FACULTAD DE ESTUDIOS ESTADÍSTICOS

GRADO EN ESTADISTICA APLICADA

Curso 2024/2025

Trabajo de Fin de Grado

TÍTULO: ESTUDIO PREDICTIVO DE LA SELECCIÓN DE RECEPTORES POR RONDA DEL NFL DRAFT

***Alumno*: DAVID THOMAS JACOBS MARITZIA**

***Tutor*: ADOLFO GÁLVEZ MORALEDA**

Febrero de 2025

UNIVERSIDAD COMPLUTENSE
MADRID

Índice

RESUMEN

En este estudio se pretende crear un modelo predictivo que estime la probabilidad de que un receptor universitario sea seleccionado en una de las siete rondas del draft de la NFL utilizando sus estadísticas universitarias y sus pruebas atléticas hechas en el NFL Combine.

Palabras clave: NFL, draft, ronda, yardas, touchdown, receptor, predicción.

ABSTRACT

In this study, the aim is to find a predictive model that estimates the probability of a college wide receiver being selected in one of the seven rounds of the NFL draft using their college stats and athletic tests conducted at the NFL Combine.

Keywords: NFL, draft, round, yards, touchdown, wide receiver. prediction.

1. INTRODUCCIÓN

El fútbol americano es un deporte de contacto derivado del rugby que consta de dos equipos que juegan once contra once y cuyo objetivo es obtener más puntos que el otro al final del partido.

El partido está dividido en cuatro cuartos de 15 minutos cada uno (60 minutos total) con descansos entre los cuartos. Entre el segundo y el tercer cuarto (la mitad) hay un descanso más largo donde los equipos se van a los vestuarios y se preparan para la segunda mitad del partido. Cada equipo obtiene 3 tiempos por mitad donde pueden parar el reloj de manera táctica. Finalmente, en el 2º y 4º cuarto hay lo que se denomina un "two-minute warning" (aviso de dos minutos) donde se pausa el partido cuando quedan dos minutos para terminar como si de un tiempo se tratase.

Los partidos son disputados en un campo de hierba rectangular como el de la ilustración 1. El ancho es de 53,33 yardas y su largo es de 120 yardas. Los denominados "end zones" son las áreas al final de cada lado del campo cuyas dimensiones son de 10 yardas y es donde se pueden anotar los touchdowns. Tenemos líneas discontinuas que marcan cada yarda, líneas continuas que son marcadas cada cinco yardas y los números que son puestos cada 10 yardas. Los números empiezan por el 10 e incrementan hasta el 50 que coincide con la mitad del campo, después vuelve a decrecer hasta 10. En el final de cada "end zone" existe un poste en forma de "Y".

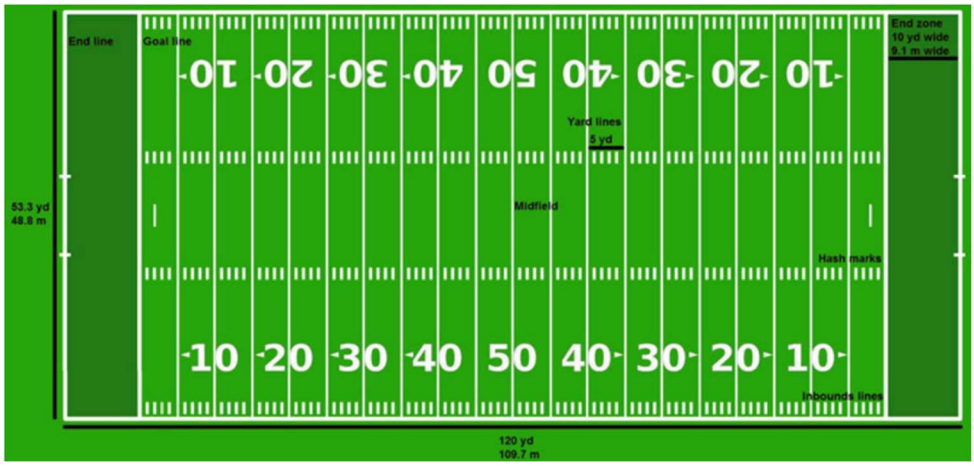

Ilustración 1.1: Campo de fútbol americano (Google User Content. (n.d.))

Los "kickoffs" (saques iniciales) se patean desde la mitad del campo (50) al inicio del partido, al inicio de la segunda mitad y tras alguna anotación.

De acuerdo con NFL Operations (n.d.a) las formas de anotar en el fútbol americano son las siguientes:

- Touchdown (6 puntos): un jugador marca un touchdown si, mientras tiene el balón en su poder, pasa la línea de anotación y entra en la end zone del otro equipo, o si atrapa el balón dentro del end zone del equipo rival.
- Field Goal (3 puntos): un equipo marca un field goal cuando un pateador patea el balón desde el campo de juego y lo hace pasar sobre el travesaño y entre los palos.
- Extra-Point (1 punto): después de anotar un touchdown, el equipo puede optar por que su pateador patee el balón sobre el travesaño y entre los palos —igual que en un Field Goal— para ganar un punto adicional. El balón se saca desde la línea de 15 yardas para el intento y la patada equivale a un gol de campo de aproximadamente 33 yardas.

- 2-point Conversion (2 puntos): después de anotar un touchdown, se puede optar por hacer una jugada desde la línea de dos yardas. El equipo gana los puntos si un corredor lleva el balón a través de la línea de anotación o lo atrapa dentro de la end zone, como ocurre al marcar un touchdown.
- Safety (2 puntos): Las defensas de la NFL pueden marcar un safety si taclean al jugador de la ofensiva que tiene el balón detrás de su propia línea de anotación o si lo hacen correr o soltar el balón fuera del campo detrás de su propia línea de anotación. También es "Safety" si los atacantes cometen una infracción en su propia end zone.

Los equipos de fútbol americano están divididos en tres secciones: offense, defense y special teams (atacantes, defensores y equipos especializados respectivamente). Los atacantes tienen como objetivo llevar el balón hasta el end zone del contrario para anotar un touchdown. Los ataques consisten en 4 "downs" que son el número de oportunidades que tiene la ofensiva en desplazarse 10 yardas. Si consiguen pasar las 10 yardas se vuelve al primer "down". La defensa tiene como objetivo intentar parar a el ataque de llegar a su "end zone".

Si la defensa consigue llegar a un cuarto down los atacantes tienen 3 opciones: La primera es arriesgar e intentar conseguir el primer down, si no lo consiguen el equipo rival comienza su ataque desde donde se terminó la jugada. La segunda opción es intentar un Field Goal; esto se hace cuando consideran que están lo suficientemente cerca del poste rival. Y finalmente si les paran cerca de su propio campo pueden hacer lo que se denomina un "punt". Un punt es cuando un jugador patea el balón lejos de su campo de forma que intenta que el rival empiece su ataque lo más alejado posible de su end zone.

Los equipos especializados son los encargados de los punts, field goals, extra points y kickoffs.

La NFL (National Football League) es la máxima categoría de fútbol americano profesional de los Estados Unidos. Esta liga contiene 32 equipos profesionales los cuales se dividen en 2 conferencias: la American Football Conference (AFC) y la National Football Conference (NFC). Los 16 equipos que pertenecen a cada conferencia se subdividen en 4 divisiones representadas por los puntos cardinales: Norte, Sur, Este, y Oeste obteniendo así 4 equipos por cada división.

La liga está estructurada en dos fases: la temporada regular (Regular Season) que está compuesto por 17 partidos con una semana de exención (en total 18 semanas). La forma en la que establecen los partidos es la siguiente, según NFL Operations (n.d.b.):

- Seis partidos contra contrincantes divisionales: dos partidos por equipo, uno de local y otro de visitante.
- Cuatro partidos contra equipos de una división dentro de su conferencia: dos partidos de local y dos de visitante.
- Cuatro partidos contra equipos de una división en la otra conferencia: dos partidos de local y dos de visitante.
- Dos partidos contra equipos de las dos divisiones restantes en su propia conferencia: un partido de local y uno de visitante (los enfrentamientos se basan en la clasificación de división de la temporada anterior).
- El partido 17 es un partido adicional contra un contrincante que no es de la conferencia de una división que el equipo no está programado para jugar (los enfrentamientos se basan en la clasificación de división de la temporada anterior)

Tras finalizar la temporada regular, 14 equipos (7 de cada conferencia) consiguen pasar a la segunda fase:

- Los campeones de cada división que se clasifican según su récord del 1 al 4 (al primero se le llama campeón de conferencia).
- Los tres equipos con mejor registro sin haber sido campeón de división siendo 5,6 y 7 respectivamente.

La segunda fase son los Playoffs: un torneo de eliminación basada en el siguiente formato:

Ilustración 1.2: Formato de los Playoffs de la NFL (Sporting news (n.d.))

Los Playoffs constan de cuatro rondas:

1) Wild Card: es la primera ronda de los playoffs. Se juegan 6 partidos (3 por conferencia) donde el segundo juega contra el séptimo, el tercero contra el sexto y el cuarto contra el quinto teniendo los campeones de división la ventaja de jugar localmente. Como podemos ver en la ilustración 2, los campeones de conferencia pasan de forma automática a la siguiente ronda.
2) Divisional: Llegados a la segunda ronda el campeón de conferencia juega contra el ganador del 4/5 y el ganador del 2/7 juega contra el ganador del 3/6. En el primer caso siempre se juega en la localidad del campeón de conferencia mientras que en el segundo partido juega en casa el ganador con mejor clasificación.
3) Conference Championship: En esta tercera ronda se decide el ganador de la conferencia. El partido se juega en casa del mejor clasificado.
4) Super Bowl. Llegamos a la final, donde juega el campeón de la NFC contra el campeón de la AFC. La localidad de este partido se decide años atrás (actualmente se sabe en qué estadios se van a jugar los siguientes 4 Super Bowls).

La Super Bowl es el evento deportivo más popular de los Estados Unidos, en febrero de 2024 la Super Bowl LVIII se convirtió en la transmisión televisiva más vista de la historia con 123,4 millones de espectadores promedio en todas las plataformas. (NFL Operations, n.d.c.)

En la NFL no existe la relegación, son los mismos 32 equipos todos los años. La forma en la que compensan este hecho es a través del "NFL Draft". El draft de la NFL es un evento en el que los distintos equipos escogen a los mayores talentos universitarios para así mejorar el nivel y las oportunidades del equipo. Cada equipo recibe una elección en cada una de las 7 rondas del draft.

El orden de selección es determinado por el orden invertido de la clasificación final de la temporada anterior. Es decir, el equipo con peor registro es el primero en escoger.

Las selecciones del 1 al 18 son para los equipos que no clasificaron a los playoffs, del 19 al 24 son para los equipos que perdieron en el Wild Card, del 25 al 28 para para los perdedores del Division,

el 29 y 30 para los perdedores de Conference Championship, el 31 y 32 es para el perdedor y ganador de la Super Bowl respectivamente.

En situaciones en las que los equipos terminaron la temporada anterior con registros idénticos, la determinación de la posición del Draft se decide por la dificultad del calendario según el porcentaje total ganador de los oponentes de un equipo. El equipo que jugó el calendario con el porcentaje ganador más bajo tendrá derecho a elegir antes. Si los equipos tienen la misma dificultad de calendario, se aplican sus registros contra oponentes comunes en su división o conferencia, si corresponde. Si no corresponden desempates de división o conferencia, se desempatará mediante el lanzamiento de una moneda. (NFL Operations, n.d.d.)

Si los equipos pierden jugadores en la Agencia Libre, la NFL puede asignar lo que se llama "Compensatory pick" en el que le asignan una elección para cubrir la vacante, las elecciones asignadas se llevan a cabo al final de la ronda tres y sigue hasta el final de la séptima. La ronda de los compensatory picks se determinan con una fórmula patentada desarrollada por el Consejo de Administración de la NFL, que considera el salario, el tiempo de juego y los honores de postemporada de un jugador. Se suma el valor de los agentes libres suplementarios que cada equipo obtiene o pierde, y se le conceden elecciones a un equipo de igual valor a la pérdida neta de los agentes libres suplementarios, hasta un máximo de cuatro. Todas las selecciones menos las compensatorias son negociables y comercializables ya sea por un jugador de la NFL o por otra selección del draft.

En este trabajo nos centraremos en los talentos seleccionados que pertenecen a la posición de receptores (Wide Receiver) cuyo objetivo en el campo es capturar el balón en el aire que reciben del quarterback (mariscal) e intentar correr campo abajo.

Los equipos le dan una importancia máxima al draft porque es la forma menos costosa y más eficiente de conseguir talento y futuras estrellas. Debido a esto han creado el NFL Combine eventos para medir las habilidades físicas y mentales de los universitarios. En el NFL Combine los jugadores universitarios van a Indianápolis y hacen las distintas pruebas físicas y mentales enfrente de todos los ojeadores, entrenadores y los directores generales de los 32 equipos.

El interés de este trabajo se basa en estudiar qué factores son considerados los más influyentes en cuanto a la estrategia de reclutamiento de los equipos. Se quiere aprender de qué forma afecta el NFL Combine y las estadísticas universitarias a la ronda en la que es seleccionada un receptor, ya que, el objetivo primordial del draft es predecir qué jugadores van a tener una carrera exitosa y formar parte del equipo a largo plazo. Este trabajo se puede considerar un primer paso hacia el estudio de la identificación de patrones, cuya finalidad es obtener una mejora en la evaluación y priorización de futuros talentos.

2. OBJETIVOS

El objetivo principal de este trabajo es lograr construir un modelo predictivo estadístico que estime la probabilidad de que un receptor universitario sea seleccionado en una de las siete rondas del draft de la NFL.

La forma de llegar a el objetivo principal de forma adecuada es a través de una estructura que nos asegure que los pasos que se toman son los correctos. Las etapas definidas para alcanzar el objetivo son las siguientes:

- Un marco teórico: en él se explicarán las diferentes técnicas estadísticas utilizadas junto con todos los conceptos necesarios para la interpretación adecuada de este trabajo.
- El conjunto de datos: la forma en la que se ha creado la base de datos y sus variables.
- La correcta depuración de datos: nos centraremos en el tratamiento de los valores perdidos, en un análisis descriptivo de las variables y finalmente en la creación de variables.
- Multicolinealidad: Se estudiará la posibilidad de encontrarnos con problemas de multicolinealidad en el conjunto de datos, la forma de detectarlo y la forma en la que se soluciona.
- Creación de los modelos: Se crearán 4 modelos basados en dos aplicaciones de Machine Learning distintos: la regresión logística multinomial y la regresión logística ordinal. El primer modelo se creará manualmente mientras que el segundo será construido a través de los métodos automáticos de selección de variables. Se analizarán, interpretarán, evaluarán y se seleccionará el óptimo. En este trabajo el nivel de significación será del 5%.
- Conclusión: Se mostrarán las conclusiones de este trabajo y unas propuestas de mejora.

La base de datos ha sido creada con Excel. Todos los análisis han sido realizados con la ayuda de Rstudio.

3. MARCO TEÓRICO

3.1. TRATAMIENTO DE VALORES PERDIDOS

"**Tipos de datos perdidos**
Aquí hay tres tipos generales de datos perdidos:

1. **Missing Completely at Random (MCAR):** Son datos perdidos completamente aleatorios, es decir, la ausencia de valores en un conjunto de datos es independiente tanto de otras variables observadas como de las no observadas.
2. **Missing at Random (MAR):** Son datos perdidos aleatorios, es decir, la probabilidad de que un valor falte puede depender de otras variables observadas, pero no de las no observadas.

A pesar de que la ausencia no es completamente aleatoria, los datos MAR son más manejables que los datos "Missing Not at Random" (MNAR), ya que, en teoría, se puede modelar y abordar la falta de manera sistemática si se conocen las variables relacionadas con la ausencia. Este es un tipo común de datos faltantes.

3. **Missing not at Random (MNAR):** Datos perdidos de forma no aleatoria, es decir, la probabilidad de que falte un valor depende de la variable no observada. Esto puede introducir sesgos en los datos y es más complicado de manejar. En otras palabras, la falta de un valor está relacionada con la información que falta y no puede ser modelada únicamente en función de las variables observadas.

Tratar con datos **MNAR** (Missing not at Random) puede ser más complicado que tratar con datos **MCAR** (Missing Completely at Random) o **MAR** (Missing at Random) porque la falta de información está relacionada con la información que falta." (Pineda, 2024)

"El algoritmo k vecinos más cercanos (KNN) es un clasificador de aprendizaje supervisado no paramétrico que utiliza la proximidad para hacer clasificaciones o predicciones sobre la agrupación de un punto de datos individual. Los conjuntos de datos tienen valores faltantes con frecuencia, pero el algoritmo KNN puede estimarlos en un proceso conocido como imputación de datos faltantes." (IBM, n.d.a.)

"knnImputation: Función que rellena todos los NA usando los k vecinos más cercanos para cada caso de NA. Por defecto usa los valores de los vecinos y obtiene una media ponderada (de la distancia al caso) de los valores para rellenar los datos faltantes. Si el método es a través de la mediana utiliza la mediana o el valor más frecuente. K usa 10 vecinos por defecto." (R Documentation, n.d.)

3.2. MULTICOLINEALIDAD

La multicolinealidad es un problema que se presenta de forma común en los modelos de regresión donde cada una o varias variables independientes son combinación lineal de otras. La multicolinealidad causa que los estimadores tengan varianzas muy altas e inestables. Esto hace que distintas muestras nos lleven a modelos muy diferentes.

Las formas en las que se determina si la multicolinealidad está presente es a través de la matriz de correlación y el cálculo del Factor de Inflación de Varianza (VIF).

El Factor de Inflación de Varianza es un indicador de multicolinealidad específica para cada variable independiente. Como norma general si el VIF>10 entonces existe multicolinealidad.

$$VIF = \frac{1}{1 - R_j^2}$$

En el caso de que el modelo presente multicolinealidad. La variable independiente con mayor VIF será excluida del modelo. Una vez se construya este nuevo modelo se calculará su VIF para ver si sigue existiendo multicolinealidad.

La matriz de correlación tiene como regla general que una correlación absoluta mayor a 0,8 se considera problemático. (Calviño, A. 2023). Estas matrices se emplearán como herramienta de visualización para secundar la existencia de la multicolinealidad.

3.3. REGRESIÓN LOGÍSTICA MULTINOMIAL

La regresión logística multinomial es un algoritmo de Machine Learning utilizado para predecir la probabilidad de una variable dependiente cualitativa con más de dos categorías. Se asume que la variable dependiente tiene un número finito de categorías que son mutuamente excluyentes y tiene una distribución multinomial cuyo primer parámetro es igual a uno.

Hay que modelizar la esperanza de la distribución de Y condicionada a los valores de las variables independientes. La esperanza coincide con las probabilidades de cada categoría.

$$P(Y = j|X_1, \dots, X_m) = \frac{e^{(\beta 0,j+\beta 1,j*X1+\cdots+\beta m,j*Xm)}}{1+\sum_{i=2}^{k} e^{(\beta 0,i+\beta 1,i*X1+\cdots+\beta m,i*Xm)}}, \forall \mathrm{j} = 2, \dots, \mathrm{K}$$

$$P(Y = 1|X_1, \dots, X_m) = \frac{1}{1+\sum_{i=2}^{k} e^{(\beta 0,i+\beta 1,i*X1+\cdots+\beta m,i*Xm)}}$$

Esta formulación proporciona K-1 conjuntos de parámetros β para todas las categorías excepto la primera (la de referencia). Se plantean cocientes de las probabilidades del modelo obtenido.

$$\frac{P(Y = j|X_1, \dots, X_m)}{P(Y = 1|X_1, \dots, X_m)} = e^{(\beta 0,j+\beta 1,j*X1+\cdots+\beta m,j*Xm)}, \forall \mathrm{j} = 2, \dots, \mathrm{K}$$

Esto permite una mejor interpretación de los parámetros y de los efectos de las variables dependientes sobre la independiente. Los parámetros van asociados a la comparación de la probabilidad de la categoría de referencia.

Una vez modelizado se ha de obtener la estimación de los parámetros, la cual está basada en el método de máxima verosimilitud que viene dada por:

$$L(\beta) = \prod_{i=1}^{n} p_{1i}^{y1i} \, p_{2i}^{y2i} \, \dots \, p_{Ki}^{yKi} \text{ con}$$

$$p_{ji} = P(Y = j|x_{1i}, \dots, x_{mi}), \forall \mathrm{i} = 1, \dots, \mathrm{n};\ \mathrm{j} = 1, \dots, \mathrm{K}$$

$$y_{ji} = \begin{cases} 1 \; si \; y_i = j \\ 0 \; en \; otro \; caso \end{cases}, \forall \mathrm{i} = 1, \dots, \mathrm{n};\ \mathrm{j} = 1, \dots, \mathrm{K}$$

Al no existir una formula explícita para obtener los parámetros se recurrirá a métodos iterativos de optimización. Una vez obtenidas las probabilidades de cada categoría se asignarán las observaciones a la categoría con mayor probabilidad predicha.

3.3.1. INTERPRETACIÓN DE LOS PARÁMETROS

Los parámetros serán interpretados a través de los odds-ratio y se asumirá que el resto de las variables se mantienen constantes. Se debe tener en cuenta si la variable independiente es cualitativa o cuantitativa ya que sus odds-ratio no se calculan de la misma manera.

"Para interpretar correctamente el parámetro de una variable regresora dicotómica debemos tener en cuenta que la inclusión de variables cualitativas se lleva a cabo a partir de la creación de variables dummy." (Calviño, A. 2023)

Se calculan los odds de los individuos para los que x=1 y para los que x=0:

$$odds(evento|x=1) = \frac{P(evento|x=1)}{1-P(evento|x=1)} = \frac{\frac{e^{(\beta_0+\beta_1)}}{1+e^{(\beta_0+\beta_1)}}}{\frac{1}{1+e^{(\beta_0+\beta_1)}}} = e^{\beta_0+\beta_1}$$

$$odds(evento|x=0) = \frac{P(evento|x=0)}{1-P(evento|x=0)} = \frac{\frac{e^{\beta_0}}{1+e^{\beta_0}}}{\frac{1}{1+e^{\beta_0}}} = e^{\beta_0}$$

El odds-ratio es el cociente de los odds mencionados anteriormente.

$$OR = \frac{odds(evento|x=1)}{odds(evento|x=0)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

La interpretación del odds-ratio si la variable independiente es cualitativa es la siguiente: "Si el parámetro está asociado a un nivel de una variable cualitativa, $e^{\beta_{i,j}}$ se interpreta como el aumento/dismunición que se observa en las posibilidades de que se dé la categoría j frente a la de referencia (la denotada como 1) cuando la variable explicativa analizada pasa de su nivel de referencia al i." (Calviño, A. 2023)

"En el caso de variables regresoras continuas, la exponencial del parámetro coincide con el odds ratio asociado al efecto de un incremento unitario en la variable" (Calviño, A. 2023)

$$OR = \frac{odds(evento|x=a+1)}{odds(evento|x=a)} = \frac{e^{\beta_0+\beta_1(a+1)}}{e^{\beta_0+\beta_1 a}} = e^{\beta_1}$$

La interpretación del odds-ratio si la variable independiente es cuantitativa es la siguiente: "Si el parámetro está asociado a una variable cuantitativa, $e^{\beta_{i,j}}$ se interpreta como el efecto multiplicativo que tiene un incremento unitario de la variable i en las posibilidades de que se dé la categoría j frente a la de referencia (la denotada como 1)." (Calviño, A. 2023)

3.3.2. ANÁLISIS DEL MODELO

En este trabajo se utilizará "stargazer" para los contrastes de los parámetros y ANOVA de tipo II para el análisis de tipo II en R. Stargazer es un paquete de R que se ha diseñado para la creación de

tablas de regresión donde se presentan los coeficientes, errores estándar y p-valores de los parámetros individuales.

Los contrastes de los parámetros permiten determinar si cada uno de los parámetros del modelo son significativamente distintos de 0 utilizando el "test t" debido a que son obtenidos a través de la máxima verosimilitud.

El análisis de tipo II contrasta la contribución que tienen las variables independientes en el modelo, lo que da como resultado unas variables conjuntamente significativas. La forma en la que se mide la significación es calculando la disminución de la verosimilitud del modelo debido a la eliminación de cada una de las variables independientes. Las comparaciones son realizadas a través del contraste de razón de verosimilitudes.

3.3.3. EVALUACIÓN DEL MODELO

Primero se empezará creando una matriz de confusión. Una matriz de confusión en Machine Learning es una matriz que tiene como objetivo visualizar el rendimiento de un modelo de clasificación.

La matriz de confusión que debe de contar con tantas filas y columnas como categorías tenga la variable dependiente. La diagonal principal representa los aciertos ya que cada $n_{i,j}$ muestra las observaciones que pertenecen a la categoría i, pero se predicen como j.

$$\begin{pmatrix} n_{11} & \cdots & n_{1j} \\ \vdots & \ddots & \vdots \\ n_{i1} & \cdots & n_{ii} \end{pmatrix}$$

A partir de la matriz de confusión se sacará la tasa de acierto del modelo y el índice de Kappa.

La tasa de acierto del modelo se calcula de la siguiente manera:

$$acc = \frac{\sum_{i=1}^{k} n_{ii}}{n}$$

El índice de Kappa es una alternativa a la tasa de acierto que reduce la influencia de la frecuencia relativa del evento, para lo cual se elimina el efecto de los aciertos que se pueden producir al azar.

$Kappa = \frac{acc - \sum_{i=1}^{k} p_{i.} * p_{.i}}{1 - \sum_{i=1}^{k} p_{i.} * p_{.i}}$, $p_{i.}$ y $p_{.i}$ son las proporciones de observaciones en la categoría i real y predicha

La interpretación del índice de Kappa se muestra en la siguiente tabla:

Índice Kappa	Interpretación
0.00	Equivalente al azar
0.01-0.20	Pobre
0.21-0,40	Justo
0.41-0.60	Moderado
0.61-0.80	Bueno
0.81-1	Excelente

Tabla 3.1

La sensibilidad y especificidad en los modelos multinomiales se calculan de manera individual para cada categoría, representando la capacidad que tiene el modelo de capturar la categoría del estudio y la ausencia de esta.

$$Sensibilidad_i = \frac{n_{ii}}{n_{i.}} \qquad Especificidad_i = \frac{n - \sum_{\substack{j=1 \\ j \neq i}}^{k} n_{ji}}{n - n_{i.}}$$

Las curvas ROC (Característica Operativa del Receptor) es una representación gráfica del rendimiento del clasificador que muestra la distribución de las fracciones de sensibilidad y 1-especificidad. En el caso multinomial se adapta esta curva de dos maneras distintas: la curva ROC pairwise y la Curva ROC one vs all.

La curva ROC pairwise obtiene tantas curvas ROC como pares de categorías exista y tiene el objetivo de ver si el modelo es capaz de distinguir entre las categorías del par. Por último, se obtiene la media como medida de evaluación.

La curva ROC one vs all obtiene tantas curvas como categorías tiene la variable dependiente. Una vez obtenido todas las curvas se calcula la media como medida de evaluación. Se pueden representan gráficamente las curvas para determinar cómo detecta el modelo cada categoría.

3.4. REGRESIÓN LOGÍSTICA ORDINAL

La regresión logística ordinal es un algoritmo de Machine Learning similar a la multinomial con la diferencia clave de que las categorías de la variable dependiente pueden ser ordenados de forma escalada. Al igual que en la multinomial se asume que la variable dependiente tiene un número finito de categorías que son mutuamente excluyentes, pero tiene una distribución logística.

El modelo recurre a valores numéricos para la predicción, cada categoría es asignada un valor de manera creciente y siguiendo el orden natural. Asume que se dispone de una variable latente cuantitativa a partir de la cual se selecciona el valor de la dependiente.

“La variable latente se construye a partir de una combinación lineal de las “m” características de los decisores (con la previa transformación de las variables cualitativas a dummies):” (Calviño, A. 2023)

$$\beta_1 X_1 + \beta_2 X_2 \ + \cdots + \beta_m X_m + \epsilon$$

Lo que equivale a que la variable dependiente Y sea:

$$Y = j, si\ \alpha_{j-1} < \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m + \epsilon \leq \alpha_j,$$
$$j = 1, \dots, K\ con\ \alpha_0 = -\infty\ y\ \alpha_K = \infty$$

La probabilidad de cada alternativa se puede calcular como:

$$p_{ji} = P(Y = j|x_{1i}, \dots, x_{mi}) = P\left(\alpha_{j-1} < \beta_1 x_1 + \cdots + \beta_m x_{mi} + \epsilon_i \leq \alpha_j\right)$$
$$= \cdots =$$
$$= \frac{e^{\alpha_j - (\beta_1 x_{1i} + \cdots + \beta_m x_{mi})}}{1 + e^{\alpha_j - (\beta_1 x_{1i} + \cdots + \beta_m x_{mi})}} - \frac{e^{\alpha_{j-1} - (\beta_1 x_{1i} + \cdots + \beta_m x_{mi})}}{1 + e^{\alpha_{j-1} - (\beta_1 x_{1i} + \cdots + \beta_m x_{mi})}}$$

Una vez obtenidas las probabilidades se clasifican con la probabilidad predicha máxima (al igual que en la regresión logística multinomial)

La estimación de los parámetros se basa en el método de máxima verosimilitud y es calculado de la misma forma que en el caso del modelo logístico multinomial.

3.4.1. INTERPRETACIÓN DE LOS PARÁMETROS

Los parámetros serán interpretados a través de los odds-ratio acumulados debido a que las categorías están relacionadas por el orden.

$$odds(Y > j|x_{1i}, \dots, x_{mi}) = \frac{P(Y > j|\ x_{1i}, \dots, x_{mi})}{P(Y \leq j|\ x_{1i}, \dots, x_{mi})}$$
$$= \frac{\frac{1}{1 + e^{\alpha_j - \beta_1 x_{1i} - \cdots - \beta_m x_{mi}}}}{\frac{e^{\alpha_j - \beta_1 x_{1i} - \cdots - \beta_m x_{mi}}}{1 + e^{\alpha_j - \beta_1 x_{1i} - \cdots - \beta_m x_{mi}}}} = \frac{1}{e^{\alpha_j - \beta_1 x_{1i} - \cdots - \beta_m x_{mi}}}$$
$$= e^{-\alpha_j + \beta_1 x_{1i} + \cdots + \beta_m x_{mi}}$$

3.4.2. ANÁLISIS DEL MODELO

El modelo se analizará de la misma manera que en el caso de la regresión logística multinomial.

3.4.3. EVALUACIÓN DEL MODELO

En cuanto a la evaluación del modelo logístico ordinal se utilizarán la matriz de confusión, la tasa de precisión, el índice de Kappa, la sensibilidad y la especificidad. Ninguno de estos evaluadores tiene en cuenta el carácter ordinal de la variable por lo que nuestro evaluador clave será el índice de Kappa ponderado.

El índice de Kappa ponderado es una modificación del índice de Kappa que da mayores pesos a los elementos que están más cercanos a la diagonal principal. Los pesos tomarán el valor de 1 en la diagonal principal y 0 en el mayor error.

$$Kappa_p = \frac{\sum_{i=1}^{K}\sum_{j=1}^{K} w_{ij*}p_{ij} - \sum_{i=1}^{K}\sum_{j=1}^{K} w_{ij*}p_{i.*}p_{.j}}{1 + \sum_{i=1}^{K}\sum_{j=1}^{K} w_{ij*}p_{i.*}p_{.j}}$$

Donde w_{ij} son los pesos

La interpretación del índice de Kappa ponderado es la misma que la del índice de Kappa.

3.5. MÉTODOS AUTOMATICOS DE SELECCIÓN DE VARIABLES

Probar todas las combinaciones de variables independientes para averiguar cuál es el mejor modelo es poco efectivo y práctico. En este trabajo se utilizará tres métodos de selección de variables muy comunes: Backwards, Forwards y Stepwise combinados con dos medidas de evaluación: AIC y BIC.

Backwards: Parte de un modelo con todas las variables y va eliminando una a una las variables que menos influyen en el modelo hasta que eliminar alguna de las variables restantes empeore la calidad del modelo. Una vez se elimina una variable no puede volver a entrar.

Forward: Parte desde cero y va añadiendo uno a uno las variables independientes con mayor influencia en el modelo hasta que añadir alguna de las variables no aporte información. Una vez se mete una variable no puede salir.

Stepwise: Mezcla backwards y forwards. Empieza igual que el método forward, pero en este caso si que se pueden eliminar las variables una vez entran al modelo basándose en el método backward. Normalmente se incluye un máximo de iteraciones para prevenir que emerjan bucles.

AIC: Akaike Information Criterion $AIC = -2 * \ln(L) + 2\tau$

BIC: Bayesian Information Criterion $BIC = -2 * \ln(L) + \tau * \ln(n)$

Donde "L" es el máximo valor de la función de verosimilitud y τ es la penalización del número de parámetro. BIC penaliza más que AIC por lo que cuenta con menos parámetros.

4. CONJUNTO DE DATOS

El conjunto de datos ha sido creado en Excel. Las variables asociadas a los jugadores, universidades a las que han ido y las pruebas del NFL Combine han sido sacadas de las siguientes bases de datos en abierto: NFL combine results (n.d.), Pro Football Reference (n.d.), y la propia página web de la NFL (n.d.). Para las variables asociadas con sus estadísticas universitarias se ha utilizado la base de datos abierta de Sports Reference CFB (n.d.) y los datos dados de las propias páginas webs universitarias. La base cuenta con 550 receptores de la NFL que se presentaron al Draft entre 2010 y 2024.

5. DEPURACIÓN DE DATOS

5.1. ANÁLISIS DESCRIPTIVO DE LAS VARIABLES

El primer paso que se va a dar para conocer de forma exhaustiva el conjunto de datos que se empleará en este trabajo es realizar un análisis descriptivo de todas las variables incluidas en la base de datos, las variables creadas tendrán su análisis descriptivo realizado en el 5.2.

Player: Variable cualitativa nominal que hace referencia al nombre del jugador. Esta es la variable identificadora.

School: Variable cualitativa nominal que hace referencia a la universidad que ha atendido el jugador el año que se presenta al draft. Hay 138 universidades distintas por lo que esta variable servirá para crear otras dos variables que serán utilizadas en los análisis.

Ht: Variable cuantitativa que hace referencia a la altura del jugador en pulgadas.

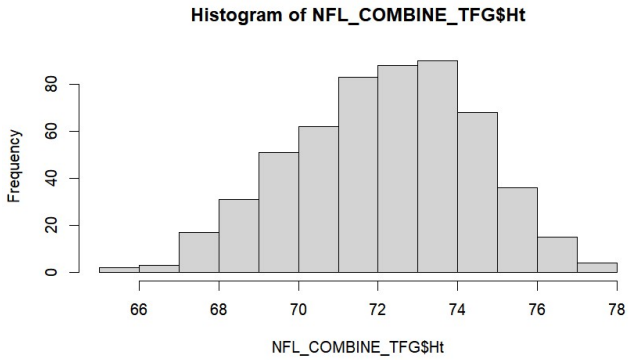

Figura 5.1

Min.	65.00
1st Qu.	71.00
Median	73.00
Mean	72.66
3rd Qu.	74.00
Max.	78.00

Tabla 5.1

Las alturas de los jugadores varían desde 65 pulgadas a 78 pulgadas, una diferencia de 13 pulgadas (unos 32 centímetros) lo cual es una diferencia considerable. Se puede ver que la altura media es de 72,66 pulgadas.

Wt: Variable cuantitativa que hace referencia a el peso en libras del jugador.

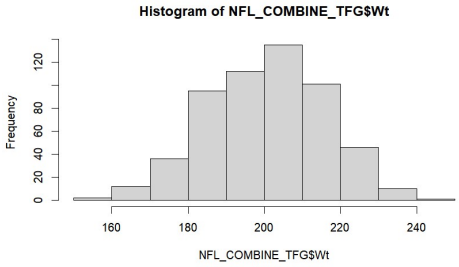

Figura 5.2

Min.	155.0
1st Qu.	190.0
Median	201.5
Mean	201.0
3rd Qu.	212.0
Max.	243.0

Tabla 5.2

Los pesos de los jugadores varían desde 155 libras a 243 libras, una diferencia de 88 libras (unos 40 kilogramos) lo cual es una diferencia considerable. Se puede ver que el peso medio es de 201 libras.

Fourty_yd: Variable cuantitativa que hace referencia a los segundos que tarda un jugador en correr cuarenta yardas. Esta prueba tiene como objetivo medir la velocidad del jugador.

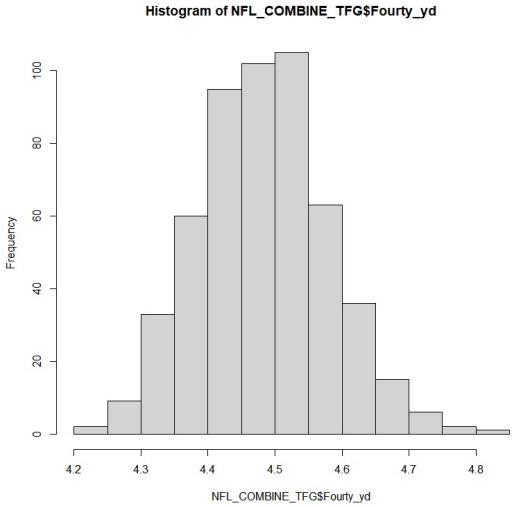

Figura 5.3

Min.	4.210
1st Qu.	4.420
Median	4.490
Mean	4.488
3rd Qu.	4.550
Max.	4.840
NA´s	21

Tabla 5.3

Los tiempos de los jugadores varían desde 4,21 segundos a 4,84 segundos, una diferencia de 0,63 segundos. La media de tiempos es de 4,488 segundos. Esta variable cuenta con 21 valores faltantes.

Al tener una diferencia de menos de un segundo, se cambiará la unidad de medida de segundos a centésimas de segundo para facilitar la interpretación de los modelos en un futuro. Este cambio será llevado a cabo en todas las variables cuya unidad de medida sean segundos.

Bench: Variable cuantitativa que hacer referencia al número de repeticiones que realiza un jugador con 225 libras (unos 102 kg) en press de banca. Esta prueba tiene como objetivo evaluar la fuerza del tren superior.

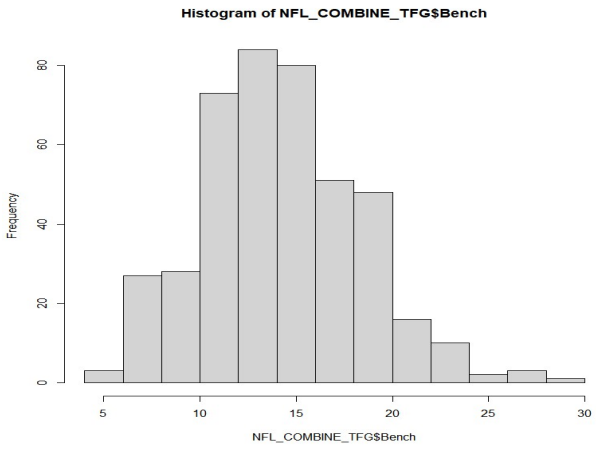

Figura 5.4

Min.	4.00
1st Qu.	12.00
Median	14.00
Mean	14.69
3rd Qu.	17.00
Max.	29.00
NA´s	124

Tabla 5.4

El número de repeticiones varia de 4 repeticiones a 29, una diferencia de 25 repeticiones lo cual es una discrepancia considerable. La media de repeticiones es de 14,69 y cuenta con 124 valores faltantes.

Vertical: Variable cuantitativa: El jugador estira los brazos hacia arriba para usar su alcance como referencia. Tras ello, desde una posición base (sin carrerilla alguna) el jugador salta lo más alto posible. La diferencia se mide en pulgadas. La prueba tiene como objetivo evaluar la explosión y la potencia del tren inferior.

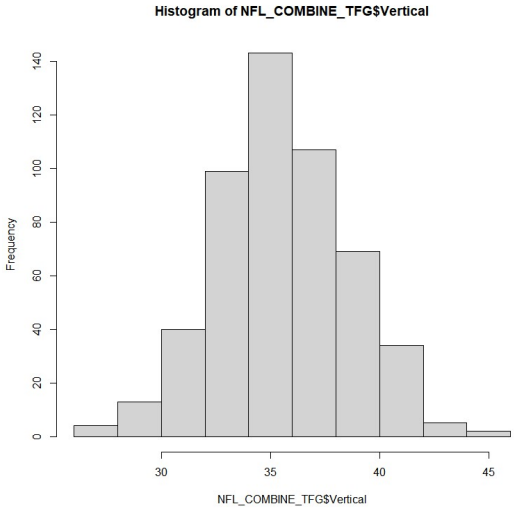

Figura 5.5

Min.	27.0
1st Qu.	33.5
Median	36.0
Mean	35.8
3rd Qu.	38.0
Max.	45.00
NA´s	34

Tabla 5.5

Los Saltos de los jugadores varían desde 27 pulgadas a 45 pulgadas, una diferencia de 18 pulgadas (unos 46 centímetros) lo cual es una diferencia considerable. Se puede ver que el salto medio es de 35,8 pulgadas y que la variable cuenta con 34 valores faltantes.

Broad_Jump: Variable cuantitativa que mide la distancia en pulgadas del salto de longitud de un jugador desde una posición base (es decir, sin carrerilla). La prueba tiene un objetivo similar al de la prueba vertical.

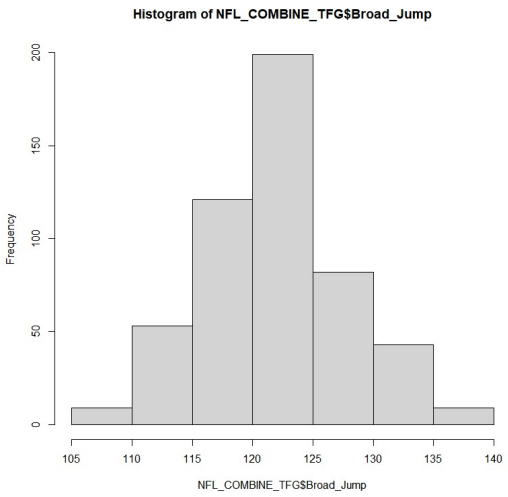

Figura 5.6

Min.	105.0
1st Qu.	119.0
Median	122.5
Mean	122.5
3rd Qu.	126.0
Max.	140.0
NA´s	34

Tabla 5.6

Los Saltos de los jugadores varían desde 105 pulgadas a 140 pulgadas, una diferencia de 35 pulgadas (unos 89 centímetros) lo cual es una diferencia considerable. Se puede ver que el salto medio es de 122,5 pulgadas y que la variable cuenta con 34 valores faltantes.

Shuttle: Variable cuantitativa que mide el tiempo que tarda un jugador en hacer el ejercicio. El jugador empieza desde la denominada "Three-point stance" que requiere tener 3 puntos de contacto con el suelo: los dos pies y una mano. El jugador tiene dos conos a cada lado con una distancia de 5 yardas. Primero el jugador corre 5 yardas hacia un cono y toca el suelo con la mano, después corre 10 yardas hacia el otro cono y toca el suelo con la mano y finalmente esprinta 5 yardas hasta cruzar la posición inicial. Esta prueba tiene como objetivo evaluar la velocidad lateral, la explosividad y la aceleración de un jugador.

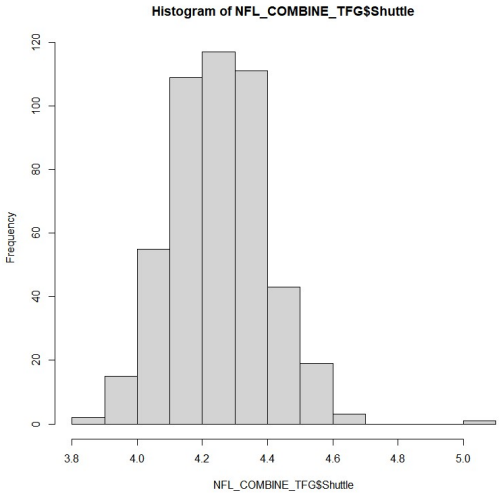

Figura 5.7

Min.	3.810
1st Qu.	4.150
Median	4.250
Mean	4.254
3rd Qu.	4.350
Max.	5.010
NA´s	75

Tabla 5.7

Los tiempos de los jugadores varían desde 3,81 segundos a 5,01 segundos, una diferencia de 1,20 segundos. La media de tiempos es de 4,254 segundos y la variable cuenta con 75 valores faltantes. Se cambiará la unidad de segundos a centésimas de segundo.

Three_Cone: Variable cuantitativa que mide el tiempo que tarda un jugador en hacer el ejercicio. El ejercicio irónicamente cuenta con 4 conos: dos de ellos tienen poca distancia entre ellos y es donde se coloca el jugador para empezar el ejercicio desde el "3-point stance" mencionado anteriormente. Los otros dos conos están colocados en forma de "L" y a 5 yardas de distancia cada una de ellas usando como referencia el cono a la derecha del jugador. El jugador empieza corriendo 5 yardas y tocando el suelo donde el primer cono, vuelve otras 5 yardas y toca el suelo donde ha iniciado el ejercicio, después corre haciendo la forma de la "L" a los otros dos conos y finalmente hace un mini-sprint a la meta. Esta prueba es utilizada para evaluar la rapidez con la que un jugador cambia de dirección mientras acelera.

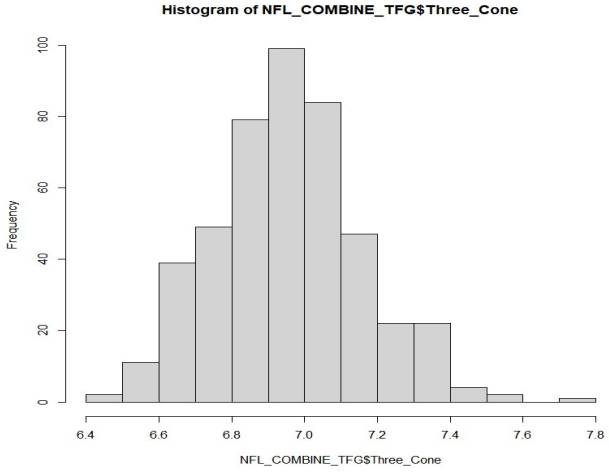

Figura 5.8

Min.	6.480
1st Qu.	6.820
Median	6.960
Mean	6.963
3rd Qu.	7.080
Max.	7.720
NA´s	89

Tabla 5.8

Los tiempos de los jugadores varían desde 6,48 segundos a 7,72 segundos, una diferencia de 1,24 segundos. La media de tiempos es de 6,963 segundos y la variable cuenta con 89 valores faltantes. Se cambiará la unidad de segundos a centésimas de segundo.

Round: Variable cualitativa que hace referencia a la ronda del NFL Draft en la que ha sido seleccionado el jugador. La variable tiene 8 categorías. Esta es la variable respuesta en este trabajo.

X1	Seleccionado en la primera ronda
X2	Seleccionado en la segunda ronda
X3	Seleccionado en la tercera ronda
X4	Seleccionado en la cuarta ronda
X5	Seleccionado en la quinta ronda
X6	Seleccionado en la sexta ronda
X7	Seleccionado en la séptima ronda
Undrafted	No ha sido seleccionado en ninguna ronda

Tabla 5.9

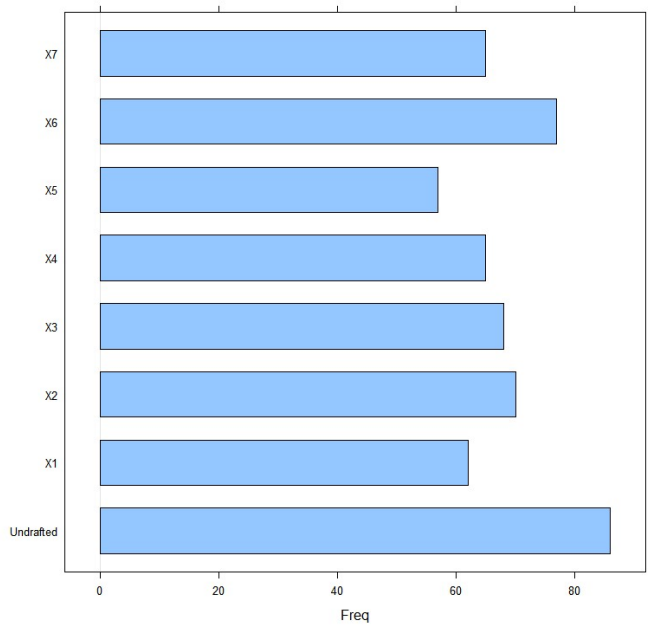

Figura 5.9

Games: Variable cuantitativa que contabiliza el número de partidos que ha jugado el jugador en su carrera universitaria.

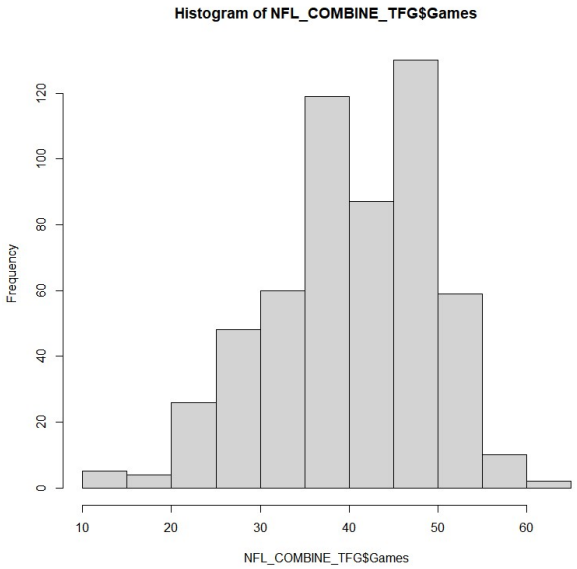

Figura 5.10

Min.	12.00
1st Qu.	35.00
Median	41.00
Mean	40.77
3rd Qu.	48.00
Max.	62.00

Tabla 5.10

Los partidos jugados varían desde 12 partidos a 62 partidos, una diferencia de 50 partidos lo cual es una discrepancia considerable. La media de partidos es de 40,77.

Receptions: Variable cuantitativa que contabiliza el número de recepciones totales que atrapa un jugador a lo largo de su carrera universitaria.

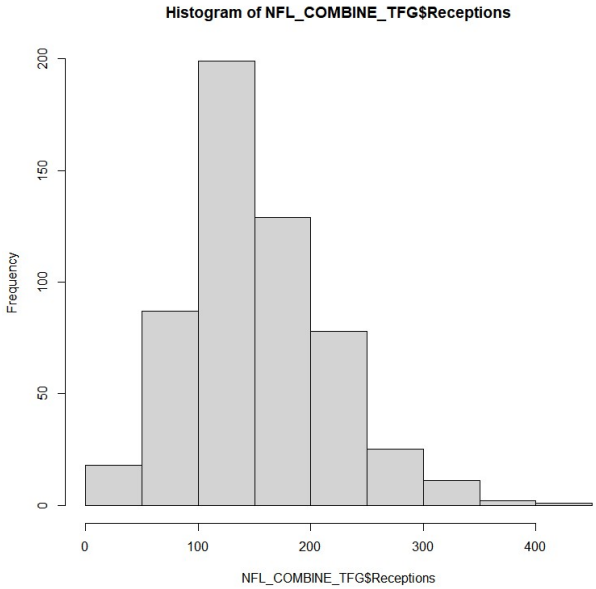

Figura 5.11

Min.	5.0
1st Qu.	111.0
Median	144.0
Mean	151.8
3rd Qu.	186.8
Max.	428.0

Tabla 5.11

Las recepciones varían desde 5 recepciones a 428 recepciones, una diferencia de 423 recepciones lo cual es una discrepancia considerable. La media de recepciones es de 151,8.

TD: Variable cuantitativa que contabiliza en número de touchdowns que consigue el jugador a lo largo de su carrera universitaria.

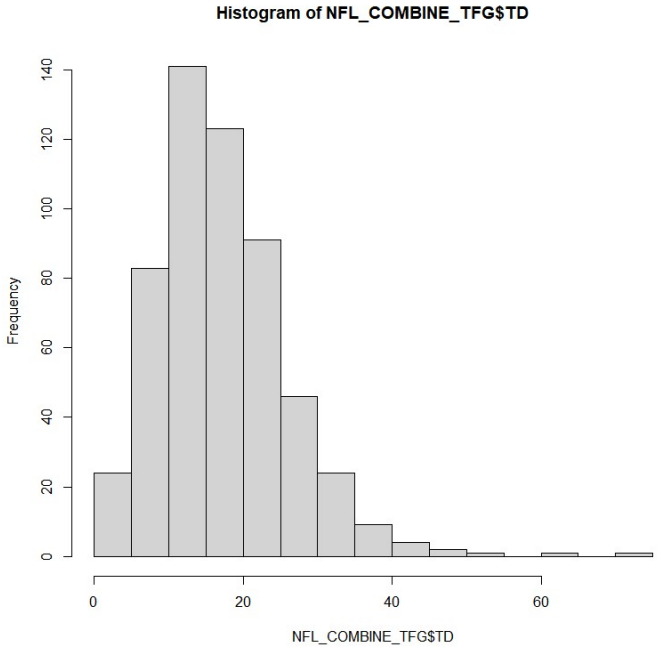

Figura 5.12

Min.	0.00
1st Qu.	12.00
Median	17.00
Mean	17.74
3rd Qu.	22.75
Max.	73.00

Tabla 5.12

Los touchdowns varían desde 0 touchdowns a 73 touchdowns, una diferencia de 73 touchdowns lo cual es una discrepancia considerable. La media de touchdowns es de 17,74.

5.2. CREACIÓN DE VARIABLES

En este apartado se crearán variables nuevas que serán utilizados para el análisis y la construcción del modelo. Estas variables son creadas a partir de la transformación de variables disponibles en el conjunto de datos inicial.

First_Division: Variable dicotómica creada a partir de la transformación de la variable School.

1	La universidad pertenece a la primera división
0	La universidad no pertenece a la primera división

Tabla 5.13

1	527
0	23

Tabla 5.14

Power_Conference: Variable dicotómica creada a partir de la transformación de la variable School.

1	La universidad pertenece a una de las cuatro Power Conferences
0	La universidad no pertenece a una de las cuatro Power Conferences

Tabla 5.15

1	403
0	147

Tabla 5.16

La primera división del fútbol americano universitario cuenta con diez conferencias y algunos equipos independientes. Dentro de estas 10 conferencias se encuentras las denominadas Power Conferences: 4 divisiones distintas que comúnmente se consideran de élite (es decir, un paso por encima de primera división).

Debido a temas financieros y televisivos, ha habido varios equipos que han intentado entrar dentro de estas conferencias de élite y esto ha llevado a 2 reestructuraciones que afectan a esta base de

datos: una a los principios de los 2010s y otro en los 2020s. Estos datos afectados fueron manualmente corregidos en el Excel, en la figura 5.13 se muestran las excepciones.

Pittsburgh Power conference since 2014 draft (Joined ACC in 2013)	Power Conference?	Rutgers Power Conference since 2015 draft (Joined BIG-10 in 2014)	Power Conference?	Utah Power conference since 2012 draft(Joined PAC-12 in 2011)	Power Conference?
Bub Means	YES	Bo Melton	YES	Devaughn Vele	YES
Tyler Boyd	YES	Leonte Carroo	YES	Kaelin Clay	YES
Devin Street	YES	Mohamed Sanu	NO	David Reed	NO
Jester Weah	YES				
Quadree Henderson	YES				
Jon Baldwin	NO				
TCU **Power conference since 2013 draft (Joined BIG-12 in 2012)**	Power Conference?	**UCF** **Power conference since 2024 draft (Joined BIG-12 in 2023)**	Power Conference?		
Quentin Johnston	YES	Javon Baker	YES		
Derius Davis	YES	Tre Nixon	NO		
Jalen Reagor	YES	Gabriel Davis	NO		
Josh Doctson	YES	Tre'Quan Smith	NO		
Kolby Listenbee	YES	Breshad Perriman	NO		
Josh Boyce	YES	Rannell Hall	NO		
Jeremy Kerley	NO	Jacob Harris	NO		

Figura 5.13

Yards_Receptions: Variable cuantitativa creada por la división entre las variables Yards y la variable Receptions.

Yards_Game: Variable cuantitativa creada por la división entre las variables Yards y la variable Game.

Receptions_Game: Variable cuantitativa creada por la división entre las variables Receptions y la variable Game.

5.3. TRATAMIENTO DE VALORES PERDIDOS

Como se ha podido ver en el análisis descriptivo de las variables, este conjunto de datos contiene valores perdidos. Primero se verá que variables contienen valores perdidos y en qué porcentaje para averiguar qué tipo de datos perdidos son y como se puede abordar el problema.

La información se proyecta en la tabla 5.17:

Fourty_yd	Bench	Vertical	Broad_Jump	Shuttle	Three_Cone
3,818182%	22,545455%	6,181818%	6,181818%	13,636364%	16,181818%

Tabla 5.17

Seis variables contienen valores perdidos, siendo Fourty_yd el que menos tiene con 3,82% y Bench el que más contiene con 22,55%. Ahora se utilizará el test de Little para ver si son MCAR y después se va a crear un gráfico que represente el porcentaje de datos “missing” y sus patrones a través del uso de VIM::aggr.

statistic	df	p.value	missing.patterns
<dbl>	<dbl>	<dbl>	<int>
421.	331	0.000583	23

Tabla 5.18

Al tener un p-valor<0,05, se rechaza la hipótesis nula de que los datos sean MCAR.

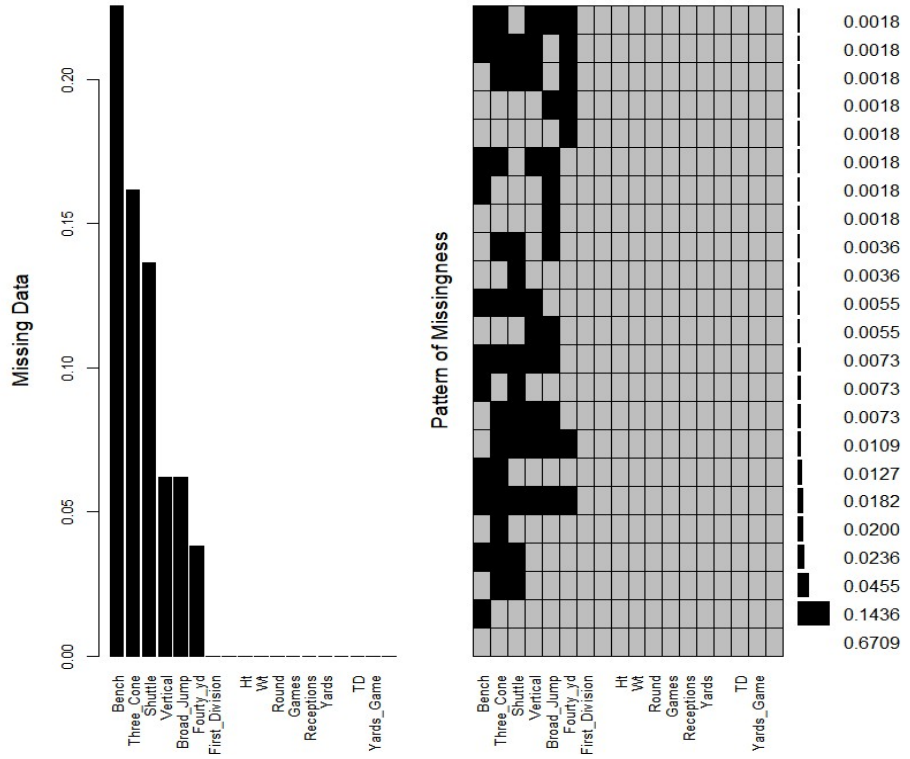

Figura 5.14

Aquí se puede ver los 23 patrones de valores perdidos que se han obtenido en la prueba de Little. En este estudio los datos serán tratados como MAR debido a que hay varias agrupaciones por lo que se trabajará bajo la asunción de que los valores faltantes dependen de otras variables observadas.

Al contar con una B.B.D.D. relativamente pequeña se evitará eliminar a los jugadores del análisis y se abordará el problema a través de la imputación.

Se utilizará la imputación knn (K-nearest neighbours) debido a su flexibilidad, su capacidad de manejar datos cuantitativos y cualitativos. En este estudio se utiliza knnImputation en R utilizando el método de la mediana y los 10 vecinos por defecto que vienen dado por la función. Este método también es utilizado en un estudio similar (Dhar, A. 2011)

Fourty_yd	Bench	Vertical	Broad_Jump	Shuttle	Three_Cone
Min. :4.210	Min. : 4.00	Min. :27.0	Min. :105.0	Min. :3.810	Min. :6.480
1st Qu.:4.420	1st Qu.:12.00	1st Qu.:33.5	1st Qu.:119.0	1st Qu.:4.150	1st Qu.:6.820
Median :4.490	Median :14.00	Median :36.0	Median :122.5	Median :4.250	Median :6.960
Mean :4.488	Mean :14.69	Mean :35.8	Mean :122.5	Mean :4.254	Mean :6.963
3rd Qu.:4.550	3rd Qu.:17.00	3rd Qu.:38.0	3rd Qu.:126.0	3rd Qu.:4.350	3rd Qu.:7.080
Max. :4.840	Max. :29.00	Max. :45.0	Max. :140.0	Max. :5.010	Max. :7.720
Sd : 0.09713	Sd : 4.12485	Sd : 3.01914	Sd : 5.84608	Sd : 0.14971	Sd : 0.20212
NA's :21	NA's :124	NA's :34	NA's :34	NA's :75	NA's :89

Tabla 5.19: métricas de las variables antes de ser imputadas.

Fourty_yd	Bench	Vertical	Broad_Jump	Shuttle	Three_Cone
Min. :4.210	Min. : 4.0	Min. :27.00	Min. :105.0	Min. :3.810	Min. :6.480
1st Qu.:4.423	1st Qu.:12.0	1st Qu.:34.00	1st Qu.:119.0	1st Qu.:4.150	1st Qu.:6.836
Median :4.490	Median :14.0	Median :36.00	Median :122.5	Median :4.258	Median :6.950
Mean :4.489	Mean :14.6	Mean :35.81	Mean :122.5	Mean :4.254	Mean :6.959
3rd Qu.:4.550	3rd Qu.:17.0	3rd Qu.:37.50	3rd Qu.:125.0	3rd Qu.:4.330	3rd Qu.:7.070
Sd : 0.09563	Sd : 3.72799	Sd : 2.93498	Sd : 5.70298	Sd : 0.14116	Sd : 0.18862
Max. :4.840	Max. :29.0	Max. :45.00	Max. :140.0	Max. :5.010	Max. :7.720

Tabla 5.20: métricas de las variables tras ser imputadas.

Se puede ver que la desviación típica y los cuartiles no varían significativamente lo cual es una buena señal de que la imputación no está alterando indebidamente la dispersión y la distribución de los datos. Un problema común en el caso de la imputación es que, al utilizar únicamente los datos existentes, se puede reducir significativamente la variabilidad afectando los análisis de los modelos predictivos.

6. MULTICOLINEALIDAD

En el estudio de la multicolinealidad utilizando el VIF se obtiene el siguiente resultado:

First_Division	Power_Conference	Ht	Wt	Fourty_yd
1.240596	1.226803	2.764239	2.947172	1.474043
Bench	**Vertical**	**Broad_Jump**	**Shuttle**	
1.379670	2.015293	2.071946	1.585615	
Three_Cone	**Games**	**Receptions**	**Yards**	
1.538014	8.131961	125.374938	120.224679	
Yards_Receptions	**TD**	**Yards_Game**	**Receptions_Game**	
6.147105	3.763302	117.050204	125.030008	

Tabla 6.1

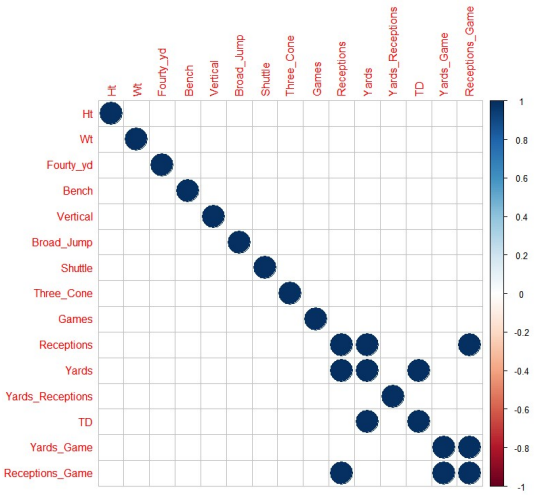

Figura 6.1

Las variables Receptions, Yards, Yards_Game y Receptions_Game presentan multicolinealidad severa. Esto se observa también en la matriz de correlaciones donde los puntos azules representan una correlación por encima del 0,8 absoluto. Receptions, Yards y Receptions_Game tienen correlación alta con otras dos variables mientras que TD y Games tienen una correlación alta con

una. Se elimina Receptions del modelo al ser la que mayor VIF tiene y se vuelve a calcular los datos para ver si se siguen teniendo problemas de multicolinealidad.

First_Division	Power_Conference	Ht	Wt
1.235589	1.221995	2.763819	2.946733
Bench	**Vertical**	**Broad_Jump**	**Shuttle**
1.376985	2.013879	2.069781	1.585604
Three_Cone	**Games**	**Fourty_yd**	**Yards**
1.537735	8.092848	1,472922	22.066386
Yards_Receptions	**TD**	**Yards_Game**	**Receptions_Game**
6.074452	3.731477	39.241186	28.034261

Tabla 6.2

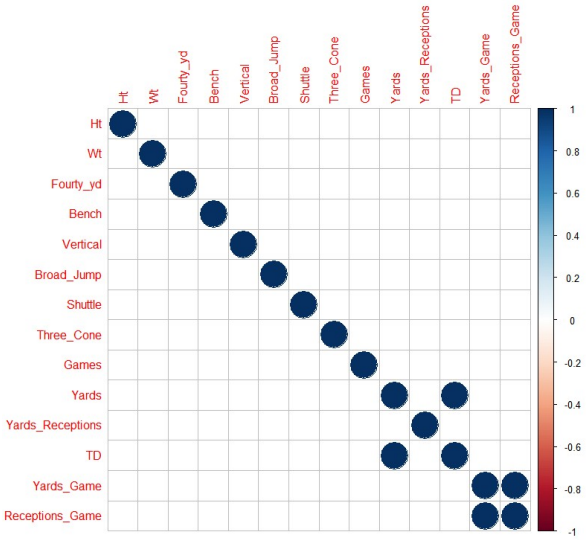

Figura 6.2

Yards, Yards_Game y Receptions_Game siguen teniendo un VIF>10 y presentan alta correlación en la matriz, por lo que se mantendrá el método anterior de eliminar la variable con mayor VIF que en este caso es Yards_Game.

First_Division	Power_Conference	Ht	Wt
1.227969	1.221316	2.760842	2.939785
Bench	**Vertical**	**Broad_Jump**	**Shuttle**
1.361853	2.013830	2.060612	1.583455
Three_Cone	**Games**	**Fourty_yd**	**Yards**
1.535979	5.703522	1.472907	15.818057
Yards_Receptions	**TD**	**Receptions_Game**	
3.482916	3.730511	12.400352	

Tabla 6.3

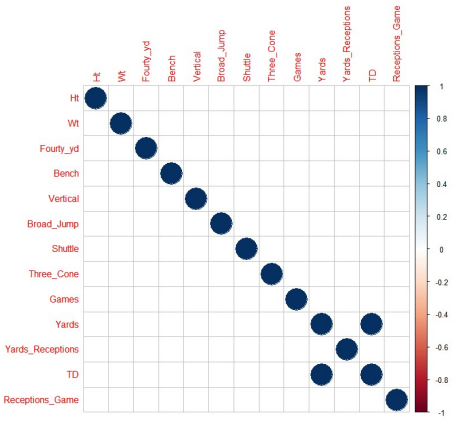

Figura 6.3

Yards y Receptions_Game siguen teniendo un VIF>10 por lo que se elimina Yards ya que tiene mayor VIF que Receptions_Game y tiene una alta correlación con TD.

First_Division	Power_Conference	Ht	Wt
1.225413	1.221239	2.759681	2.926037
Bench	**Vertical**	**Broad_Jump**	**Shuttle**
1.361845	2.011130	2.059821	1.583450
Three_Cone	**Games**	**Fourty_yd**	
1.535711	1.742479	1.441425	
Yards_Receptions	**TD**	**Receptions_Game**	
1.846655	2.8549033	3.096477	

Tabla 6.4

Finalmente las 14 variables independientes de la tabla 6.4 son las utilizadas para calcular el modelo logístico multinomial y el modelo logístico ordinal.

7. MACHINE LEARNING

Machine Learning es una forma de inteligencia artificial que permite a los ordenadores aprender. Cuando es expuesto a nuevos datos, las máquinas pueden esencialmente auto enseñarse programas. Técnicamente Machine Learning es una subcategoría de inteligencia artificial que trata especialmente con reconocimiento de patrones, procesamiento de lenguajes naturales, y redes neuronales. Machine Learning se construye a partir de los análisis predictivos, añadiendo mejor monitoreo y retroalimentación. (Regina Luttrell et al., 2022)

7.1. REGRESIÓN LOGÍSTICA MULTINOMIAL

Primero se va a mirar el reparto de las categorías pertenecientes a nuestra variable dependiente Round.

X1	Undrafted	X2	X3	X4	X5	X6	X7
0.1127273	0.1563636	0.1272727	0.1236364	0.1181818	0.1036364	0.1400000	0.1181818

Tabla 7.1

En la tabla 7.1 se puede observar que existen observaciones en todas las categorías y que tienen unos porcentajes relativamente similares que varían entre 10% y 15%. Se tomará la categoría X1 como referencia para la estimación de los parámetros de la regresión multinomial.

Se ha utilizado 2025 como semilla para generar números pseudoaleatorios y así realizar la división entrenamiento-prueba de los cuales el 80% de los datos han sido asignados a entrenamiento y el 20% restante ha sido asignado a prueba.

Se va a construir un modelo utilizando todas las variables independientes. El modelo creado cuenta con 105 parámetros.

Se obtienen los contrastes de significación individual a través de "Stargazer" en R, estos se muestran en la siguiente tabla.

	Dependent variable:						
	Undrafted (1)	X2 (2)	X3 (3)	X4 (4)	X5 (5)	X6 (6)	X7 (7)
First_Division1	-5.513***	-21.367***	-23.339***	-22.800***	-25.072***	-24.583***	-25.884***
	(0.001)	(0.128)	(0.688)	(0.089)	(0.629)	(0.598)	(0.539)
Power_Conference1	-3.327***	-2.041***	-3.104***	-2.817***	-3.421***	-3.545***	-3.587***
	(0.316)	(0.388)	(0.320)	(0.328)	(0.342)	(0.296)	(0.328)
Ht	-0.057	0.060	-0.089	-0.022	0.129	-0.139	0.104
	(0.164)	(0.145)	(0.155)	(0.156)	(0.171)	(0.159)	(0.167)
Wt	-0.035	-0.025	-0.009	-0.041*	-0.063**	-0.023	-0.057**
	(0.025)	(0.022)	(0.023)	(0.023)	(0.026)	(0.024)	(0.026)
Fourty_yd	0.232***	0.035	0.086***	0.105***	0.154***	0.130***	0.178***
	(0.028)	(0.025)	(0.026)	(0.027)	(0.028)	(0.027)	(0.028)
Bench	-0.046	0.191***	0.006	0.021	-0.046	0.003	-0.012
	(0.079)	(0.069)	(0.074)	(0.075)	(0.079)	(0.076)	(0.079)
Vertical	-0.060	0.028	-0.067	-0.109	0.014	-0.095	-0.086
	(0.120)	(0.108)	(0.114)	(0.116)	(0.123)	(0.117)	(0.122)
Broad_Jump	-0.147**	-0.035	-0.046	-0.114**	-0.170***	-0.138**	-0.089
	(0.061)	(0.055)	(0.056)	(0.058)	(0.063)	(0.060)	(0.062)
Shuttle	-0.001	0.023	-0.008	-0.012	-0.016	-0.004	-0.011
	(0.020)	(0.017)	(0.019)	(0.020)	(0.022)	(0.020)	(0.021)
Three_Cone	0.016	-0.009	0.0002	-0.001	0.003	0.018	0.007
	(0.015)	(0.013)	(0.014)	(0.015)	(0.016)	(0.015)	(0.016)
Games	0.072**	0.012	0.042	0.040	0.117***	0.043	0.066*
	(0.037)	(0.035)	(0.036)	(0.036)	(0.039)	(0.036)	(0.037)
Yards_Receptions	-0.340**	-0.298**	-0.404***	-0.225*	-0.132	-0.378***	-0.414***
	(0.137)	(0.124)	(0.131)	(0.130)	(0.139)	(0.132)	(0.137)
TD	-0.162***	-0.052	-0.065	-0.074*	-0.199***	-0.163***	-0.158***
	(0.049)	(0.039)	(0.041)	(0.042)	(0.050)	(0.047)	(0.049)
Receptions_Game	-1.214***	-0.596**	-0.957***	-1.110***	-0.817**	-1.100***	-1.245***
	(0.304)	(0.279)	(0.286)	(0.301)	(0.319)	(0.296)	(0.309)
Constant	-64.239***	13.467***	17.177***	19.695***	-7.756***	5.526***	-21.281***
	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Akaike Inf. Crit.	1,688.473	1,688.473	1,688.473	1,688.473	1,688.473	1,688.473	1,688.473
Note:							*p<0.1; **p<0.05; ***p<0.01

Tabla 7.2

Al contar la variable respuesta con muchas categorías, un asunto recurrente es que algunas variables independientes no tengan parámetros significativos en una o varias categorías como es el caso de TD, que no tiene parámetros significativos ni en X2, ni en X3, ni en X4.

Las variables cuyos parámetros son significativos para todas las categorías son First_Division, Power_Conference, y Receptions_Game. Estas categorías serán utilizadas en el modelo final.

A continuación, se hará un ANOVA de tipo II para ver qué variables son conjuntamente significativas y así añadir más variables al modelo final si fuera el caso.

```
Analysis of Deviance Table (Type II tests)

Response: Round
                  LR Chisq Df Pr(>Chisq)
First_Division      32.235  7  3.674e-05 ***
Power_Conference    36.109  7  6.914e-06 ***
Ht                   7.750  7  0.3551363
Wt                  10.420  7  0.1659948
Fourty_yd           72.462  7  4.693e-13 ***
Bench               18.732  7  0.0090712 **
Vertical             5.446  7  0.6056464
Broad_Jump          15.101  7  0.0347230 *
Shuttle              7.774  7  0.3529144
Three_Cone           6.339  7  0.5007995
Games               14.584  7  0.0417170 *
Yards_Receptions    20.365  7  0.0048330 **
TD                  28.127  7  0.0002085 ***
Receptions_Game     24.321  7  0.0010005 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Tabla 7.3

Finalmente se añadirán Fourty_yd, Bench, Broad_Jump, Games, Yards_Reception y TD porque son conjuntamente significativas.

```
formula = Round ~ First_Division + Power_Conference + Fourty_yd + Bench +
                  Broad_Jump + Games + Yards_Receptions + TD + Receptions_Game
```

Una vez creado el modelo multinomial con las variables seleccionadas sale que tiene 70 parámetros. Se observará de nuevo el ANOVA en la tabla 7.4, los contrastes de significación individual en la tabla 7.5 y después se obtendrán los ODDS-Ratio de los efectos del modelo para cuantificar el efecto de las variables independientes sobre Round.

```
Analysis of Deviance Table (Type II tests)

Response: Round
                  LR Chisq Df Pr(>Chisq)
First_Division      28.064  7  0.0002141 ***
Power_Conference    61.887  7  6.334e-11 ***
Fourty_yd           79.014  7  2.188e-14 ***
Bench               33.347  7  2.281e-05 ***
Broad_Jump          25.781  7  0.0005512 ***
Games               21.197  7  0.0034888 **
Yards_Receptions    32.050  7  3.976e-05 ***
TD                  39.715  7  1.427e-06 ***
Receptions_Game     28.253  7  0.0001978 ***

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Tabla 7.4

```
================================================================================================
                                          Dependent variable:
                    ------------------------------------------------------------------------
                    Undrafted      X2         X3         X4         X5         X6         X7
                       (1)         (2)        (3)        (4)        (5)        (6)        (7)
------------------------------------------------------------------------------------------------
First_Division1    -2.240***  -13.863*** -15.661*** -14.997*** -17.163*** -16.661*** -18.054***
                    (0.002)    (1.022)    (0.681)    (1.012)    (0.780)    (0.673)    (0.608)

Power_Conference1 -3.372***   -1.984**   -3.085***  -2.846***  -3.402***  -3.599***  -3.616***
                    (0.804)    (0.775)    (0.770)    (0.787)    (0.814)    (0.788)    (0.811)

Fourty_yd           0.202***    0.022*    0.063***   0.066***   0.112***   0.096***   0.142***
                    (0.014)    (0.013)    (0.013)    (0.014)    (0.015)    (0.014)    (0.014)

Bench               -0.110     0.151**     -0.027     -0.051     -0.119*     -0.049     -0.090
                    (0.071)    (0.061)    (0.067)    (0.069)    (0.072)    (0.069)    (0.071)

Broad_Jump         -0.165***    -0.029    -0.065*   -0.143***  -0.158***  -0.170***   -0.108**
                    (0.044)    (0.038)    (0.040)    (0.041)    (0.045)    (0.043)    (0.044)

Games               0.070**     0.012      0.046      0.044     0.123***    0.042      0.068*
                    (0.036)    (0.034)    (0.035)    (0.035)    (0.038)    (0.035)    (0.036)
```

Yards_Receptions	-0.352***	-0.293**	-0.408***	-0.251**	-0.140	-0.392***	-0.420***
	(0.132)	(0.124)	(0.129)	(0.127)	(0.135)	(0.127)	(0.133)
TD	-0.155***	-0.042	-0.060	-0.066	-0.196***	-0.157***	-0.154***
	(0.048)	(0.039)	(0.041)	(0.042)	(0.050)	(0.046)	(0.048)
Receptions_Game	-1.156***	-0.598**	-0.882***	-1.017***	-0.692**	-1.026***	-1.150***
	(0.307)	(0.283)	(0.291)	(0.303)	(0.324)	(0.298)	(0.313)
Constant	-53.685***	14.625***	7.928***	14.239***	-5.284***	10.019***	-16.993***
	(0.002)	(0.017)	(0.008)	(0.007)	(0.002)	(0.002)	(0.003)
Akaike Inf. Crit.	1,649.737	1,649.737	1,649.737	1,649.737	1,649.737	1,649.737	1,649.737
Note:							*p<0.1; **p<0.05; ***p<0.01

Tabla 7.5

	First_Division1	Power_Conference1	Fourty_yd	Bench	Broad_Jump	Games	Yards_Receptions	TD	Receptions_Game
Undrafted	1.06E-01	0.03431747	1.22421	0.896057	0.8481157	1.07255	0.7030548	0.856558	0.3146646
X2	9.53E-07	0.1375764	1.02254	1.162912	0.9714839	1.01206	0.745773	0.959254	0.5500628
X3	1.58E-07	0.04571517	1.06552	0.9733	0.9366307	1.04681	0.6651673	0.941558	0.4138199
X4	3.07E-07	0.05807034	1.06802	0.949818	0.8669405	1.04452	0.7777648	0.935867	0.3617035
X5	3.52E-08	0.03331175	1.11862	0.887453	0.8535706	1.13088	0.8692878	0.822156	0.5007836
X6	5.81E-08	0.02733802	1.10076	0.952368	0.8433366	1.04253	0.6754653	0.854923	0.3585499
X7	1.44E-08	0.02687805	1.15308	0.913906	0.8972287	1.07035	0.6571014	0.857564	0.3166833

Tabla 7.6

La interpretación de los odds-ratio de la tabla 7.6 es la siguiente:

- **First_Division:** La categoría de referencia es 0.
  - **Undrafted:** El odds-ratio es inferior a uno por lo que los jugadores de primera división tienen un 89,35% menos de posibilidades de no ser seleccionados frente a ser seleccionados en la primera ronda, en comparación con los que no juegan en primera división.
  - **X2-X7:** El odds-ratio es prácticamente nulo por lo que los jugadores de primera división no tienen posibilidad de ser seleccionados en otra ronda frente a ser seleccionados en la primera ronda, en comparación con los que no juegan en primera división.
- **Power_Conference:** La categoría de referencia es 0.
  - **Undrafted:** El odds-ratio es inferior a uno por lo que los jugadores de las conferencias de élite tienen un 96,57% menos de posibilidades de no ser seleccionados frente a ser seleccionados en la primera ronda, en comparación con los que no juegan en las conferencias de élite.
  - **X2:** El odds-ratio es inferior a uno por lo que los jugadores de las conferencias de élite tienen un 86,24% menos de posibilidades de ser seleccionado en la segunda ronda frente a ser seleccionados en la primera ronda, en comparación con los que no juegan en las conferencias de élite.
  - **X3:** El odds-ratio es inferior a uno por lo que los jugadores de las conferencias de élite tienen un 95,43% menos de posibilidades de ser seleccionados en la tercera ronda frente a ser seleccionados en la primera ronda, en comparación con los que no juegan en las conferencias de élite.
  - **X4:** El odds-ratio es inferior a uno por lo que los jugadores de las conferencias de élite tienen un 94,19% menos de posibilidades de ser seleccionados en la cuarta ronda frente a ser seleccionados en la primera ronda, en comparación con los que no juegan en las conferencias de élite.
  - **X5:** El odds-ratio es inferior a uno por lo que los jugadores de las conferencias de élite tienen un 96,67% menos de posibilidades de ser seleccionados en la quinta

ronda frente a ser seleccionados en la primera ronda, en comparación con los que no juegan en las conferencias de élite.
    - **X6:** El odds-ratio es inferior a uno por lo que los jugadores de las conferencias de élite tienen un 97,27% menos de posibilidades de ser seleccionados en la sexta ronda frente a ser seleccionados en la primera ronda, en comparación con los que no juegan en las conferencias de élite.
    - **X7:** El odds-ratio es inferior a uno por lo que los jugadores de las conferencias de élite tienen un 97,31% menos de posibilidades de ser seleccionados en la séptima ronda frente a ser seleccionados en la primera ronda, en comparación con los que no juegan en las conferencias de élite.
- **Fourty_yd:**
    - **Undrafted:** el odds-ratio es superior a 1, por lo que el aumento del tiempo aumenta las posibilidades de no ser seleccionado frente a ser seleccionado en la primera ronda. Concretamente, por cada centésima de segundo adicional, estas posibilidades aumentan en un 22,42%.
    - **X2:** El parámetro no es significativo, por lo que el tiempo tardado no influye en el hecho de ser seleccionado en la segunda ronda o ser seleccionado en la primera ronda.
    - **X3:** el odds-ratio es superior a 1, por lo que el aumento del tiempo aumenta las posibilidades de ser seleccionado en la tercera ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada centésima de segundo adicional, estas posibilidades aumentan en un 6,55%.
    - **X4:** el odds-ratio es superior a 1, por lo que el aumento del tiempo aumenta las posibilidades de ser seleccionado en la cuarta ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada centésima de segundo adicional, estas posibilidades aumentan en un 6,80%.
    - **X5:** el odds-ratio es superior a 1, por lo que el aumento del tiempo aumenta las posibilidades de ser seleccionado en la quinta ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada centésima de segundo adicional, estas posibilidades aumentan en un 11,86%.
    - **X6:** el odds-ratio es superior a 1, por lo que el aumento del tiempo aumenta las posibilidades de ser seleccionado en la sexta ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada centésima de segundo adicional, estas posibilidades aumentan en un 10,07%.
    - **X7:** el odds-ratio es superior a 1, por lo que el aumento del tiempo aumenta las posibilidades de ser seleccionado en la séptima ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada centésima de segundo adicional, estas posibilidades aumentan en un 15,31%.
- **Bench:**
    - **Undrafted, X3-X7:** El parámetro no es significativo, por lo que el número de repeticiones no influye en el hecho de no ser seleccionado, ser seleccionado entre la tercera y séptima ronda o ser seleccionado en la primera ronda.
    - **X2:** el odds-ratio es superior a 1, por lo que el aumento de las repeticiones aumenta las posibilidades de ser seleccionado en la segunda ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada repetición adicional, estas posibilidades aumentan en un 16,29%.
- **Broad_Jump:**
    - **Undrafted:** el odds-ratio es inferior a 1, por lo que el aumento del salto disminuye las posibilidades de no ser seleccionado frente a ser seleccionado en la primera ronda. Concretamente, por cada pulgada adicional, estas posibilidades disminuyen en un 15,19%.

  - **X2-X3:** El parámetro no es significativo, por lo que el salto no influye en el hecho de ser seleccionado en la segunda ronda, en la tercera ronda o ser seleccionado en la primera ronda.
  - **X4:** el odds-ratio es inferior a 1, por lo que el aumento del salto disminuye las posibilidades de ser seleccionado en la cuarta ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada pulgada adicional, estas posibilidades disminuyen en un 13,31%.
  - **X5:** el odds-ratio es inferior a 1, por lo que el aumento del salto disminuye las posibilidades de ser seleccionado en la quinta ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada pulgada adicional, estas posibilidades disminuyen en un 14,64%.
  - **X6:** el odds-ratio es inferior a 1, por lo que el aumento del salto disminuye las posibilidades de ser seleccionado en la sexta ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada pulgada adicional, estas posibilidades disminuyen en un 15,67%.
  - **X7:** el odds-ratio es inferior a 1, por lo que el aumento del salto disminuye las posibilidades de ser seleccionado en la séptima ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada pulgada adicional, estas posibilidades disminuyen en un 10,28%.
- **Games**:
  - **Undrafted:** el odds-ratio es superior a 1, por lo que el aumento de partidos aumenta las posibilidades de no ser seleccionado frente a ser seleccionado en la primera ronda. Concretamente, por cada partido adicional, estas posibilidades aumentan en un 7,26%.
  - **X2-X4, X6-X7:** El parámetro no es significativo, por lo que el número de partidos jugados no influye en el hecho de ser seleccionado en la segunda, tercera, cuarta, sexta o séptima ronda, o ser seleccionado en la primera ronda.
  - **X5:** el odds-ratio es superior a 1, por lo que el aumento de partidos aumenta las posibilidades de ser seleccionado en la quinta ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada partido adicional, estas posibilidades aumentan en un 13,08%.
- **Yards_Receptions:**
  - **Undrafted:** el odds-ratio es inferior a 1, por lo que el aumento de yardas por recepción disminuye las posibilidades de no ser seleccionado frente a ser seleccionado en la primera ronda. Concretamente, por cada yarda por recepción adicional, estas posibilidades disminuyen en un 29,69%.
  - **X2:** el odds-ratio es superior a 1, por lo que el aumento de yardas por recepción disminuye las posibilidades de ser seleccionado en la segunda ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada yarda por recepción adicional, estas posibilidades disminuyen en un 25,42%.
  - **X3:** el odds-ratio es inferior a 1, por lo que el aumento de yardas por recepción disminuye las posibilidades de ser seleccionado en la tercera ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada yarda por recepción adicional, estas posibilidades disminuyen en un 33,48%.
  - **X4:** el odds-ratio es inferior a 1, por lo que el aumento de yardas por recepción disminuye las posibilidades de ser seleccionado en la cuarta ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada yarda por recepción adicional, estas posibilidades disminuyen en un 22,22%.
  - **X5:** El parámetro no es significativo, por lo que las yardas por recepción no influyen en el hecho de ser seleccionado en la quinta ronda o ser seleccionado en la primera ronda.

  - **X6:** el odds-ratio es inferior a 1, por lo que el aumento de yardas por recepción disminuye las posibilidades de ser seleccionado en la sexta ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada yarda por recepción adicional, estas posibilidades disminuyen en un 32,45%.
  - **X7:** el odds-ratio es inferior a 1, por lo que el aumento de yardas por recepción disminuye las posibilidades de ser seleccionado en la séptima ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada yarda por recepción adicional, estas posibilidades disminuyen en un 34,29%.
- **TD:**
  - **Undrafted:** el odds-ratio es inferior a 1, por lo que el aumento de touchdowns disminuye las posibilidades de no ser seleccionado frente a ser seleccionado en la primera ronda. Concretamente, por cada touchdown adicional, estas posibilidades disminuyen en un 14,34%.
  - **X2-X4:** El parámetro no es significativo, por lo que número de touchdowns no influye en el hecho de ser seleccionado en la segunda ronda, en la tercera ronda, en la cuarta ronda o ser seleccionado en la primera ronda.
  - **X5:** el odds-ratio es inferior a 1, por lo que el aumento de touchdowns disminuye las posibilidades de ser seleccionado en la quinta ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada touchdown adicional, estas posibilidades disminuyen en un 17,78%.
  - **X6:** el odds-ratio es inferior a 1, por lo que el aumento de touchdowns disminuye las posibilidades de ser seleccionado en la sexta ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada touchdown adicional, estas posibilidades disminuyen en un 14,51%.
  - **X7:** el odds-ratio es inferior a 1, por lo que el aumento de touchdowns disminuye las posibilidades de ser seleccionado en la séptima ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada touchdown adicional, estas posibilidades disminuyen en un 14,24%.
- **Receptions_Game:**
  - **Undrafted:** el odds-ratio es inferior a 1, por lo que el aumento de recepciones por partido disminuye las posibilidades de no ser seleccionado frente a ser seleccionado en la primera ronda. Concretamente, por cada recepción por partido adicional, estas posibilidades disminuyen en un 68,53%.
  - **X2:** el odds-ratio es superior a 1, por lo que el aumento de recepciones por partido disminuye las posibilidades de ser seleccionado en la segunda ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada recepción por partido adicional, estas posibilidades disminuyen en un 44,99%.
  - **X3:** el odds-ratio es inferior a 1, por lo que el aumento de recepciones por partido disminuye las posibilidades de ser seleccionado en la tercera ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada recepción por partido adicional, estas posibilidades disminuyen en un 58,61%.
  - **X4:** el odds-ratio es inferior a 1, por lo que el aumento de recepciones por partido disminuye las posibilidades de ser seleccionado en la cuarta ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada recepción por partido adicional, estas posibilidades disminuyen en un 63,83%.
  - **X5:** el odds-ratio es inferior a 1, por lo que el aumento de recepciones por partido disminuye las posibilidades de ser seleccionado en la quinta ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada recepción por partido adicional, estas posibilidades disminuyen en un 49,92%.
  - **X6:** el odds-ratio es inferior a 1, por lo que el aumento de recepciones por partido

disminuye las posibilidades de ser seleccionado en la sexta ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada recepción por partido adicional, estas posibilidades disminuyen en un 64,15%.

- **X7:** el odds-ratio es inferior a 1, por lo que el aumento de recepciones por partido disminuye las posibilidades de ser seleccionado en la séptima ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada recepción por partido adicional, estas posibilidades disminuyen en un 68,33%.

La última parte del estudio del modelo es hacer una evaluación de este.

Se crea una matriz de confusión y a partir de ella se sacan las siguientes conclusiones:

	X1	Undrafted	X2	X3	X4	X5	X6	X7
X1	8	0	8	1	2	1	1	3
Undrafted	0	7	2	2	2	2	0	2
X2	2	1	1	4	2	1	4	0
X3	2	1	2	1	0	1	2	4
X4	0	0	1	1	2	1	3	0
X5	0	3	0	2	3	1	0	1
X6	0	4	0	2	2	2	4	1
X7	0	1	0	0	0	2	1	2

Tabla 7.7

El eje horizontal representa los casos reales mientras que el eje vertical representa las predicciones.

La forma de interpretar esta tabla es sabiendo que la diagonal principal representa las observaciones clasificadas correctamente, mientras que el resto de las celdas corresponden con fallos. Por ejemplo, en el caso de X1 se han clasificado correctamente 8 observaciones, mientras que 8 fueron clasificada como X2, 1 como X3, 2 como X4, 1 como X5, 1 como X6 y 3 como X7. La matriz de confusión asigna las observaciones a aquella categoría con mayor probabilidad predicha, lo que puede dar resultados malos para las categorías minoritarias.

Accuracy	Kappa
0.2407407	0.1309979

Tabla 7.8

El modelo tan solo consigue clasificar correctamente el 24,07% de los datos y da lugar a un índice de Kappa de 0,13 lo que implica una calidad pobre. Mientras que los datos de entrenamiento clasifican correctamente el 36,65% de los datos y da lugar a un índice de Kappa de 0,27 lo cual es un aumento considerable. Esta diferencia es un indicio bastante claro de sobreajuste.

“En el machine learning, el sobreajuste ocurre cuando un algoritmo se ajusta demasiado o incluso exactamente a sus datos de entrenamiento, lo que da como resultado un modelo que no puede hacer predicciones o conclusiones precisas a partir de ningún otro dato que no sea el de entrenamiento.” (IBM, n.d.b.)

	Sensitivity	Specificity
Class:X1	0.66666667	0.8333333
Class: Undrafted	0.41176471	0.8901099
Class: X2	0.07142857	0.8510638
Class: X3	0.07692308	0.8736842
Class: X4	0.15384615	0.9368421
Class: X5	0.09090909	0.9072165
Class: X6	0.26666667	0.8817204
Class: X7	0.15384615	0.9578947

Tabla 7.9

La interpretación de la sensibilidad y especificidad mostrada en la tabla 7.9 es la siguiente:

- **Sensibilidad:** Identificar los verdaderos positivos P(+|E)
  - **X1:** El modelo identifica correctamente el 66,66% de los casos que pertenecen a X1.
  - **X2:** El modelo identifica correctamente el 7,14% de los casos que pertenecen a X2.
  - **X3:** El modelo identifica correctamente el 7,69% de los casos que pertenecen a X3.
  - **X4:** El modelo identifica correctamente el 15,38% de los casos que pertenecen a X4.
  - **X5:** El modelo identifica correctamente el 9,09% de los casos que pertenecen a X5.
  - **X6:** El modelo identifica correctamente el 26,67% de los casos que pertenecen a X6.
  - **X7:** El modelo identifica correctamente el 15,38% de los casos que pertenecen a X7.
  - **Undrafted:** El modelo identifica correctamente el 41,18% de los casos que pertenecen a Undrafted.

- **Especificidad:** Identifica los verdaderos negativos P(-|E^c)
  - **X1**: El modelo identifica correctamente los casos que no pertenecen a X1 con un acierto del 83,33%.
  - **X2:** El modelo identifica correctamente los casos que no pertenecen a X2 con un acierto del 85,11%.
  - **X3:** El modelo identifica correctamente los casos que no pertenecen a X3 con un acierto del 87,37%
  - **X4:** El modelo identifica correctamente los casos que no pertenecen a X4 con un acierto del 93,68%
  - **X5:** El modelo identifica correctamente los casos que no pertenecen a X5 con un acierto del 90,72%
  - **X6:** El modelo identifica correctamente los casos que no pertenecen a X6 con un acierto del 88,17%
  - **X7:** El modelo identifica correctamente los casos que no pertenecen a X7 con un acierto del 95,79%
  - **Undrafted:** El modelo identifica correctamente los casos que no pertenecen a Undrafted con un acierto del 89,01%.

La sensibilidad y especificidad se interpreta de la misma manera a lo largo del trabajo, por lo que no se volverá a escribir la interpretación en los modelos posteriores.

El modelo funciona de manera asimétrica debido a la diferencia en las frecuencias de las categorías. Esto se puede ver en la matriz de confusión y en las diferencias de sensibilidad de las categorías.

Las curvas ROC es otra herramienta que se va a usar para evaluar el modelo y esta no depende de la estrategia de clasificación. No se puede obtener una única curva ROC porque es multinomial, pero se mostrarán las dos alternativas disponibles: pairwise y one vs all.

- Curva ROC pairwise: `Multi-class area under the curve: 0.7081`
  El AUC medio es de 0,7081 por lo que la calidad es aceptable pero baja.
- Curva ROC one vs all: `0.7076548`

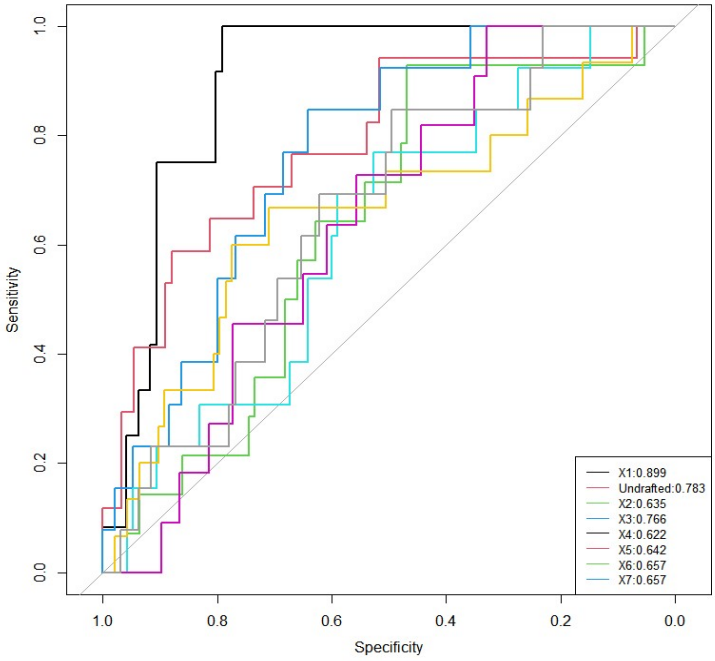

Figura 7.1

La media de los AUC es de 0.7076548, donde X1 es la que mejor se modeliza con un AUC de 0,899 y X5 es la que peor se modeliza con un AUC de 0,622.

Dado que el modelo construido no es convincente se van a construir métodos automáticos de selección de variables para poder comparar distintos modelos. Se usan los métodos backward, forward y stepwise para cada una de las medidas de evaluación AIC y BIC. El número de parámetros que contienen cada uno de los modelos que se han construido son los siguientes:

```
28 56 28 56 70 70
```

Los seis procesos dan como resultados tres modelos distintos, uno con 28 parámetros, otro con 56 y otro con 70:

```
formula = Round ~ TD + Fourty_yd + Power_Conference

formula = Round ~ TD + Fourty_yd + Power_Conference + Games + Bench +
                  First_Division + Broad_Jump

formula = Round ~ First_Division + Power_Conference + Fourty_yd + Bench +
                  Broad_Jump + Games + Yards_Reception + TD + Receptions_Game
```

Coincidentemente el tercer modelo construido y el modelo que se ha construido manualmente son idénticos por lo que se compararán los tres modelos automáticos recurriendo a la validación cruzada repetida, gracias a la cual se puede evaluar los modelos mediante la tasa de acierto, el índice de Kappa y el AUC one vs. all.

La validación cruzada repetida se basa en dividir el conjunto de datos de entrenamiento en submuestras. En cada iteración, el modelo se construye utilizando todas las observaciones menos la de una submuestra determinada, que es utilizada para evaluar el modelo. Este método asegura que todas las observaciones sean utilizadas para hacer predicciones sin que formen parte del modelo en una iteración específica, pero que a la vez contribuyan al entrenamiento del modelo en las demás iteraciones. Esta técnica se considera importante porque su evaluación es más robusta. (Calviño, 2023)

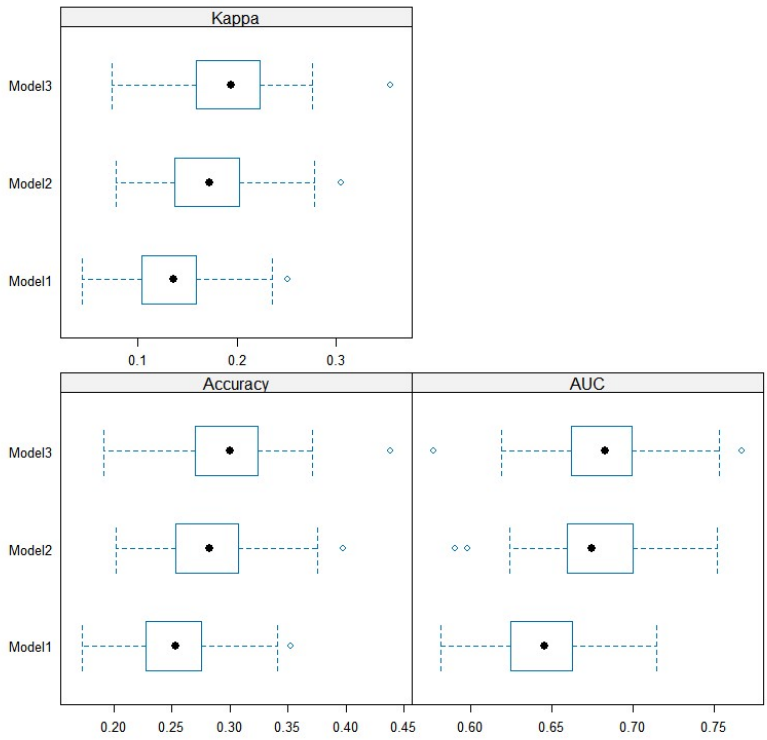

Figura 7.2

En el gráfico superior Model1 representa el modelo de 28 parámetros, el Model2 representa el de 56 parámetros y el Model3 representa el de 70 parámetros.

AUC

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
Model1	0.5822462	0.6250098	0.6456485	0.6447572	0.6630100	0.7147463	0
Model2	0.5905636	0.6598836	0.6752030	0.6778132	0.7002397	0.7519595	0
Model3	0.5775357	0.6623910	0.6830245	0.6823748	0.6990549	0.7676128	0

Kappa

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
Model1	0.04410194	0.1048753	0.1366800	0.1355556	0.1581308	0.2507842	0
Model2	0.07899723	0.1378458	0.1721346	0.1747005	0.2024659	0.3043997	0
Model3	0.07452340	0.1594624	0.1937593	0.1924158	0.2225636	0.3543239	0

Accuracy

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
Model1	0.1724138	0.2272727	0.2528736	0.2529203	0.2750095	0.3522727	0
Model2	0.2022472	0.2528736	0.2824949	0.2836684	0.3068182	0.3977273	0
Model3	0.1910112	0.2696629	0.3000000	0.2979311	0.3231273	0.4382022	0

Tabla 7.10

Debido a que son asimétricas se le da más credibilidad al índice Kappa y al AUC. Sabiendo esto se descarta el Model1 debido a que presentan valores bastante inferiores. En este caso se va a optar por el Model2 debido a que obtiene los resultados similares a Model3, pero es más sencillo ya que cuenta con 14 parámetros menos. De esta forma se puede comparar las evaluaciones de

ambos modelos.

Ahora se interpretará el modelo utilizando las mismas herramientas que fueron usadas en el modelo manual.

	Dependent variable:						
	Undrafted (1)	X2 (2)	X3 (3)	X4 (4)	X5 (5)	X6 (6)	X7 (7)
TD	0.763*** p = 0.000	0.907*** p = 0.001	0.862*** p = 0.00001	0.855*** p = 0.00000	0.783*** p = 0.000	0.770*** p = 0.000	0.759*** p = 0.000
Fourty_yd	1.210*** p = 0.000	1.020** p = 0.046	1.060*** p = 0.00000	1.054*** p = 0.00000	1.106*** p = 0.000	1.092*** p = 0.000	1.143*** p = 0.000
Power_Conference1	0.056*** p = 0.0002	0.189** p = 0.026	0.070*** p = 0.0003	0.096*** p = 0.002	0.051*** p = 0.0002	0.044*** p = 0.00004	0.044*** p = 0.0001
Games	1.163*** p = 0.00000	1.065** p = 0.023	1.124*** p = 0.0001	1.117*** p = 0.0002	1.184*** p = 0.00000	1.127*** p = 0.00003	1.164*** p = 0.00000
Bench	0.908 p = 0.166	1.173*** p = 0.009	0.987 p = 0.843	0.959 p = 0.539	0.894 p = 0.114	0.966 p = 0.609	0.929 p = 0.295
First_Division1	0.089*** p = 0.000	0.000*** p = 0.000	0.000*** p = 0.000	0.000*** p = 0.000	0.000*** p = 0.000	0.000*** p = 0.000	0.000*** p = 0.000
Broad_Jump	0.867*** p = 0.0004	0.978 p = 0.510	0.945 p = 0.114	0.892*** p = 0.002	0.881*** p = 0.002	0.855*** p = 0.00004	0.910** p = 0.017
Constant	0.000*** p = 0.000	194,765.300*** p = 0.000	27.660*** p = 0.000	400,018.400*** p = 0.000	0.008*** p = 0.000	448.338*** p = 0.000	0.000*** p = 0.000
Akaike Inf. Crit.	1,654.629	1,654.629	1,654.629	1,654.629	1,654.629	1,654.629	1,654.629
Note:							*p<0.1; **p<0.05; ***p<0.01

Tabla 7.10

```
Analysis of Deviance Table (Type II tests)

Response: Round
                LR Chisq Df Pr(>Chisq)
TD                118.718  7  < 2.2e-16 ***
Fourty_yd          67.715  7  4.273e-12 ***
Power_Conference   32.697  7  3.014e-05 ***
Games              43.495  7  2.676e-07 ***
Bench              25.757  7  0.0005566 ***
First_Division     21.994  7  0.0025468 **
Broad_Jump         20.074  7  0.0054122 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Se puede ver que todas las variables son conjuntamente significativas y que Bench y Broad_Jump no tienen parámetros significativos para todas las categorías. Se explicarán los odds-ratio de las tres variables más contribuyentes al modelo, que en este caso son TD, Fourty_yd y Games.

	TD	Fourty_yd	Power_Conference1	Games	Bench	First_Division1	Broad_Jump
Undrafted	0.762919	1.20996	0.05607627	1.16257	0.90838	8.90E-02	0.867408
X2	0.906678	1.02016	0.18944251	1.06525	1.173165	3.37E-09	0.9777671
X3	0.862121	1.06035	0.06963981	1.12377	0.986964	4.72E-10	0.9449266
X4	0.855294	1.05421	0.096463	1.11706	0.959435	8.76E-10	0.8920962
X5	0.783196	1.1063	0.05115344	1.18446	0.893544	1.16E-10	0.8808101
X6	0.770221	1.09158	0.04379318	1.12693	0.96618	1.67E-10	0.8545985
X7	0.759359	1.1427	0.043759	1.16445	0.929391	4.33E-11	0.9099048

Tabla 7.12

- **TD: número de touchdowns totales.**
  - **Undrafted:** el odds-ratio es inferior a 1, por lo que el aumento de touchdowns disminuye las posibilidades de no ser seleccionado frente a ser seleccionado en la primera ronda. Concretamente, por cada touchdown adicional, estas posibilidades disminuyen en un 23,71%.
  - **X2:** el odds-ratio es inferior a 1, por lo que el aumento de touchdowns disminuye las posibilidades de ser seleccionado en la segunda ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada touchdown adicional, estas posibilidades disminuyen en un 9,33%.
  - **X3:** el odds-ratio es inferior a 1, por lo que el aumento de touchdowns disminuye las posibilidades de ser seleccionado en la tercera ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada touchdown adicional, estas posibilidades disminuyen en un 13,79%.
  - **X4:** el odds-ratio es inferior a 1, por lo que el aumento de touchdowns disminuye las posibilidades de ser seleccionado en la cuarta ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada touchdown adicional, estas posibilidades disminuyen en un 14,47%.
  - **X5:** el odds-ratio es inferior a 1, por lo que el aumento de touchdowns disminuye las posibilidades de ser seleccionado en la quinta ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada touchdown adicional, estas posibilidades disminuyen en un 21,68%.
  - **X6:** el odds-ratio es inferior a 1, por lo que el aumento de touchdowns disminuye las posibilidades de ser seleccionado en la sexta ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada touchdown adicional, estas posibilidades disminuyen en un 22,98%.
  - **X7:** el odds-ratio es inferior a 1, por lo que el aumento de touchdowns disminuye las posibilidades de ser seleccionado en la séptima ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada touchdown adicional, estas posibilidades disminuyen en un 24,06%.

- **Fourty_yd: tiempo que tarda un receptor en correr cuarenta yardas.**
  - **Undrafted:** el odds-ratio es superior a 1, por lo que el aumento del tiempo aumenta las posibilidades de no ser seleccionado frente a ser seleccionado en la primera ronda. Concretamente, por cada centésima de segundo adicional, estas posibilidades aumentan en un 21%.
  - **X2:** el odds-ratio es superior a 1, por lo que el aumento del tiempo aumenta las posibilidades de ser seleccionado en la segunda ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada centésima de segundo adicional, estas posibilidades aumentan en un 2,02%.
  - **X3:** el odds-ratio es superior a 1, por lo que el aumento del tiempo aumenta las posibilidades de ser seleccionado en la tercera ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada centésima de segundo adicional, estas posibilidades aumentan en un 6,03%.
  - **X4:** el odds-ratio es superior a 1, por lo que el aumento del tiempo aumenta las posibilidades de ser seleccionado en la cuarta ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada centésima de segundo adicional, estas posibilidades aumentan en un 5,42%.
  - **X5:** el odds-ratio es superior a 1, por lo que el aumento del tiempo aumenta las posibilidades de ser seleccionado en la quinta ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada centésima de segundo adicional, estas posibilidades aumentan en un 10,63%.

  - **X6:** el odds-ratio es superior a 1, por lo que el aumento del tiempo aumenta las posibilidades de ser seleccionado en la sexta ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada centésima de segundo adicional, estas posibilidades aumentan en un 9,16%.
  - **X7:** el odds-ratio es superior a 1, por lo que el aumento del tiempo aumenta las posibilidades de ser seleccionado en la séptima ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada centésima de segundo adicional, estas posibilidades aumentan en un 14,27%.

- **Games**: **número de partidos jugados.**
  - **Undrafted:** el odds-ratio es superior a 1, por lo que el aumento de partidos aumenta las posibilidades de no ser seleccionado frente a ser seleccionado en la primera ronda. Concretamente, por cada partido adicional, estas posibilidades aumentan en un 16,26%.
  - **X2:** el odds-ratio es superior a 1, por lo que el aumento de partidos aumenta las posibilidades de ser seleccionado en la segunda ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada partido adicional, estas posibilidades aumentan en un 6,52%.
  - **X3:** el odds-ratio es superior a 1, por lo que el aumento de partidos aumenta las posibilidades de ser seleccionado en la tercera ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada partido adicional, estas posibilidades aumentan en un 12,38%.
  - **X4:** el odds-ratio es superior a 1, por lo que el aumento de partidos aumenta las posibilidades de ser seleccionado en la cuarta ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada partido adicional, estas posibilidades aumentan en un 11,71%.
  - **X5:** el odds-ratio es superior a 1, por lo que el aumento de partidos aumenta las posibilidades de ser seleccionado en la quinta ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada partido adicional, estas posibilidades aumentan en un 18,45%.
  - **X6:** el odds-ratio es superior a 1, por lo que el aumento de partidos aumenta las posibilidades de ser seleccionado en la sexta ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada partido adicional, estas posibilidades aumentan en un 12,69%.
  - **X7:** el odds-ratio es superior a 1, por lo que el aumento de partidos aumenta las posibilidades de ser seleccionado en la séptima ronda frente a ser seleccionado en la primera ronda. Concretamente, por cada partido adicional, estas posibilidades aumentan en un 16,44%.

Se procede con la evaluación del modelo creado a través de la selección automática de variables.

	X1	Undrafted	X2	X3	X4	X5	X6	X7
X1	8	1	7	1	2	0	1	3
Undrafted	0	8	2	1	2	2	1	2
X2	3	0	3	4	2	2	4	0
X3	0	2	1	2	1	3	2	3
X4	1	0	1	2	2	1	1	1
X5	0	2	0	1	2	0	1	1
X6	0	3	0	2	2	1	5	1
X7	0	1	0	0	0	2	0	2

Tabla 7.13

Accuracy	Kappa
0.2777778	0.1724138

Tabla 7.14

El modelo clasifica correctamente el 27,78% de los datos y da lugar a un índice de Kappa de 0,17 lo que implica una calidad pobre, pero con mejor estadísticos que la del modelo manual.

	Sensitivity	Specificity
Class:X1	0.6666667	0.8437500
Class: Undrafted	0.4705882	0.8901099
Class: X2	0.2142857	0.8404255
Class: X3	0.1538462	0.8736842
Class: X4	0.1538462	0.9263158
Class: X5	0.0000000	0.9278351
Class: X6	0.3333333	0.9032258
Class: X7	0.1538462	0.9684211

Tabla 7.15

- Curva ROC pairwise: `Multi-class area under the curve: 0.7049`
  El AUC medio es de 0.7049 que es menor, aunque prácticamente igual que el AUC del modelo manual.

- Curva ROC one vs all: `0.7039579`

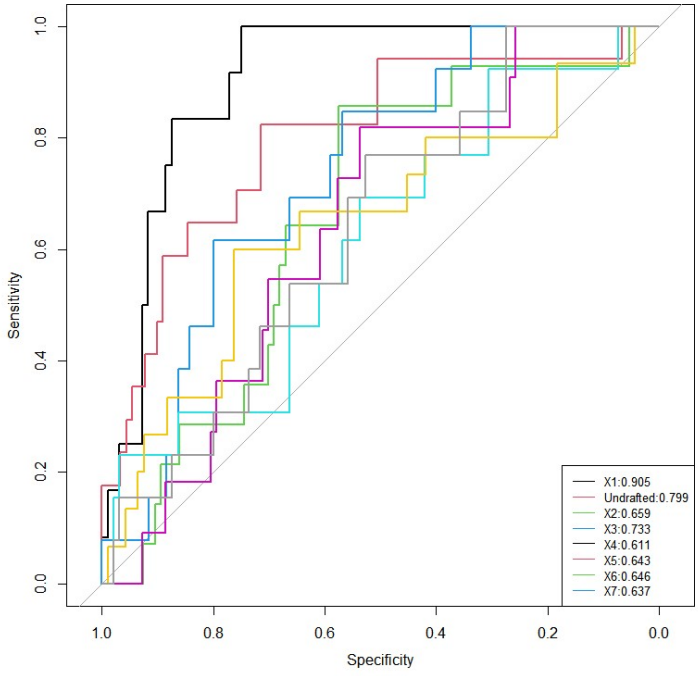

Figura 7.3

La media de los AUC es de 0.7039579, donde X1 es la que mejor se modeliza con un AUC de 0,905 y X4 es la que peor se modeliza con un AUC de 0,611.

En resumen, ambos modelos dan resultados muy similares por lo que me quedaría con el más simple. El modelo es calidad aceptable pero pobre y deja que desear.

7.2. REGRESIÓN LOGÍSTICA ORDINAL

La distinción entre la regresión múltiple multinomial y la ordinal es que las categorías de las rondas de la NFL se ordenarán de manera natural. En este caso irán de X1 a X7 y la última será Undrafted ya que significa que no ha sido seleccionado en ninguna de las rondas.

La semilla 2025 se utilizará también en este caso para entrenamiento-prueba. Se construye un primero modelo con todas las variables independientes, da como resultado un modelo de 21 parámetros.

Se recurre a la función de "Stargazer" en R, la cual se representa en la siguiente tabla:

	Dependent variable:
	Round
First_Division1	-1.995***
	p = 0.0001
Power_Conference1	-0.980***
	p = 0.00001
Ht	-0.005
	p = 0.929
Wt	-0.015
	p = 0.103
Fourty_yd	0.089***
	p = 0.000
Bench	-0.035
	p = 0.197
Vertical	-0.064
	p = 0.118
Broad_Jump	-0.028
	p = 0.190
Shuttle	-0.006
	p = 0.431
Three_Cone	0.006
	p = 0.317
Games	0.030**
	p = 0.017
Yards_Receptions	-0.045
	p = 0.333
TD	-0.085***
	p = 0.00000
Receptions_Game	-0.292***
	p = 0.006
Observations	442
Note:	*p<0.1; **p<0.05; ***p<0.01

Tabla 7.16

Se cuentan con menos parámetros que en la multinomial lo que hará que el modelo sea más rígido.

Las variables First_Division, Power_Conference, Fourty_yd, Games, TD y Receptions_Game tienen parámetros significativos.

Ahora se hará un ANOVA de tipo II para ver la significación conjunta de los parámetros y confirmar que las 6 variables anteriores deben de estar incluidas en el modelo.

```
Analysis of Deviance Table (Type II tests)

Response: Round
                 LR Chisq Df Pr(>Chisq)
First_Division     15.991  1  6.363e-05 ***
Power_Conference   21.495  1  3.548e-06 ***
Ht                  0.008  1   0.930978
Wt                  2.401  1   0.121250
Fourty_yd          63.246  1  1.824e-15 ***
Bench               1.668  1   0.196460
Vertical            2.271  1   0.131775
Broad_Jump          1.689  1   0.193720
Shuttle             0.640  1   0.423592
Three_Cone          1.097  1   0.294913
Games               5.499  1   0.019031 *
Yards_Receptions    0.940  1   0.332222
TD                 25.078  1  5.507e-07 ***
Receptions_Game     7.712  1   0.005485 **
```

Tabla 7.17

Finalmente serán incluidas en el modelo todas las variables mencionadas anteriormente.

```
formula = Round ~ First_Division + Power_Conference + Fourty_yd + Games + TD +
Receptions_Game
```

Crea un modelo de 13 parámetros donde se volverá a comprobar que las variables sean conjuntamente significativas.

```
Analysis of Deviance Table (Type II tests)

Response: Round
                 LR Chisq Df Pr(>Chisq)
First_Division     10.228  1  0.0013833 **

Power_Conference   24.465  1  7.566e-07 ***
Fourty_yd          87.377  1  < 2.2e-16 ***
Games              14.360  1  0.0001509 ***
TD                 44.117  1  3.094e-11 ***
Receptions_Game     3.442  1  0.0635515 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Tabla 7.18

Al crear el modelo reducido, se ve en la tabla 7.18 que Receptions_Game deja de ser conjuntamente significativo, por lo que será eliminada y se hará el modelo con las cinco variables restantes. El nuevo modelo cuenta con 12 parámetros.

```
Analysis of Deviance Table (Type II tests)

Response: Round
                 LR Chisq Df Pr(>Chisq)
First_Division     10.545  1   0.001165 **
Power_Conference   22.634  1  1.959e-06 ***
Fourty_yd          83.894  1  < 2.2e-16 ***
Games              26.458  1  2.694e-07 ***
TD                 95.390  1  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Tabla 7.19

```
================================================
                        Dependent variable:
                    ----------------------------
                               Round
------------------------------------------------
First_Division1              -1.518***
                              (0.468)

Power_Conference1            -0.969***
                              (0.205)

Fourty_yd                    0.086***
                              (0.001)

Games                        0.050***
                              (0.010)

TD                           -0.109***
                              (0.012)

------------------------------------------------
Observations                    442
================================================
```

Tabla 7.20

First_Division1	Power_Conference1	Fourty_yd	Games	TD
0.2191173	0.3804245	1.089947	1.051066	0.897537

Tabla 7.21

La interpretación de los odds-ratio de la variable multinomial mostrados en la tabla 7.21 es la siguiente:

- **First_Division**: el odds-ratio es inferior a uno, por lo que los jugadores que juegan en primera división tienen un 78% menos de posibilidades de ser seleccionados más tarde que aquellos que no pertenecen a primera división.
- **Power_Conference**: el odds-ratio es inferior a uno, por lo que los jugadores que juegan en una de las cuatro conferencias de élite tienen un 62% menos de posibilidades de ser seleccionados más tarde que aquellos que no pertenecen a una conferencia de élite.
- **Fourty_yd:** El odds-ratio es superior a uno, por lo que, por cada milisegundo adicional, las posibilidades de que un jugador sea seleccionado más tarde aumentan en un 8,99%.
- **Games:** El odds-ratio es superior a uno, por lo que, por cada partido adicional, las posibilidades de que un jugador sea seleccionado más tarde aumentan en un 5,11%.
- **TD:** el odds-ratio es inferior a uno, por lo que, por cada touchdown adicional, las posibilidades de que un jugador sea seleccionado más tarde disminuyen en un 10%.

A continuación, se hará una evaluación del modelo.

	X1	X2	X3	X4	X5	X6	X7	Undrafted
X1	6	4	1	1	1	1	0	1
X2	5	5	3	1	0	2	1	0
X3	1	1	3	5	3	4	2	0
X4	0	0	1	0	1	0	0	0
X5	0	0	0	0	0	0	0	0
X6	0	3	3	4	1	6	3	5
X7	0	0	0	0	0	0	0	0
Undrafted	0	1	2	2	5	2	7	11

Tabla 7.22

Es de notar que tanto X5 como X7 no tiene predicción alguna.

Accuracy	Kappa
0.2870370	0.1756542

Tabla 7.23

	value	ASE	z	Pr(>|z|)
Unweighted	0.1757	0.04627	3.568	3.592e-04
Weighted	0.4262	0.05814	7.181	6.918e-13

Tabla 7.24

```
Weights:
          [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]      [,8]
[1,] 1.0000000 0.8571429 0.7142857 0.5714286 0.4285714 0.2857143 0.1428571 0.0000000
[2,] 0.8571429 1.0000000 0.8571429 0.7142857 0.5714286 0.4285714 0.2857143 0.1428571
[3,] 0.7142857 0.8571429 1.0000000 0.8571429 0.7142857 0.5714286 0.4285714 0.2857143
[4,] 0.5714286 0.7142857 0.8571429 1.0000000 0.8571429 0.7142857 0.5714286 0.4285714
[5,] 0.4285714 0.5714286 0.7142857 0.8571429 1.0000000 0.8571429 0.7142857 0.5714286
[6,] 0.2857143 0.4285714 0.5714286 0.7142857 0.8571429 1.0000000 0.8571429 0.7142857
[7,] 0.1428571 0.2857143 0.4285714 0.5714286 0.7142857 0.8571429 1.0000000 0.8571429
[8,] 0.0000000 0.1428571 0.2857143 0.4285714 0.5714286 0.7142857 0.8571429 1.0000000
```

Tabla 7.25

Se puede observar en la tabla 7.24 que el índice de capa clásico (Unweighted) y el índice de Kappa ajustado (Weighted). Ambos En la tbala 7.25 es posible ver los pesos aplicados y como se reducen los considerados aciertos al equivocarse en los niveles. En el caso de equivocarte un nivel, se considera 0,8571429 de acierto y 0,1428571 de error.

La diferencia entre los dos índices de Kappa es notable, pero comprobando la matriz de confusión se puede ver que esto tiene sentido. Los contrastes son significativamente distintos de 0, por lo que la clasificación se puede considerar mejor que la que se obtendría al azar.

	Sensitivity	Specificity
Class:X1	0.5000000	0.9062500
Class: X2	0.3571429	0.8723404
Class: X3	0.2307692	0.8315789
Class: X4	0.0000000	0.9789474
Class: X5	0.0000000	1.00
Class: X6	0.4000000	0.7849462
Class: X7	0.0000000	1.00
Class: Undrafted	0.5882353	0.7912088

Tabla 7.26

Ahora se crearán modelos utilizando los procesos de selección automática de variables para poder comparar estos con el modelo manual.

```
14 15 14 15 14 15
```

Los seis procesos dan como resultado dos modelos distintos, uno con 14 parámetros y otro con 15.

```
formula = Round ~ Fourty_yd + TD + Power_Conference + Games + First_Division +
Vertical + Wt
```

```
formula = Round ~ Fourty_yd + TD + Power_Conference + Games + First_Division +
Vertical + Wt + Receptions_Game
```

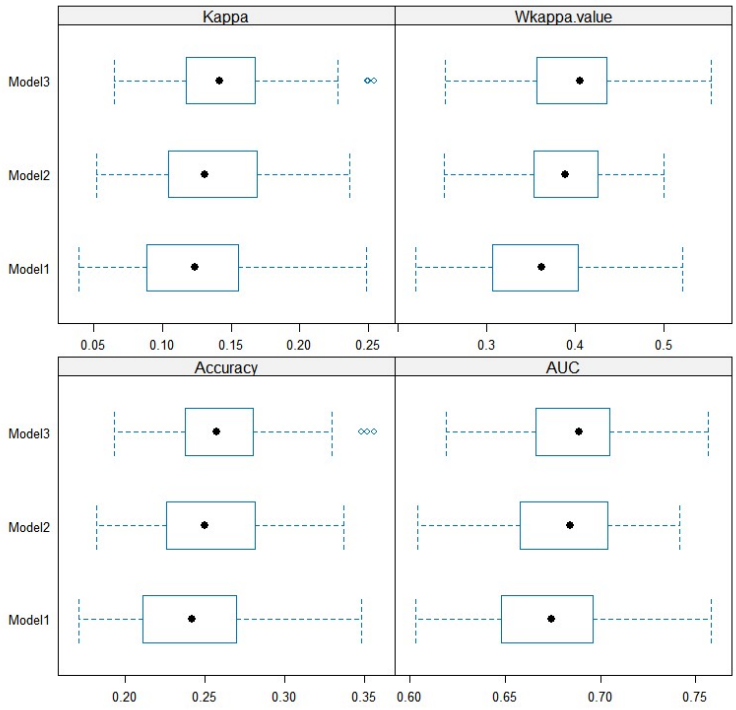

Figura 7.4

El Model1 representa el modelo creado manualmente, Model2 representa el modelo de 14 parámetros y el Model3 representa el modelo de 15 parámetros.

AUC

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
Model1	0.5994965	0.6483088	0.6732774	0.6719049	0.6946038	0.7572427	0
Model2	0.6088721	0.6575651	0.6840073	0.6812439	0.7022436	0.7441212	0
Model3	0.6171510	0.6652008	0.6875822	0.6865963	0.7039932	0.7587693	0

Kappa

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
Model1	0.03832335	0.09152802	0.1242064	0.1224596	0.1538364	0.2486172	0
Model2	0.05178090	0.10593157	0.1310969	0.1352709	0.1673122	0.2361071	0
Model3	0.06453062	0.11799244	0.1443575	0.1470627	0.1695892	0.2681248	0

Accuracy

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
Model1	0.1704545	0.2134831	0.2415688	0.2417406	0.2696629	0.3483146	0
Model2	0.1818182	0.2272727	0.2485955	0.2521984	0.2808989	0.3370787	0
Model3	0.1931818	0.2386364	0.2598953	0.2620432	0.2808989	0.3678161	0

Wkappa.value

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
Model1	0.2179262	0.3075768	0.3604795	0.3565014	0.4033152	0.5082873	0
Model2	0.2525173	0.3496777	0.3866491	0.3847979	0.4205614	0.5050468	0
Model3	0.2511327	0.3613176	0.4027059	0.3974198	0.4316161	0.5657508	0

Tabla 7.27

Se puede ver en la figura 7.4 que el Model3 es el modelo que mejores estadísticos muestra por lo que se seleccionará para estudiar su interpretación y evaluación.

Se procede con la prueba individual sobre los parámetros y el ANOVA:

	Dependent variable:
	Round
Fourty_yd	1.097***
	p = 0.000
TD	0.910***
	p = 0.000
Power_Conference1	0.400***
	p = 0.00002

```
Games                        1.039***
                             p = 0.001

First_Division1              0.150***
                            p = 0.0002

Vertical                     0.897***
                            p = 0.0002

Wt                           0.980***
                             p = 0.001

Receptions_Game               0.819**
                             p = 0.015

-----------------------------------------------
Observations                    442
===============================================
Note:              *p<0.1; **p<0.05; ***p<0.01
```

Tabla 7.28

Todos los parámetros son significativos.

```
Analysis of Deviance Table (Type II tests)

Response: Round
                 LR Chisq Df Pr(>Chisq)
Fourty_yd          71.890  1  < 2.2e-16 ***
TD                 45.747  1  1.346e-11 ***
Power_Conference   19.359  1  1.083e-05 ***
Games              12.145  1  0.0004922 ***
First_Division     14.899  1  0.0001134 ***
Vertical           10.938  1  0.0009421 ***
Wt                 11.449  1  0.0007153 ***
Receptions_Game     5.975  1  0.0145095 *
---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Tabla 7.29

Todas las variables son conjuntamente significativas así que se miran los odds-ratio: se interpretarán las tres variables que más contribuyen al modelo.

Fourty_yd	TD	Power_Conference1	Games	First_Division1	Vertical	Wt	Receptions_Game
1.096621	0.909531	0.4003199	1.038804	0.1495442	0.897357	0.979549	0.8187732

Tabla 7.30

- **Tiempo en correr cuarenta yardas (Fourty_yd):** El odds-ratio es superior a uno, por lo que, por cada milisegundo adicional, las posibilidades de que un jugador sea seleccionado más tarde aumentan en un 9,66%.
- **Número totales de touchdowns (TD):** el odds-ratio es inferior a uno, por lo que, por cada touchdown adicional, las posibilidades de que un jugador sea seleccionado más tarde disminuyen en un 9,05%.
- **Juega en una conferencia de élite (Power_Conference):** el odds-ratio es inferior a uno, por lo que los jugadores que juegan en una de las cuatro conferencias de élite tienen un 60% menos de posibilidades de ser seleccionados más tarde que aquellos que no pertenecen a una conferencia de élite.

Se procede a evaluar el modelo:

	X1	X2	X3	X4	X5	X6	X7	Undrafted
X1	7	6	1	1	1	1	0	1
X2	5	1	4	1	0	0	1	0
X3	0	4	3	4	1	5	0	1
X4	0	0	2	0	0	3	0	0
X5	0	0	0	0	0	0	0	0
X6	0	2	1	5	5	3	5	1
X7	0	0	0	0	0	0	0	0
Undrafted	0	1	2	2	4	3	7	14

Tabla 7.31

Accuracy	Kappa
0.2592593	0.1437066

Tabla 7.32

	value	ASE	z	Pr(>|z|)
Unweighted	0.1437	0.04279	2.858	4.259e-03
Weighted	0.4781	0.05531	8.227	1.927e-16

Tabla 7.33

Los pesos aplicados son los mismos que en el modelo manual. La tasa de acierto disminuye a 25,93% pero el índice de Kappa ponderada aumenta a 0,4781. Por lo que la calidad del modelo es media.

	Sensitivity	Specificity
Class:X1	0.58333333	0.8854167
Class: X2	0.07142857	0.8829787
Class: X3	0.23076923	0.8421053
Class: X4	0.00000000	0.9473684
Class: X5	0.00000000	1.00
Class: X6	0.20000000	0.7956989
Class: X7	0.00000000	1.00
Class: Undrafted	0.82352941	0.7912088

Tabla 7.34

Aunque el modelo creado por la selección automática de variable tenga una menor precisión, será seleccionada por encima del modelo manual debido a que tiene una kappa ponderada mayor y este es el indicador clave en cuanto a los modelos logísticos ordinales.

8. CONCLUSIONES

El objetivo principal de este trabajo es lograr construir un modelo predictivo estadístico que estime la probabilidad de que un receptor universitario sea seleccionado en una de las siete rondas del draft de la NFL.

El desarrollo se ha llevado a cabo a través de dos algoritmos de Machine Learning para encontrar el mejor modelo predictivo. Primero se han creado algunas variables a partir de las existentes de la base de datos, después se han tratado los valores perdidos a través de la imputación utilizando knn (Los k vecinos cercanos). Una vez consolidado los datos, se ha detectado y eliminado la multicolinealidad para proceder a la construcción de modelos. El primer modelo se ha construido manualmente mientras que el segundo modelo ha sido construido por el mejor evaluado de los métodos automáticos de selección de variables.

Se ha observado que las variables más asociadas a la ronda en la que es escogido un jugador son el estar jugando en la primera división universitaria, el pertenecer a una de las cuatro conferencias de élite, la carrera de cuarenta yardas, el número total de partidos jugados a nivel universitario y el número de touchdowns totales que consigue el jugador a nivel universitario. Todas estas variables aparecen en los cuatro modelos finales.

La variable referente a la carrera de las cuarenta yardas tiene un odds-ratio mayor que uno en los cuatro modelos por lo que se concluye que cuanto más rápido hace un individuo la prueba más posibilidades existen de que sea seleccionado antes.

La variable referente a si el jugador pertenece a primera división tiene un odds-ratio considerablemente menor que uno por lo que los jugadores que juegan en primera división tienen mucha más probabilidad de ser seleccionados antes que los que no juegan en primera división. Esta misma lógica es aplicada a la variable vinculada a si un jugador compite en una conferencia de élite, donde se comparan los jugadores que pertenecen a una de las cuatro conferencias de élite y a los que no.

La variable games tiene un odds-ratio mayor que uno por lo que cada partido de menos que juegue un receptor aumenta su probabilidad de ser escogido más pronto. Una teoría de por qué se da este fenómeno es que los jugadores pueden presentarse al draft en su tercer año de carrera. Esto conlleva a que los jugadores que se presentan de forma más prematura al proceso suelen tener más talento y más oportunidades de ser seleccionados en las primeras rondas.

La variable TD tiene un odds-ratio menor que uno por lo que cada touchdown de más aumenta la probabilidad de que sea seleccionado más temprano. El touchdown es la máxima forma de puntuar en el fútbol americano por lo que es algo a lo que se le da importancia.

Las evaluaciones de los modelos nos muestran que en el caso de este trabajo utilizar los modelos de regresión logística ordinal es óptimo en comparación con aplicar los modelos de regresión logística multinomial. Esto concuerda bastante debido a que la variable dependiente sigue un orden natural.

El modelo multinomial construido a través de los métodos automáticos de selección de variables obtiene mejor evaluación que el construido manualmente con un aumento del 3,7% en la tasa de aciertos y una mejora de 0,0414 en el índice de kappa, con unas medias de ROC equivalentes, pero contando con 14 parámetros menos. Aun así, dicho modelo tiene un índice de kappa de 0,1724 < 0,20 por lo que se concluye que la calidad del modelo es pobre.

El modelo multinomial construido a través de los métodos automáticos de selección de variables también obtiene mejor evaluación que el construido manualmente. La tasa de aciertos disminuye un 2,77% y tiene un empeoramiento de 0,0319 el en índice de kappa, pero su kappa ponderada se encuentra 0,0519 puntos por encima. Esto demuestra la importancia de los pesos en la distancia del fallo. Este modelo obtiene un índice de kappa ponderada de 0,4781 por lo que se concluye que la calidad del modelo es moderada.

En todos los modelos se puede observar que tienen alta especificidad, lo que significa que los modelos identifican correctamente los casos que no pertenecen a la ronda real. Otra observación es la diferencia que hay en cuanto a la sensibilidad entre las categorías de Round. En el modelo ordinal creado por la selección automática de variables tiene como resultados 0% de sensibilidad en X4, X5 y X7 mientras que otras categorías como X1 y no ser seleccionado tienen 50% y 82,35% de sensibilidad respectivamente.

8.1. PROPUESTAS DE MEJORA DEL ESTUDIO

Dado que el modelo tiene una calidad moderada y una tasa de acierto por debajo del 30%, se van a proponer unas propuestas de mejora del estudia que se podrían implementar en el caso de querer profundizar más en este trabajo.

La primera propuesta de mejora es aumentar el número de observaciones de la base de datos. Actualmente se cuenta con 550 jugadores lo cual es una muestra relativamente pequeña. El aumento del número de observaciones conlleva a aumento de observaciones tanto en entrenamiento como en prueba, lo que ayudaría a mejorar la precisión del modelo y a estabilizar la sensibilidad.

Otra propuesta de mejora sería añadir más variables independientes a el modelo. Algunas variables como el IMC de los jugadores o si han jugado en los playoffs universitarios (BCS schools) son ejemplos de variables han sido utilizadas tanto por Mulholland y Jensen (2014) como por Fenn y Berri (2018).

Utilizar técnicas de tratamiento de valores perdidos distintos, por ejemplo, en algunos estudios se recurre a la imputación múltiple (Murran, 2018; Fenn y Berri, 2018) y en otros a excluir casos por pares (Teramoto et al. 2016)

Finalmente, se pueden añadir al estudio más técnicas de Machine Learning como los árboles de clasificación. Los árboles tienen la ventaja de ser más robustos, no necesitar que las variables independientes cualitativas sean convertidas en variables dummy y manejar los datos ausentes de manera eficiente. (Calviño, 2023)

9. BIBLIOGRAFÍA

Apuntes de la asignatura Técnicas de segmentación y tratamientos de encuestas de la profesora Aída Calviño.

Dhar, A. (2011): *Drafting NFL Wide Receivers: Hit or miss?* Disponible en: https://www.stat.berkeley.edu/~aldous/157/Old_Projects/Amrit_Dhar.pdf.

Fenn, A.J. y Berri, D (2018): *Drafting a Successful Wide Receiver in the NFL-Hail Mary?* Disponible en: https://jvlone.com/sportsdocs/DraftingWideReceiverinNFL2018.pdf.

Google User Content. (n.d.): *Dimensiones de un campo de fútbol americano.* Disponible en: https://lh7-us.googleusercontent.com/BXGAkLrCx9f_t8gqw9jxJ3BETDURxzY6OKNMxtZpHaOZbwLQ3LIXiz6cYo-n2Cy34SVehTEbYqNAK_fX_PmRfTCicgCtRqzvSNVSLKSi1rZKo6DDLamV1PsZ0SPSE9gZm-MqQFl3W5JYQcIkDHwW3OI [Consulta: 27 de octubre de 2024].

IBM (n.d.a.): *KNN*. Disponible en: https://www.ibm.com/mx-es/topics/knn.

IBM (n.d.b.): *Overfitting.* Disponible en: https://www.ibm.com/es-es/topics/overfitting

Luttrel, R., Emerick, S.F. y Wallace, A. (2022): *Digital Strategies: Data-Driven Public Relations, Marketing, and Advertising,* Nueva York, Oxford University Press.

Mulholland, J. y Jensen, S.T. (2014): *Predicting the draft and career success of tight ends in the National Football League.* Disponible en: https://www.degruyter.com/document/doi/10.1515/jqas-2013-0134/html.

Murran, D. (2018): *Ahead of the Curve, Analysis of the NFL Draft: Technical Report.* Disponible en: https://norma.ncirl.ie/3454/1/danielmurran.pdf.

NFL. (n.d.): *NFL Draft.* Disponible en: https://www.nfl.com/draft/ [Consulta: 15 de octubre 2024].

NFL combine results. (n.d.): *NFL combine results.* Disponible en: https://nflcombineresults.com/ [Consulta: 15 de octubre 2024].

NFL Operations. (n.d.a.): *Glosario de términos*. Disponible en: https://operations.nfl.com/es/aprenda-el-juego/lo-basico-de-la-nfl/glosario-de-terminos/ [Consulta: 29 de octubre 2024].

NFL Operations. (n.d.b.): *Creación del calendario de la nfl.* Disponible en: https://operations.nfl.com/es/dia-del-partido/horario-de-la-nfl/creacion-del-calendario-de-la-nfl/ [Consulta: 29 de octubre 2024].

NFL Operations. (n.d.c.): *Super Bowl LVIII.* Disponible en: https://operations.nfl.com/updates/the-game/super-bowl-lviii-is-most-watched-telecast-in-history/#:~:text=Record%2DSetting%20Viewership,previous%20record%20(115.1%20million [Consulta: 29 de octubre 2024].

NFL Operations. (n.d.d.): *NFL Draft.* Disponible en: https://operations.nfl.com/es/paso-a-la-nfl/el-draft-de-la-nfl/las-reglas-para-el-draft/ [Consulta: 29 de octubre 2024].

Pineda San Juan, Silvia (2024): *Depuración de Datos – TFG Estadística.* Disponible en: https://silvia-pineda.gitbook.io/depuracion-de-datos-tfg3.

Pro Football Reference. (n.d.): *Draft.* Disponible en: https://www.pro-football-reference.com/draft/ [Consulta: 15 de octubre 2024].

R Documentation (n.d.): *KnnImputation.* Disponible en: https://www.rdocumentation.org/packages/DMwR/versions/0.4.1/topics/knnImputation

Sporting news (n.d.): *Playoff Bracket*. Disponible en: https://library.sportingnews.com/styles/crop_style_16_9_desktop_webp/s3/2023-12/GFX-1240%20NFL%20Playoff%20Bracket.jpg.webp [Consulta: 29 de octubre 2024].

Sports reference CFB. (n.d.): *CFB.* Disponible en: https://www.sports-reference.com/cfb/ [Consulta: 15 de octubre 2024].

Teramoto, M., Cross, C.L. y Willick, S.E. (2016): *Predictive Value of National Football League Scouting Combine on Future Performance of Running Backs and Wide Receivers.* Disponible en: https://journals.lww.com/nsca-jscr/fulltext/2016/05000/Predictive_Value_of_National_Football_League.25.aspx?platform=hootsuite.

UNIVERSIDAD COMPLUTENSE MADRID

Declaración Responsable sobre Autoría y Uso Ético de Herramientas de Inteligencia Artificial (IA)

Yo, JACOBS MARITZIA, DAVID

Con DNI/NIE/PASAPORTE:

declaro de manera responsable que el/la presente:

- ☐ **Trabajo de Fin de Grado (TFG)**
- ☐ Trabajo de Fin de Máster (TFM)
- ☐ Tesis Doctoral

Titulado/a

ESTADÍSTICA APLICADA

es el resultado de mi trabajo intelectual personal y creativo, y ha sido elaborado de acuerdo con los principios éticos y las normas de integridad vigentes en la comunidad académica y, más específicamente, en la Universidad Complutense de Madrid.

Soy, pues, autor del material aquí incluido y, cuando no ha sido así y he tomado el material de otra fuente, lo he citado o bien he declarado su procedencia de forma clara -incluidas, en su caso, herramientas de inteligencia artificial-. Las ideas y aportaciones principales incluidas en este trabajo, y que acreditan la adquisición de competencias, son mías y no proceden de otras fuentes o han sido reescritas usando material de otras fuentes.

Asimismo, aseguro que los datos y recursos utilizados son legítimos, verificables y han sido obtenidos de fuentes confiables y autorizadas. Además, he tomado medidas para garantizar la confidencialidad y privacidad de los datos utilizados, evitando cualquier tipo de sesgo o discriminación injusta en el tratamiento de la información.

En Madrid a 19/02/2025

DAVID THOMAS JACOBS MARITZIA