jugadores, mayores los tiempos en las 40 yardas, es decir son más lentos.

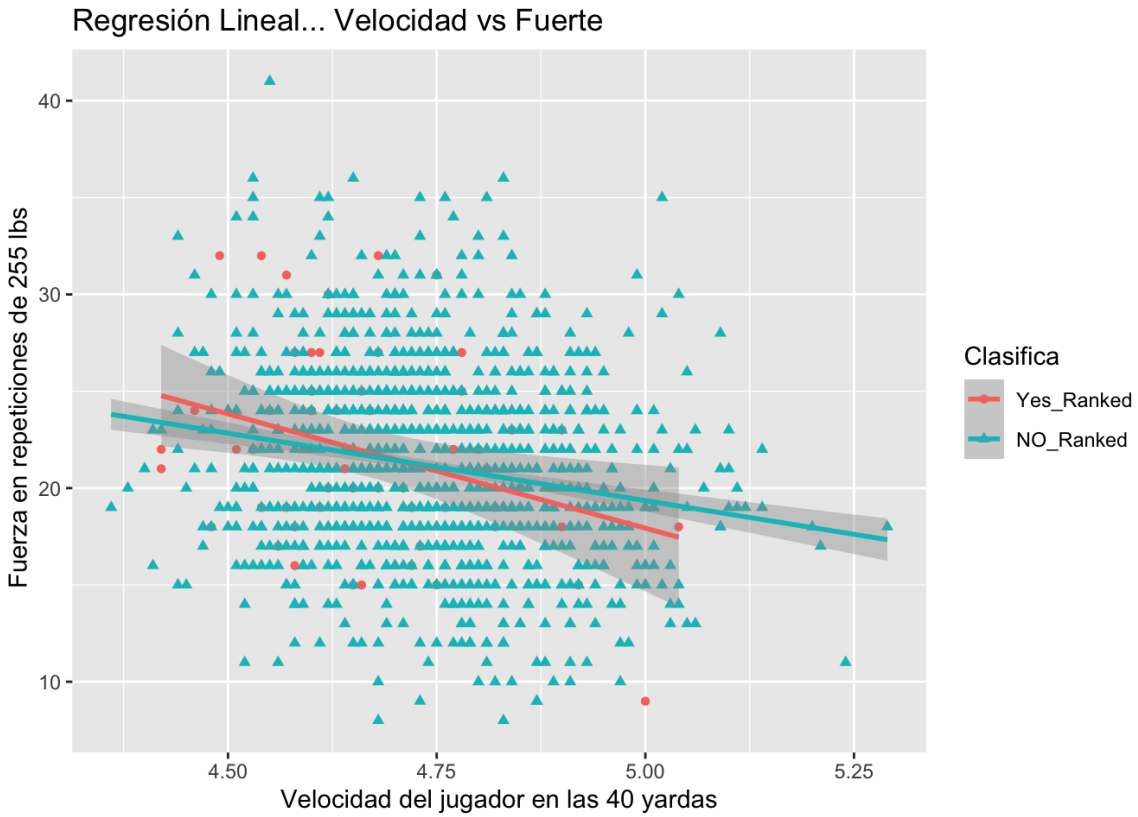
**Observación**, los jugadores rankeados son más veloces y más ligeros, como lo muestra la línea roja(tenue).

**Explorando los datos usando regresión lineal**

Se estudia con el uso de la regresión lineal, la correlación entre la velocidad y la fuerza; la Fuerza es medida por el número de repeticiones que puedan hacer los jugadores en bench press vs velocidad de los atletas en las 40 yardas. Se aprecia 15 de 18 jugadores (elite) rápidos y pueden ejecutar 15 o más repeticiones en bench press

```
pp <- ggplot(mydata_trained, aes( x = X40.Yard, y = Bench.Press,
                                  colour = Clasifica))
pp <- pp + geom_point(aes(shape = Clasifica ))
pp <- pp + xlab("Velocidad del jugador en las 40 yardas") +
           ylab(" Fuerza en repeticiones de 255 lbs") +
```

```
          ggtitle("Regresión Lineal... Velocidad vs Fuerte")
pp <- pp + geom_smooth(method = "lm")
pp
```




**Se observa** la correlación, los jugadores más fuertes, también tienden a ser más veloces y ágiles; sin embargo los mejores jugadores no siempre son más fuerte, se cruza la gráfica de regresión.

**Explorando los datos con tablas**

***Nota hallazgo***

Dentro del combinado se ha encontrado una variable llamada WONDERLIC, en ella se plasma una calificación a los jugadores considerados como las mejores selección, sin embargo, la métrica de Wonder sólo ha sido exitosa (acorde a este estudio) en 3 ocaciones de 24 predicciones realizadas, es decir es un 0.125 confiable

```
wondervsrank <- (mydata_trained %>%
                  filter(!is.na(Wonderlic)) %>%
                 select(Wonderlic, Clasifica))
```

```
wondervsrank
```

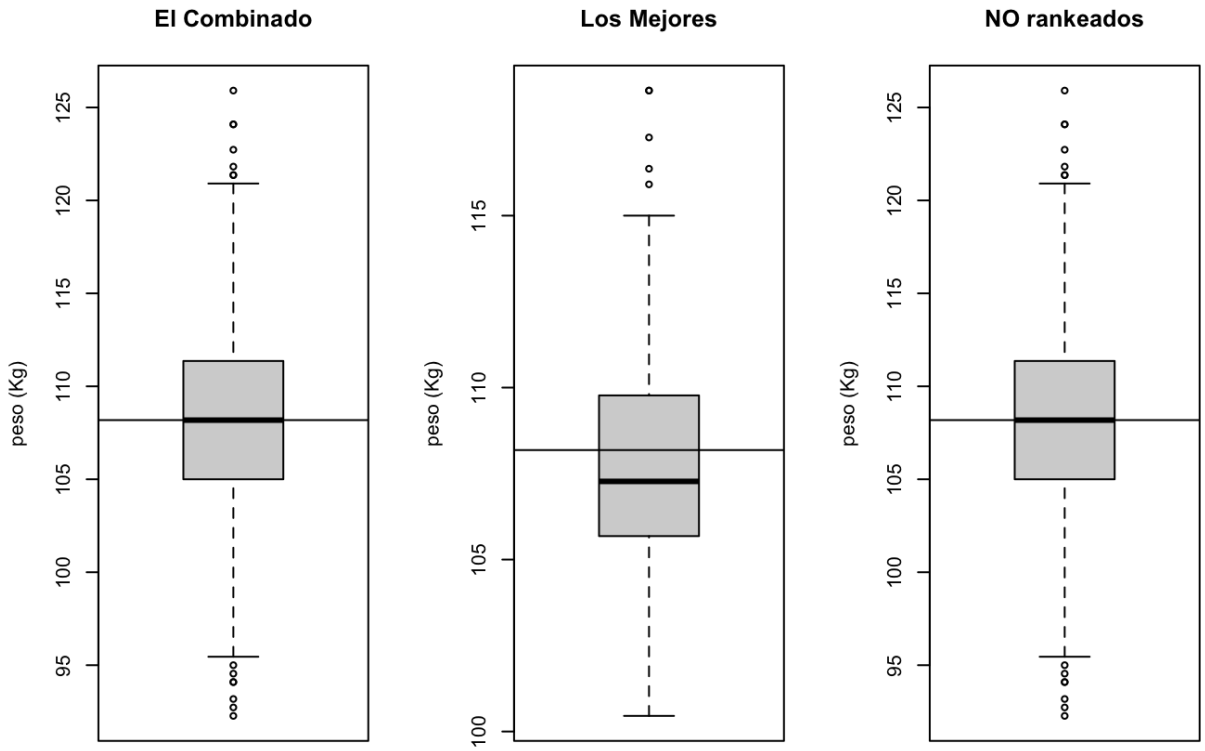
```
##    Wonderlic  Clasifica
## 1         17 Yes_Ranked
## 2         28 Yes_Ranked
## 3         14 Yes_Ranked
## 4         15  NO_Ranked
## 5         23  NO_Ranked
## 6         27  NO_Ranked
## 7         23  NO_Ranked
## 8         23  NO_Ranked
## 9         20  NO_Ranked
## 10        17  NO_Ranked
## 11        25  NO_Ranked
## 12        13  NO_Ranked
## 13        16  NO_Ranked
## 14        14  NO_Ranked
## 15        32  NO_Ranked
## 16        23  NO_Ranked
## 17        20  NO_Ranked
## 18        14  NO_Ranked
## 19        29  NO_Ranked
## 20        16  NO_Ranked
## 21        29  NO_Ranked
## 22        22  NO_Ranked
## 23        25  NO_Ranked
## 24        23  NO_Ranked
```

***Explorando los datos con gráficas***

Vistas comparativa con tres perspectivas, se aprecia el peso, la vista on todos los jugadores del combinado y se marca la media (promedio) la segunda vista “Los Mejores Jugadores”, y se ve por debajo de la media general, demostrando que son más ligeros los jugadores (elite)

```
par(mfrow=c(1,3))
boxplot(mydata1987a2022$PesoKg,
        main = "El Combinado",
        ylab = "peso (Kg)")
abline(h=mean(mydata1987a2022$PesoKg))
boxplot(myRANKPlayers$PesoKg,
        main = "Los Mejores",
        ylab = "peso (Kg)")
abline(h=mean(mydata1987a2022$PesoKg))
boxplot(my_NO_RANKPlayers$PesoKg,
        main = "NO rankeados",
        ylab = "peso (Kg)")
```

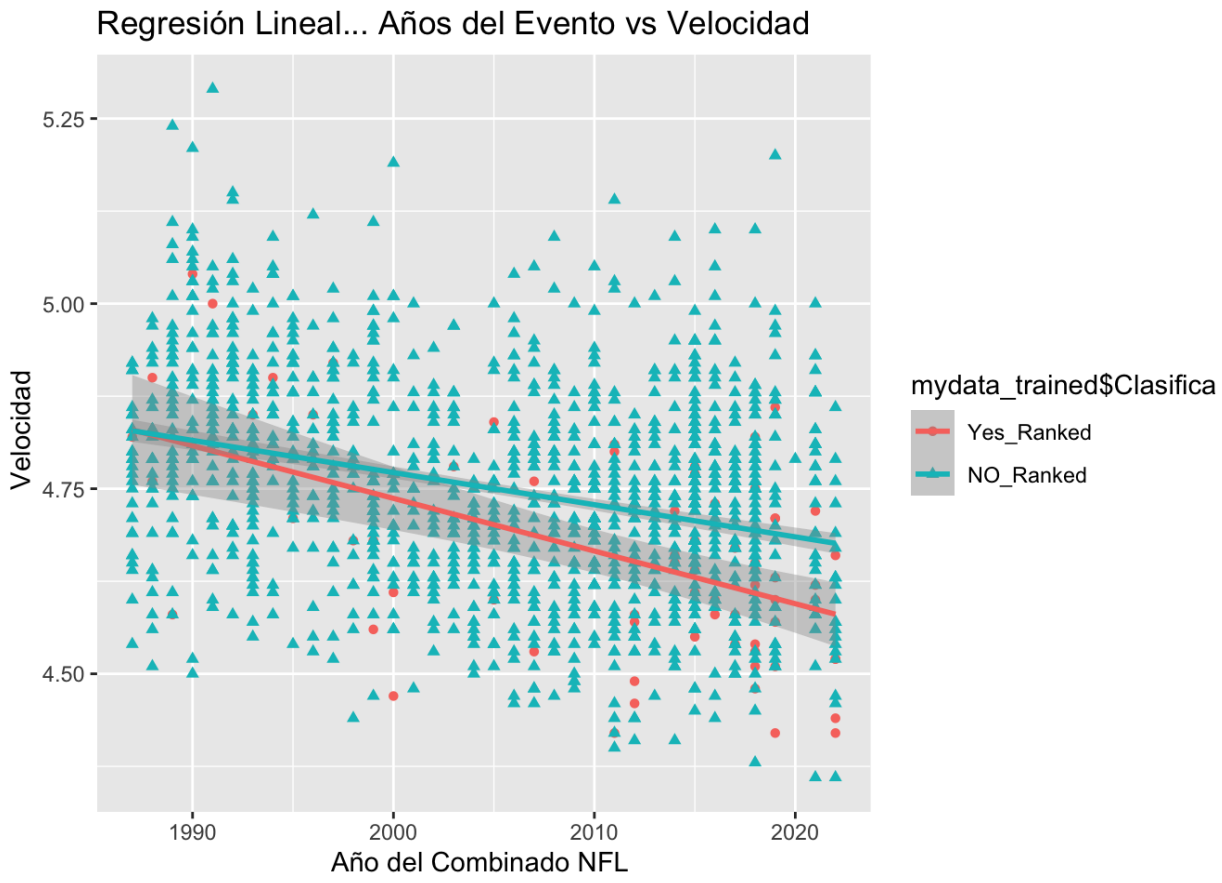
```
abline(h=mean(mydata1987a2022$PesoKg))
```




**Explorando los datos usando la regresión lineal**

La tendencia dictan que, en los años más recientes de las pruebas, los jugadores han reducido los tiempos en las 40 yardas

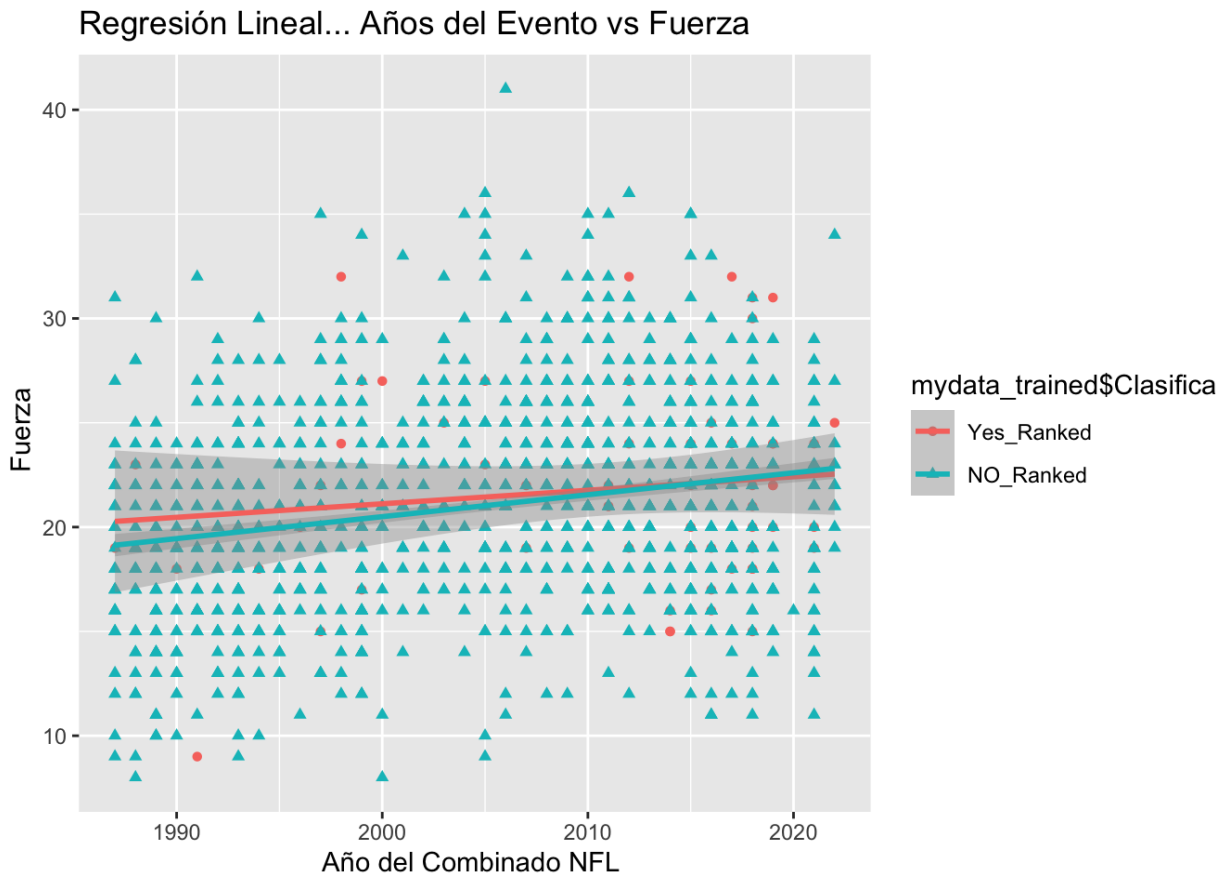
```
pp1 <- ggplot(mydata_trained, aes( x = mydata_trained$Year,
                                   y = mydata_trained$X40.Yard,
                                   colour = mydata_trained$Clasifica))
pp1 <- pp1 + geom_point(aes(shape = mydata_trained$Clasifica ))
pp1 <- pp1 + xlab("Año del Combinado NFL ") +
             ylab(" Velocidad") +
             ggtitle("Regresión Lineal... Años del Evento vs Velocidad")
pp1 <- pp1 + geom_smooth(method = "lm")
pp1
```




Observaciones, los jugadores en epocas recientes son más rápidos y también muestarn mayor fuerza.

Nota: Dentro de los resultados de la regresión lineal, muestra una marcada tendencia con los jugadores elite más rápidos que el resto.

```
pp2 <- ggplot(mydata_trained, aes( x = mydata_trained$Year,
                                   y = mydata_trained$Bench.Press,
                                   colour = mydata_trained$Clasifica))
pp2 <- pp2 + geom_point(aes(shape = mydata_trained$Clasifica ))
pp2 <- pp2 + xlab("Año del Combinado NFL ") +
             ylab(" Fuerza") +
             ggtitle("Regresión Lineal... Años del Evento vs Fuerza")
pp2 <- pp2 + geom_smooth(method = "lm")
pp2
```




Observaciones, los jugadores en epocas recientes son más fuertes y no hay gran diferencia entre la elite y la media

**Evaluación, correlación estadística peso vs velocidad**

En probabilidad y estadística, la correlación indica la fuerza y la dirección de una relación lineal y proporcionalidad entre dos atributos estadísticos. Se determina que las dos variables están correlacionadas cuando los valores de alguna de ellas, se comporta de manera homogénea y sistemáticamente igual con respecto a los valores de la otra, existe correlación entre ellas si al disminuir los valores de una lo hacen también los de la otra variable y viceversa.

```
cor(mydata_limpia_no_NA$PesoKg, mydata_limpia_no_NA$X40.Yard)
```

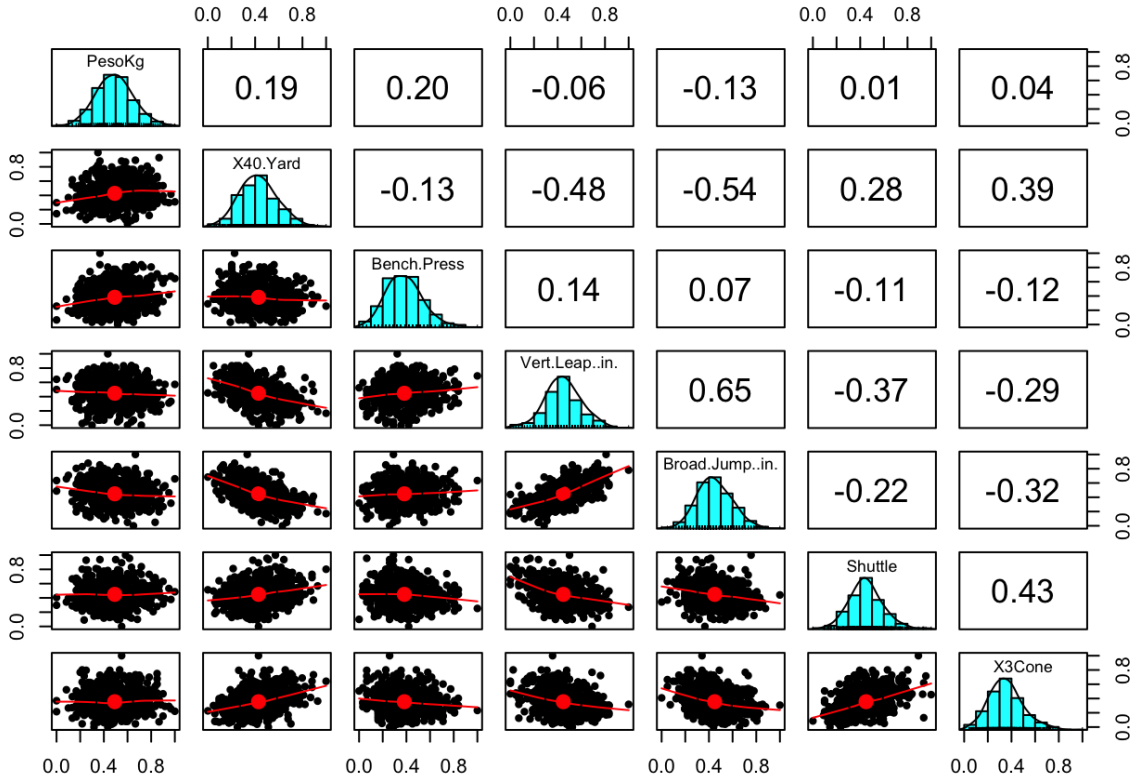
```
## [1] 0.1938715
```

El resultado que se aprecia es de 0.1938715 entre estas dos variables; como siguiente estudiará en forma matricial las variables.

**Matriz correlacional de eventos**

La siguiente figura, contiene la matriz de diagrama de dispersión entre los pares. La diagonal contiene histogramas que representan la distribución de valores para cada característica. Los diagramas de dispersión debajo de la diagonal se presentan con información visual adicional; el objeto de forma ovalada en cada diagrama de dispersión es una elipse de correlación; proporciona una visualización de la fuerza de la correlación. Cuanto más se estira la elipse, más fuerte es la correlación.

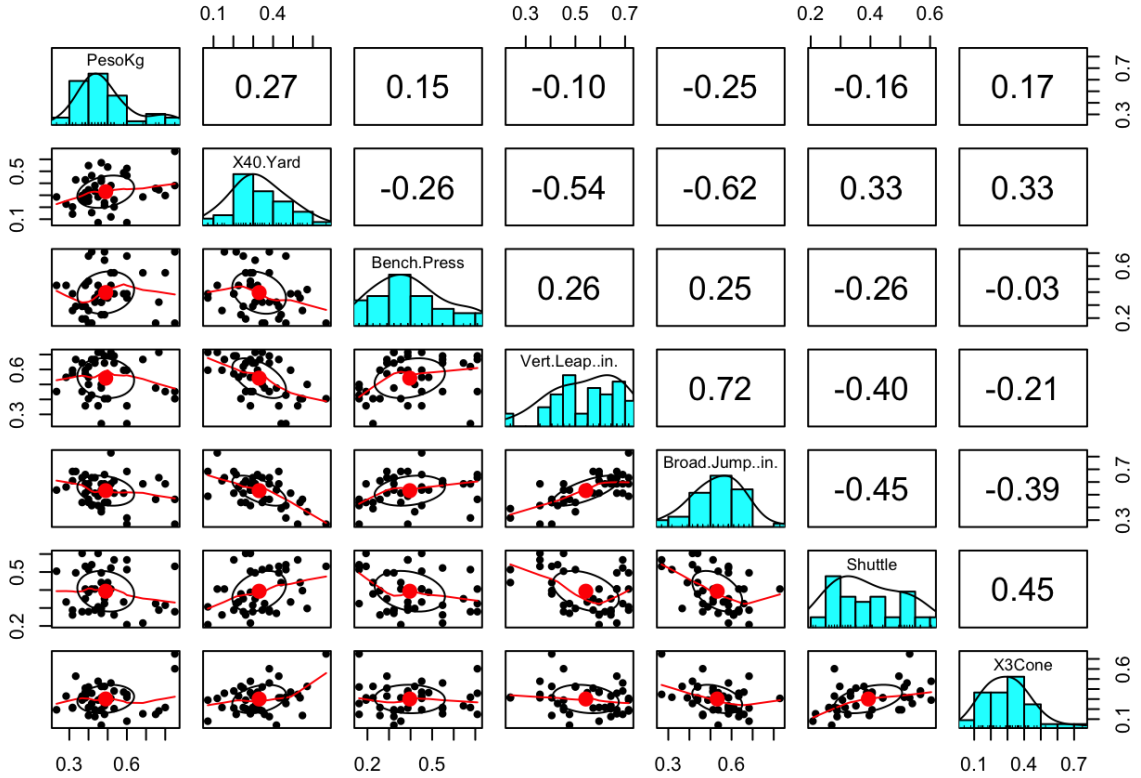
```
pairs.panels(mydata_trained_n_short[c("PesoKg", "X40.Yard","Bench.Press",
                "Vert.Leap..in.", "Broad.Jump..in.", "Shuttle", "X3Cone")])
```




**Pares de regresión lineal**

Analizando a los jugadores de la “elite” (aquellos rankeados como los mejores). Las gráficas ayudan a visualizar las tendencias de las variables comparadas en pares. Se identifica en los ovalos la tendencia de la regresión lineal, así como la distribuciión de los jugadores en el histogramo, adicionalmente se expresa el factor de correlación entre variable (COR), como ejemplo individual, se observa que el peso vs la velocidad es de 0.19

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



## ANEXO K

### Modelo k-nn, lazy y supervisado

#### K Nearest Neighbors

Técnica de aprendizaja “supervisado”, considerada como “lazy”, es la técnica de clasificación K-Nearest Neighbors que nos permite agrupar a elementos similares, como lo pueden ser los jugadores elite con sus propspectos más cercanos, esto ayuda a generar un modelo al cual se le alimente con datos de lo futuros resultados en los combinados y se le otorgue una probabilidad de ser exitoso y de conocer en que posición ILB o OLB tiene mayor posibilidades de destacar.

#### Preparación | transformación | normalización

Se crea la tabla de entrada llamada (mydata_trained),

a) Todos los campos deben estar completos, no faltantes o (NA),

b) Llevar a cabo la normalización de los datos, para homogenizar la muestra

c) Se segmenta en mydata_trained para (entrenamiento y pruebas)

Se logran obtener 42 registros entrenados con los jugadores elite.

#### Eliminar na

```
mydata_limpia_no_NA <- mydata_trained %>%
    filter(X3Cone != is.na(X3Cone)) %>%
    filter(X40.Yard != is.na(X40.Yard)) %>%
    filter(Bench.Press != is.na(Bench.Press)) %>%
    filter(Vert.Leap..in. != is.na(Vert.Leap..in.)) %>%
    filter(Broad.Jump..in. != is.na(Broad.Jump..in.)) %>%
    filter(Shuttle != is.na(Shuttle)) %>%
    select(Year, Name, College, POS, Height..in., PesoKg, X40.Yard, Bench.Pr
ess,Vert.Leap..in., Broad.Jump..in., Shuttle, X3Cone, Clasifica) %>%
    view()

  table(mydata_limpia_no_NA$Clasifica)
```

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

```
class(mydata_limpia_no_NA)
```

```
## [1] "data.frame"
```

```
str(mydata_limpia_no_NA)
```

```
## 'data.frame':    807 obs. of  13 variables:
##  $ Year           : int  1998 1999 1999 2000 2005 2005 2007 2011 2011 2012
##  $ Name           : chr  "Jeremiah Trotter" "Al Wilson" "Joey Porter"
##  $ College        : chr  "Stephen F. Austin (TX)" "Tennessee" "Colorado
##  $ POS            : chr  "ILB" "ILB" "OLB" "OLB" ...
##  $ Height..in.    : num  72.5 71.8 74.5 75.9 71.9 ...
##  $ PesoKg         : num  119 109 110 117 108 ...
##  $ X40.Yard       : num  4.68 4.56 4.68 4.61 4.84 4.6 4.53 4.8 4.42 4.46 #
#  $ Bench.Press    : int  32 17 27 27 23 27 22 22 21 24 ...
##  $ Vert.Leap..in. : num  34 33 39 34 35 38.5 39 32 37 39.5 ...
##  $ Broad.Jump..in.: int  117 116 122 122 113 122 119 109 126 132 ...
##  $ Shuttle        : num  4.14 4.25 4.41 4.18 4.16 4.07 4.46 4.46 4.06 4.28
##  $ X3Cone         : num  7.62 7.31 7.37 6.94 7.31 6.83 7.23 7.17 6.7 7.1
##  $ Clasifica      : Factor w/ 2 levels "Yes_Ranked","NO_Ranked": 1 1 1 1 1
```

***Normalizar, crear la función en R para normalizar***

Dado que las cifras expresadas en la recolección de datos es dispareja, es decir hay números muy grandes y otros muy pequeños, se decide transformar y aplicarles una normalización con la siguiente formula, eso permitirá dar precisión al estudio.

```
normalize <- function(x){
      return((x-min(x)) / (max(x) - min(x)))
      }
```

***Comprobación***

```
normalize(c(10,20,30,40,50))
```

```
## [1] 0.00 0.25 0.50 0.75 1.00
```

```
  normalize(c(1,2,3,4,5))
```

```
## [1] 0.00 0.25 0.50 0.75 1.00
```

**Normalizar**

```
mydata_trained_n <-mydata_limpia_no_NA %>%
        mutate(Height..in. = normalize(Height..in.)) %>%
```

```
        mutate(PesoKg = normalize(PesoKg)) %>%
        mutate(X40.Yard = normalize(X40.Yard)) %>%
        mutate(Bench.Press = normalize(Bench.Press)) %>%
        mutate(Vert.Leap..in. = normalize(Vert.Leap..in.)) %>%
        mutate(Broad.Jump..in. = normalize(Broad.Jump..in.)) %>%
        mutate(Shuttle = normalize(Shuttle)) %>%
        mutate(X3Cone = normalize(X3Cone))
str(mydata_trained_n[43:80,])
```

```
## 'data.frame':    38 obs. of  13 variables:
##  $ Year           : int  1997 1997 1997 1997 1997 1997 1997 1997 1997 1997
##  $ Name           : chr  "Lyron Cobbins" "John Fiala" "Jon Hesse" "Anthony
##  $ College        : chr  "Notre Dame" "Washington" "Nebraska" "Middle
##  $ POS            : chr  "ILB" "ILB" "ILB" "ILB" ...
##  $ Height..in.    : num  0.36 0.6 0.817 0.6 0.486 ...
##  $ PesoKg         : num  0.65 0.367 0.65 0.5 0.533 ...
##  $ X40.Yard       : num  0.571 0.643 0.69 0.738 0.655 ...
##  $ Bench.Press    : num  0.3871 0.3548 0.0968 0.5806 0.4839 ...
##  $ Vert.Leap..in. : num  0.381 0.31 0.5 0.452 0.262 ...
##  $ Broad.Jump..in.: num  0.463 0.268 0.341 0.39 0.22 ...
##  $ Shuttle        : num  0.505 0.351 0.73 0.459 0.631 ...
##  $ X3Cone         : num  0.591 0.466 0.67 0.665 0.801 ...
##  $ Clasifica      : Factor w/ 2 levels "Yes_Ranked","NO_Ranked": 2 2 2 2 2
```

```
mydata_trained_n_short <- mydata_trained_n%>%
  select(Height..in., PesoKg, X40.Yard, Bench.Press, Vert.Leap..in., Broad.J
ump..in., Shuttle, X3Cone)
str(mydata_trained_n_short)
```

```
## 'data.frame':    807 obs. of  8 variables:
##  $ Height..in.    : num  0.429 0.349 0.657 0.817 0.358 ...
##  $ PesoKg         : num  0.85 0.483 0.517 0.8 0.467 ...
##  $ X40.Yard       : num  0.381 0.238 0.381 0.298 0.571 ...
##  $ Bench.Press    : num  0.71 0.226 0.548 0.548 0.419 ...
##  $ Vert.Leap..in. : num  0.452 0.405 0.69 0.452 0.5 ...
##  $ Broad.Jump..in.: num  0.463 0.439 0.585 0.585 0.366 ...
##  $ Shuttle        : num  0.279 0.378 0.523 0.315 0.297 ...
##  $ X3Cone         : num  0.602 0.426 0.46 0.216 0.426 ...
```

```
str(mydata_trained_n)
```

```
## 'data.frame':    807 obs. of  13 variables:
##  $ Year           : int  1998 1999 1999 2000 2005 2005 2007 2011 2011 2012
##  $ Name           : chr  "Jeremiah Trotter" "Al Wilson" "Joey Porter"
##  $ College        : chr  "Stephen F. Austin (TX)" "Tennessee" "Colorado
##  $ POS            : chr  "ILB" "ILB" "OLB" "OLB" ...
##  $ Height..in.    : num  0.429 0.349 0.657 0.817 0.358 ...
##  $ PesoKg         : num  0.85 0.483 0.517 0.8 0.467 ...
##  $ X40.Yard       : num  0.381 0.238 0.381 0.298 0.571 ...
##  $ Bench.Press    : num  0.71 0.226 0.548 0.548 0.419 ...
##  $ Vert.Leap..in. : num  0.452 0.405 0.69 0.452 0.5 ...
```

```
##  $ Broad.Jump..in.: num  0.463 0.439 0.585 0.585 0.366 ...
##  $ Shuttle        : num  0.279 0.378 0.523 0.315 0.297 ...
##  $ X3Cone         : num  0.602 0.426 0.46 0.216 0.426 ...
##  $ Clasifica      : Factor w/ 2 levels "Yes_Ranked","NO_Ranked": 1 1 1 1 1
```

**Guardando campo calificador tipo factor (para pruebas de knn)**

**Pruebas nuevas del knn tratando de modificar las etiquetas**

mydata_trained_n, contiene toda la estructura limpia de espacios (NA) y cifras normalizadas como se apreciaa continuación

```
summary(mydata_trained_n)
```

```
##       Year          Name              College              POS
##  Min.   :1997   Length:807         Length:807         Length:807
##  1st Qu.:2006   Class :character   Class :character   Class :character
##  Median :2011   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2011
##  3rd Qu.:2016
##  Max.   :2022
##   Height..in.         PesoKg          X40.Yard       Bench.Press
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.4434   1st Qu.:0.3833   1st Qu.:0.3214   1st Qu.:0.2903
##  Median :0.5543   Median :0.5000   Median :0.4167   Median :0.3871
##  Mean   :0.5499   Mean   :0.4914   Mean   :0.4283   Mean   :0.3822
##  3rd Qu.:0.6571   3rd Qu.:0.6000   3rd Qu.:0.5238   3rd Qu.:0.4839
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  Vert.Leap..in.   Broad.Jump..in.     Shuttle          X3Cone
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.3571   1st Qu.:0.3415   1st Qu.:0.3514   1st Qu.:0.2557
##  Median :0.4524   Median :0.4390   Median :0.4414   Median :0.3409
##  Mean   :0.4474   Mean   :0.4484   Mean   :0.4493   Mean   :0.3551
##  3rd Qu.:0.5476   3rd Qu.:0.5366   3rd Qu.:0.5315   3rd Qu.:0.4375
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##       Clasifica
##  Yes_Ranked: 42
##  NO_Ranked :765
##
#
```

```
str(mydata_trained_n)
```

```
## 'data.frame':    807 obs. of  13 variables:
##  $ Year           : int  1998 1999 1999 2000 2005 2005 2007 2011 2011 2012
##  $ Name           : chr  "Jeremiah Trotter" "Al Wilson" "Joey Porter"
##  $ College        : chr  "Stephen F. Austin (TX)" "Tennessee" "Colorado
##  $ POS            : chr  "ILB" "ILB" "OLB" "OLB" ...
```

```
##  $ Height..in.    : num  0.429 0.349 0.657 0.817 0.358 ...
##  $ PesoKg         : num  0.85 0.483 0.517 0.8 0.467 ...
##  $ X40.Yard       : num  0.381 0.238 0.381 0.298 0.571 ...
##  $ Bench.Press    : num  0.71 0.226 0.548 0.548 0.419 ...
##  $ Vert.Leap..in. : num  0.452 0.405 0.69 0.452 0.5 ...
##  $ Broad.Jump..in.: num  0.463 0.439 0.585 0.585 0.366 ...
##  $ Shuttle        : num  0.279 0.378 0.523 0.315 0.297 ...
##  $ X3Cone         : num  0.602 0.426 0.46 0.216 0.426 ...
##  $ Clasifica      : Factor w/ 2 levels "Yes_Ranked","NO_Ranked": 1 1 1 1 1
```

***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]
```

Existen 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 supervisado**

***K Nearest Neighbors***

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

table(myPrediction)
```

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

```
table(mydata_train_n_labels)
```

```
## mydata_train_n_labels
## Yes_Ranked  NO_Ranked
##         30          0
```

**Comprobación del modelo k-nn, lazy y supervisado**

***K Nearest Neighbors***

Comparando el modelo que ha sido entrenado vs test: El modelo encontrando los 12 mejores jugadores correctamente (Yes_Ranked son embargo expresa que sólo esta 40% seguro de la redicción) también identifica el resto como NO rankeados, con el 60% de confianza Cabe mencionar que el modelo acertó al 100%..

Nota, el modelo de entrenamiento le entregó las muestras de los jugadores elite

```
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 |
## ---------------------|------------|------------|
##
```

## ANEXO L

## Árbol de Decisión

El “Árbol de Decisión” entrega la probabilidad de que el apoyador, sea probablemete exitoso como Interno ILB o externo OLB. Predice el resultado probabilistico del evento dado un camino seleccionado. Se han considerado lo siguiente, todos los jugadores apoyadores en el “Combinado Scout NFL” desde 1982 hasta 2022, y estudiado con el uso de las siguientes variables: X40.Yard + Vert.Leap..in.+ Bench.Press + X3Cone + PesoKg + Broad.Jump..in.

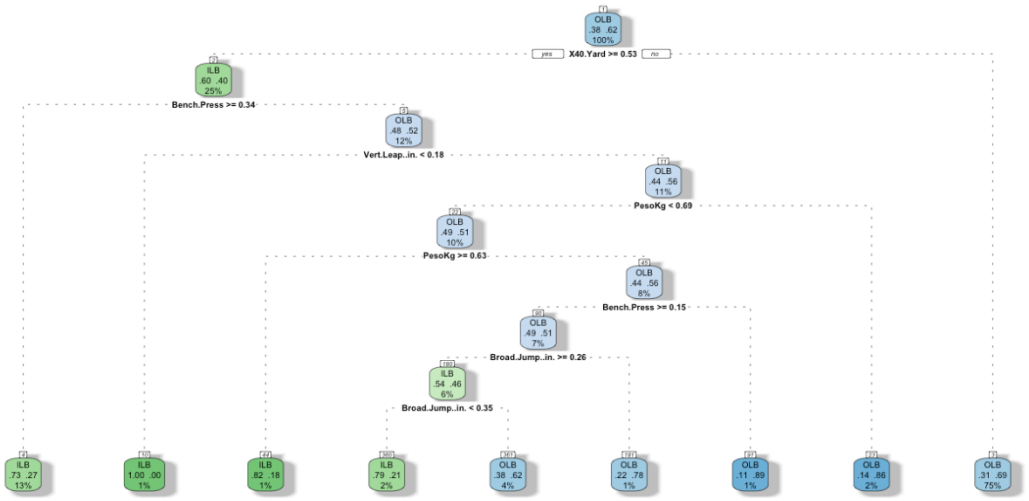
El modelo se entrena con 30 jugadores exitosos y una muestra combinada con 12 jugadores elite y el resto jugadores promedio

### 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)
```




Rattle 2022-Dec-11 21:33:25 mauricioarriaga

Se observa que el 62% de los participantes más rápidos, en la prueba de velocidad (40 yardas) son apoyadores externos (OLB), lo cual coincide en el desempeño esperado al cubrir pases y evitar las carreras por fuera de los tackles, en las llamadas “sweeps”, “rápidas” y pases al “hook” y al “flat”, tambien se observa en el siguiente nivel de la izquierda, que el apoyador ILB en un 60% de las veces es más fuerte, lo cual le ayudará a contra-bloquear a los linieros ofensivos, así como podrá romper bloqueos del fullback, bloqueos dobles y de trayectoria de trampa.  Es remarcable también que de aquellos jugadores en el tercer nivel, que siendo 38% más lentos que el OLB, un 40% de ellos, no son muy fuertes, dado ésta característica su tendencia es ser OLB

### Árbol de decisión

Predice el resultado con la probabilidad que pueda suceder el evento, dado un camino por analizar seleccionado. La matriz considera sólo a los jugadores de “elite” se anaizan las

siguientes variables: X40.Yard + Vert.Leap..in.+ Bench.Press + X3Cone + PesoKg + Broad.Jump..in.

La muestra de entrenamiento contiene a los 42 mejores jugadores con los datos normalizados.

***Preparación de los datos***

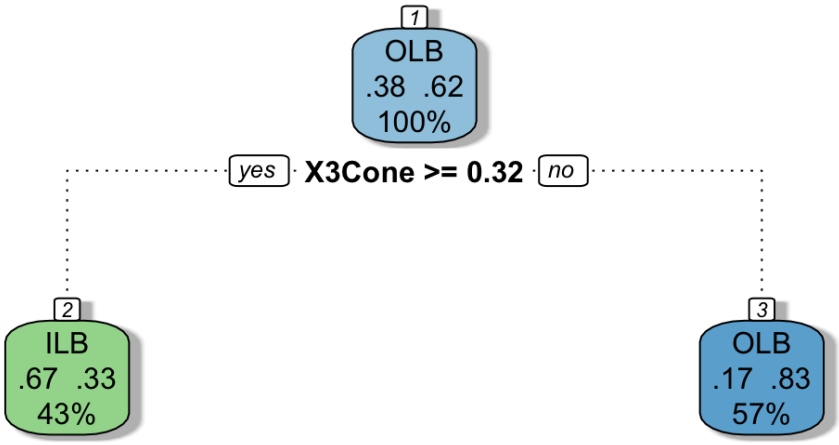
```
data1 <- mydata_trained_n
data1_trained <- data1[1:42,]
```

***Creación del modelo***

```
arbol <- rpart(
  formula = POS ~ X40.Yard + Vert.Leap..in.+ Bench.Press + X3Cone + PesoKg +
Broad.Jump..in. ,
  data = data1_trained,
  method = "class")
```

***Gráficando el Árbol***

```
fancyRpartPlot(arbol)
```




Rattle 2022-Dec-11 21:33:26 mauricioarriaga

Se identifica una (1) característica que es dominante : “X3Cone”

## ANEXO M

### Naive Bayes

Se extrae el campo clasificador y se guarda en diferentes archivos, serán usados en los modelos tanto de entrenamientos como en los de prueba

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

Se entrena el modelo con 30 jugadores de elite, identificados en la variable, Yes_Ranked, lo que implica que el evento dado es que han sido exitosos.

```
train2 <- mydata_trained_n[1:30,]
```

Se entrena los datos para probar contienen 12 jugadores de elite, el resto (18) no lo son.

```
test2 <- mydata_trained_n[31:60,]
```

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)
m
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = train2, y = mydata_train_n_labels, laplace = 0)
##
## A-priori probabilities:
## mydata_train_n_labels
## Yes_Ranked  NO_Ranked
##          1          0
##
## Conditional probabilities:
##                       Year
## mydata_train_n_labels     [,1]     [,2]
##            Yes_Ranked 2013.067 7.129024
##            NO_Ranked        NA       NA
##
##                       Name
## mydata_train_n_labels  Al Wilson Ben Niemann Bobby Wagner Brian Urlacher
##            Yes_Ranked 0.03333333  0.03333333   0.03333333     0.03333333
##            NO_Ranked
##                       Name
## mydata_train_n_labels Cory Littleton DeMarcus Ware Demario Davis
```

```
##            Yes_Ranked     0.03333333    0.03333333    0.03333333
##            NO_Ranked
##                       Name
## mydata_train_n_labels DeVondre Campbell Drue Tranquill   EJ Speed Ernest Jones
##            Yes_Ranked        0.03333333     0.03333333 0.03333333   0.03333333
##            NO_Ranked
##                       Name
## mydata_train_n_labels Fred Warner Greg Lloyd JaWhaun Bentley Jeremiah Trotter
##            Yes_Ranked  0.03333333 0.03333333      0.03333333       0.03333333
##            NO_Ranked
##                       Name
## mydata_train_n_labels Jerome Baker Joey Porter Jordan Hicks Josey Jewell
##            Yes_Ranked   0.03333333  0.03333333   0.03333333   0.03333333
##            NO_Ranked
##                       Name
## mydata_train_n_labels Kaden Elliss Kwon Alexander Lavonte David
##            Yes_Ranked   0.03333333     0.03333333    0.03333333
##            NO_Ranked
##                       Name
## mydata_train_n_labels Leighton Vander Esch Lofa Tatupu Luke Kuechly Nick Bolton
##            Yes_Ranked           0.03333333  0.03333333   0.03333333  0.03333333
##            NO_Ranked
##                       Name
## mydata_train_n_labels Patrick Willis Pete Werner Sione Takitaki Von Miller
##            Yes_Ranked     0.03333333  0.03333333     0.03333333 0.03333333
##            NO_Ranked
##
##                       College
## mydata_train_n_labels Arkansas State Boise State Boston College Brigham Young
##            Yes_Ranked     0.03333333  0.03333333     0.03333333    0.06666667
##            NO_Ranked
##                       College
## mydata_train_n_labels Colorado State Connecticut      Idaho       Iowa
##            Yes_Ranked     0.03333333  0.03333333 0.03333333 0.06666667
##            NO_Ranked
##                       College
## mydata_train_n_labels Louisiana State  Minnesota Mississippi   Missouri
##            Yes_Ranked      0.03333333 0.03333333  0.03333333 0.03333333
##            NO_Ranked
##                       College
## mydata_train_n_labels   Nebraska New Mexico Notre Dame Ohio State     Purdue
##            Yes_Ranked 0.03333333 0.03333333 0.03333333 0.06666667 0.03333333
##            NO_Ranked
##                       College
## mydata_train_n_labels South Carolina Southern California Stephen F. Austin (TX)
##            Yes_Ranked     0.03333333          0.03333333             0.03333333
##            NO_Ranked
##                       College
## mydata_train_n_labels Tarleton State  Tennessee      Texas  Texas A&M
##            Yes_Ranked     0.03333333 0.03333333 0.03333333 0.03333333
##            NO_Ranked
##                       College
## mydata_train_n_labels  Troy (AL) Utah State Washington
##            Yes_Ranked 0.03333333 0.03333333 0.03333333
##            NO_Ranked
```

```
##
##                      POS
## mydata_train_n_labels       ILB       OLB
##             Yes_Ranked 0.3333333 0.6666667
##             NO_Ranked
##
##                      Height..in.
## mydata_train_n_labels      [,1]      [,2]
##             Yes_Ranked 0.5915429 0.1600157
##             NO_Ranked         NA        NA
##
##                      PesoKg
## mydata_train_n_labels      [,1]      [,2]
##             Yes_Ranked 0.4866667 0.1461118
##             NO_Ranked         NA        NA
##
##                      X40.Yard
## mydata_train_n_labels      [,1]      [,2]
##             Yes_Ranked 0.3234127 0.1230081
##             NO_Ranked         NA        NA
##
##                      Bench.Press
## mydata_train_n_labels      [,1]      [,2]
##             Yes_Ranked 0.4150538 0.1524221
##             NO_Ranked         NA        NA
##
##                      Vert.Leap..in.
## mydata_train_n_labels      [,1]      [,2]
##             Yes_Ranked 0.5396825 0.1402796
##             NO_Ranked         NA        NA
##
##                      Broad.Jump..in.
## mydata_train_n_labels      [,1]      [,2]
##             Yes_Ranked 0.5373984 0.1236191
##             NO_Ranked         NA        NA
##
##                      Shuttle
## mydata_train_n_labels      [,1]      [,2]
##             Yes_Ranked 0.3975976 0.1197924
##             NO_Ranked         NA        NA
##
##                      X3Cone
## mydata_train_n_labels      [,1]      [,2]
##             Yes_Ranked 0.2912879 0.1360934
##             NO_Ranked         NA        NA
##
##                      Clasifica
## mydata_train_n_labels Yes_Ranked NO_Ranked
##             Yes_Ranked          1         0
##             NO_Ranked
```

***Se evalúa el 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 probabilidad # de ser 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 pruebas de los "Combinados Scouts NFL", así entregará la probabilidad de ser jugadores exitosos como profesionales.