COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

MÁSTER UNIVERSITARIO EN INGENIERÍA INDUSTRIAL

TRABAJO FIN DE GRADO

MODELO PREDICTIVO DE COMPETICIONES DE PATINAJE SOBRE HIELO

Autor: Laura Gil Martínez

Director: Dr. Antonio García y Garmendia
Madrid

Julio 2023

Declaro bajo mi responsabilidad que el Proyecto presentado de titulo

Modelo predictivo de competiciones de patinaje sobre hielo

en la ETS de Ingeniería - ICAI de la Universidad Pontificia de Comillas en elcurso académico 2022/23 ha sido realizado por mí, se trata de una copia inédita y aún no se ha utilizado para otros fines.

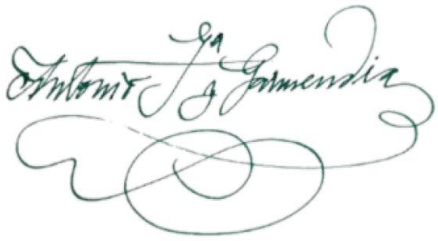
El proyecto no está plagiado y la información tomada de otros documentosse encuentra referenciada adecuadamente.

Fdo.: Laura Gil Martinez Fecha: 28/06/2023

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Fdo.: Dr. Antonio García y Garmendia Fecha: 28/06/2023



MODELO PREDICTIVO DE COMPETICIONES DE PATINAJE SOBRE HIELO

Autor: Laura Gil Martínez
Director: Dr. García y Garmendia, Antonio
Entidad Colaboradora: ICAI – Universidad Pontificia Comillas

RESUMEN DEL PROYECTO

Este trabajo se centra en la aplicación de herramientas estadísticas y la inteligencia artificial para la creación de modelos predictivos en el ámbito de las competiciones de patinaje artístico. Durante el proyecto emplean tanto la regresión lineal múltiple como como las máquinas de vectores de soporte para la creación de distintos modelos, con el objetivo de comparar los resultados obtenidos a partir de estos y seleccionar el método óptimo.

Palabras clave: patinaje artístico, regresión lineal múltiple, máquinas de vectores de soporte.

En primer lugar, se hace un breve y clara introducción del patinaje artístico sobre hielo. Este se trata de un deporte muy complejo en el que muchos factores tanto objetivos como subjetivos que pueden alterar las puntuaciones finales, dando lugar a resultados inesperados. En un intento de evitar estas situaciones se pretende crear un modelo capaz de predecir las puntuaciones finales de las patinadoras tanto en el programa corto como en el programa largo, las dos partes que componen una competición de patinaje artístico, sin tener en cuenta la dificultad de los elementos técnicos, ya que se pretende considerar solo aquellos factores que influyen de manera menos obvia para los espectadores menos expertos.

Una vez explicado el contexto del proyecto y el motivo de su creación, así como su adecuación a los Objetivos de Desarrollo Sostenible, se lleva a cabo un estudio de la literatura existente acerca de los modelos predictivos en el mundo del deporte y de las actividades que combinen elementos objetivos y subjetivos. Durante este proceso se han identificado dos técnicas interesantes en este contexto, la regresión lineal múltiple y las máquinas de vectores de soporte. Para confirmar su utilidad en las predicciones de patinaje sobre hielo se decide crear distintos modelos empleando ambas técnicas.

El primer paso que se lleva a cabo en el diseño de ambos modelos es la selección de variables que puedan influir en la puntuación final de una patinadora, como se ha mencionado previamente, sin tener en cuenta aquellas que pueden resultar más obvias como la dificultad de los elementos técnicos. Una vez seleccionadas, se crea una base de datos que contenga la mayor información posible acerca de estas variables, para a continuación, poder formular sus hipótesis correspondientes.

Una vez completada esta parte del trabajo se realiza una explicación exhaustiva de ambas técnicas, detallando tanto del fundamento teórico como la creación de los modelos. Dedicando un capítulo a la regresión lineal múltiple, la cual emplea el método

de los mínimos cuadrados, y otro a la aplicación de las máquinas de vectores de soporte a un modelo de regresión y al kernel utilizado, necesario durante el desarrollo de esta técnica.

A continuación, se comprueba que los resultados obtenidos por ambos métodos muestran modelos de gran calidad. Para ello se emplean el coeficiente de regresión y la raíz del error cuadrático medio, medidas utilizadas para comprobar la utilidad del modelo predictivo creado por la regresión lineal múltiple y las máquinas de vectores de soporte y que se muestran en las siguientes tablas, respectivamente:

Medida	**Programa corto**	**Programa largo**
R^2 entrenamiento	0,8551	0,857
RMSE entrenamiento	0.1507803	0.1371814
R^2 prueba	0.8431	0.8488
RMSE prueba	0.1516	0.1387

Medida	**Programa corto**	**Programa largo**
R^2 entrenamiento	0.999	0.9999
R^2 prueba	0.979	0.9648
RMSE entrenamiento	0.0009	7.319e-05
RMSE prueba	0.0210	0.0318

A partir de esta información es posible comparar los modelos creados, pudiéndose así seleccionar la herramienta óptima. En este caso, muestran mejores resultados los modelos creados por las máquinas de vectores de soporte, seleccionándose este método como el óptimo a la hora de diseñar modelos de predicción para las competiciones de patinaje sobre hielo.

El trabajo continúa con la aplicación del modelo creado a un ejemple práctico, concretamente a las diez mejores patinadoras participantes en los Juegos Olímpicos de Pekín en 2022. A partir de este se comprueba la precisión de este, comparando sus resultados con los obtenidos por las patinadoras en la vida real.

Por último, se presentan posibles campos de aplicación para el modelo creado. Entre ellos destaca el mundo de las apuestas deportivas, el cual se explica en mayor profundidad. Este se trata de un mundo muy lucrativo y con un amplio abanico de aplicaciones para el modelo creado en este proyecto, las cuales se desarrollan brevemente.

PREDICTIVE MODEL FOR ICE SKATING COMPETITIONS

Author: Laura Gil Martínez

Director: Dr. García y Garmendia, Antonio

Collaborating Institution: ICAI - Pontifical Comillas University

PROJECT SUMMARY

The purpose of this research is to develop predictive models for figure skating contests using statistical techniques and artificial intelligence. The project uses support vector machines and multiple linear regression to build various models with the goal of comparing the output from each of them and choosing the best approach.

Keywords: figure skating, multiple linear regression, support vector machines.

First, a brief and complete introduction to ice figure skating is given. This is a very complex sport where a variety of objective and subjective elements may have an unexpected impact on the final scores. The objective is to develop a model that can forecast skaters' final results in both the short program and the free program, which are the two events that make up a figure skating competition, in an effort to prevent these occurrences. The objective is to simply take into account those characteristics that have a less evident impact on less experienced viewers, therefore the complexity of technical parts is not taken into consideration.

Following an overview of the project's context, goal, and alignment with the SDGs, a review of the literature on predictive models used in sports and other activities that integrate objective and subjective components is carried out. Multiple linear regression and support vector machines are two noteworthy methods that are discovered throughout this process. By combining these approaches, various models are developed to demonstrate their efficacy in predicting ice skating results.

The very first step when developing these models is choosing the variables that, as previously said, can affect a skater's ultimate score, omitting those that can be more visible, including the difficulty of technical parts. After these variables are chosen, a database is made with as much data as possible about them in order to develop the associated hypotheses.

After this section of the project is finished, both methodologies are thoroughly explained, including information on how the models were made and their theoretical underpinnings. Multiple linear regression, which employs the least squares method, is covered in one chapter, and the support vector machine application to a regression model and the kernel utilized, which is essential during the development of this methodology, is covered in another chapter.

Next, it is confirmed that both approaches' findings display high-quality models. To test the usefulness of the predictive model produced by multiple linear regression and support vector machines, respectively, the coefficient of regression and the root mean square error are used. The following tables display the findings:

Measure	Short Program	Free program
R^2 training	0,8551	0,857
RMSE trainig	0.1507803	0.1371814
R^2 test	0.8431	0.8488
RMSE test	0.1516	0.1387

Measure	Short Program	Free program
R^2 trainig	0.999	0.9999
R^2test	0.979	0.9648
RMSE training	0.0009	7.319e-05
RMSE test	0.0210	0.0318

The models that have been built can be compared using this data to choose the best tool. Support vector machine models produce better outcomes in this situation, and this method is chosen as the best strategy for developing prediction models for ice skating contests.

The next step in the project is to apply the developed model to a real-world example, in this case, the top ten skaters competing in the 2022 Beijing Olympic Games. By contrasting the model's predictions with the skaters' actual performance, the model's accuracy is confirmed.

Finally, the potential applications for the developed model are discussed. The area of sports betting stands out among them and is discussed in further detail. The model created in this project has a wide range of applications in this extremely lucrative industry, which are briefly reviewed.

Contenido

CAPÍTULO 1: INTRODUCCIÓN

En 1860 el profesor de ballet Jackson Haines decidió añadir elementos de danza a la popular actividad de patinar sobre hielo, surgiendo así la combinación entre mezcla y arte que hoy en día se conoce como patinaje artístico sobre hielo[1]. En poco tiempo el deporte ganó cierta popularidad hasta la creación en 1892 de la Unión Internacional de Patinaje[2], encargada de establecer los primeros reglamentos oficiales[3].

Con los años su popularidad y complejidad ha aumentado hasta convertirse en uno de los deportes más importantes del programa de los Juegos Olímpicos de invierno[4]. Hoy en día el patinaje artístico cuenta oficialmente con cuatro modalidades diferentes[5]:

- Individual: Esta categoría consiste en un único patinador o patinadora realizando un programa al ritmo de la música, en el que se combina tanto técnica como presentación al público[6]. Está dividida a su vez en las subcategorías masculina y femenina[7].

Figura 1: Kamila Valieva en las Olimpiadas de Pekín 2022

Fuente: Dupre, E., et.al. (2022)

[1] Carreño, F. M. (2018)
[2] International Skating Union (ISU)
[3] Carreño, F. M. (2018)
[4] Ibid.
[5] *Figure Skating - International Skating Union*. (s. f.-b)
[6] Riedell. (2023)
[7] Ibid.

Figura 2: Patinador español Javier Fernández

Fuente: Cano, V. (2018)

- Parejas: Se trata de una modalidad en la que un patinador y una patinadora llevan a cabo, de manera sincronizada, elementos similares que en el patinaje individual. Al involucrar dos patinadores es posible incluir elementos como saltos lanzados y elevaciones[8]

Figura 3: Sui Wenjing y Han Cong en las Olimpiadas de Pekín 2022

Fuente: McCarvel, N. (2022)

- Danza: A pesar de involucrar dos patinadores como en la categoría anterior, esta modalidad se centra en la coreografía y pasos en vez de en saltos y piruetas[9]. Es

[8] Ibid.
[9] Ibid.

decir, se evalúa el nivel de precisión y limpieza a la hora de deslizarse por el hielo y transmitir la actuación al público[10].

Figura 4: Tessa Virtue y Scott Moir en las Olimpiadas de PyeongChang 2018

Fuente: Staff, B. (2023)

- Patinaje sincronizado: Esta última es una adaptación de la categoría de danza a un grupo de patinadores[11]. Los dieciséis deben realizar de manera precisa y perfectamente sincronizada todos los elementos exigidos en la competición[12].

[10] Ibid.
[11] Ibid.
[12] Ibid.

Figura 5: Helsinki Rockettes en el mundial de Canadá 2022

Fuente: Jura syncho (2022)

En este trabajo nos centramos en la categoría de patinaje individual femenino, donde las patinadoras se enfrentan en numerosas competiciones tanto nacionales como internacionales[13]. A nivel internacional existen un gran número de eventos, como el campeonato de Europa o el Grand Prix, entre los que destaca el mundial senior[14]. Este se celebra anualmente y enfrenta a las mejores patinadoras de los diferentes países miembros de la ISU[15].

Todas estas competiciones se dividen en dos jornadas:

- Programa corto (PC): En esta primera parte las deportistas deben llevar a cabo ocho elementos determinados, elegidos por la ISU cada año, en un tiempo máximo de 2 minutos y 40 segundos[16]. Estos elementos incluyen tanto piruetas como saltos y pasos y pueden ser ejecutados en el orden en el que la patinadora desee y acompañados por una música de su elección[17].
  Aquellas patinadoras que consiguen la mejor puntuación pasan a la siguiente ronda.
- Programa largo (PL): En esta segunda parte las participantes deben patinar durante unos cuatro minutos, en los que deben mostrar de la mejor forma

[13] Sánchez, E. (2023, 21 mar).
[14] Encyclopedia Britannica (2023).
[15] Ibid.
[16] Ibid.
[17] Ibid.

posible tanto sus habilidades técnicas como artísticas[18]. En este caso no se exigen elementos específicos, sino que se incluyen una serie de pautas para asegurarse programas equilibrados[19]. Estas pautas pueden ser limitaciones como un máximo de seis saltos triples diferentes o cuatro piruetas distintas[20].

Basándose en estos dos programas se decide la puntuación total de las patinadoras[21]. Esta puntuación es decidida por dos grupos de profesionales encargados de analizar y evaluar las diferentes partes involucradas en un programa[22].

- Por un lado, el panel técnico es el encargado de evaluar de forma objetiva la parte más técnica del programa[23]. Su trabajo consiste tanto en identificar los elementos realizados durante los programas como el nivel que les corresponde, para lo que utilizan los extensos reglamentos oficiales publicados por la ISU[24]. Este grupo está formado por cinco integrantes: el controlador técnico, el especialista técnico, el operador de datos, el operador de repetición de videos y el árbitro[25].
- Por otro lado, se encuentra el grupo encargado de evaluar de forma más subjetiva la calidad tanto de los elementos como de la interpretación de la patinadora[26]. Es decir, este panel de jueces está encargado de la parte artística del programa: habilidades de la patinadora, actuación, transiciones entre los distintos elementos del programa, composición del programa e interpretación de la música[27]. Está formado por nueve profesionales entrenados en diferentes seminarios y cursos, procedentes de diferentes nacionalidades[28].

Esta organización permite regular el sistema de puntuación en las competiciones de alto nivel, evitando malentendidos y conflictos[29]. Sin embargo, esto no siempre es así, como se pudo comprobar en el año 2022[30].

Los Juegos Olímpicos de Pekín 2022 se vieron afectados por numerosos escándalos como la sonada descalificación de cinco esquiadoras acusadas de usar trajes no permitidos en la prueba de salto[31].

---

[18] Ibid.
[19] *ISU Judging System - International Skating Union*. (2022).
[20] Ibid.
[21] Encyclopedia Britannica (2023).
[22] Word press (2016)
[23] Ibid.
[24] *ISU Judging System - International Skating Union*. (2022).
[25] Encyclopedia Britannica (2023).
[26] Ibid.
[27] Today, C. B. U. (2022)
[28] Encyclopedia Britannica (2023).
[29] Ibid.
[30] No se trata de los primeros Juegos Olímpicos en los que el mundo del patinaje es el centro de atención. En 1994 tuvo lugar el mediático conflicto entre Tonya Harding y Nancy Kerrigan.
[31] Jaimes, K. (2022).

Figura 6: Katharina ALthanus, una de las esquiadoras descalificadas en Pekín 2022

Fuente: Agencia AFP. (2022)

Entre estos sucesos recibió especial atención el equipo de patinaje artístico sobre hielo ruso, el cual se vio involucrado en el sonado escándalo de dopaje relacionado con la joven patinadora Kamila Valieva[32].

Sin embargo, la historia que más nos llama la atención es el conflicto creado entre las dos compañeras, Anna Shcherbakova y Alexandra Trúsova, entrenadas por Eteri Tutberidze. Ambas patinadoras mostraron una gran técnica y elegancia en el hielo obteniendo el primer y segundo puesto, algo que terminó suponiendo un drama[33].

Durante la segunda parte de la competición Trúsova fue capaz de aterrizar cinco cuádruples[34] en un mismo programa, convirtiéndose en la primera mujer en la historia en conseguir este logro[35]. Esto le permitió alcanzar una puntuación de 177.13 en el programa largo, su mejor marca hasta la fecha[36].

[32] Martín, A.(2022).
[33] Dávila, J. (2022)
[34] Salto en el que la patinadora realiza cuatro rotaciones en el aire, elemento poco común en la categoría femenina debido a la fuerza y altura necesarias para llevarlo a cabo.
[35] Dávila, J. (2022).
[36] Ibid.

Figura 7: Alexandra Trusova en los Juegos Olímpicos de Pekín 2022

Fuente: The Associated Press (2022)

Sin embargo, aunque este logro permitió a Alexandra obtener el primer puesto en la segunda parte de la competición, sus 251.73 puntos no fueron suficientes para superar a su rival en la competición global, con una puntuación total de 255.95[37]. Esto llevó a Trusova a desesperarse y acabar el evento entre lágrimas, lo cual afectó enormemente al mundo del patinaje artístico[38].

A raíz de estos resultados, a los fans del patinaje artístico nos surgió la siguiente duda: ¿Cómo una patinadora puede terminar en segunda posición después de una actuación histórica y ejecutando los elementos más complejos vistos en la competición? La respuesta a esta pregunta, según los expertos[39], es la importancia de la parte artística[40].

La superioridad técnica demostrada por Trusova no fue suficiente para compensar las carencias artísticas que presentaban sus programas[41]. Esto llevó a Anna a causar mejor impresión en los jueces, logrando así el primer puesto[42].

Toda esta situación nos lleva a plantearnos la importancia de la subjetividad en un deporte con tantos elementos artísticos como el patinaje sobre hielo. Es decir, como la

[37] The Associated Press (2022).
[38] Dávila, J. (2022).
[39] Expertos como Pedro Lamelas, director de Hielo Español.
[40] El Financiero (2022).
[41] Ibid.
[42] Ibid.

parte menos regulada de las competiciones puede terminar siendo decisiva en los resultados.

A partir de esto surge la idea de crear modelos de regresión capaces de predecir las puntuaciones de las patinadoras en los dos programas que componen una competición, basándonos en variables no relacionadas con la complejidad de los elementos ejecutados. Surgiendo así un proyecto con los siguientes objetivos:

- Estudio de las técnicas existentes en la actualidad relacionadas con las predicciones en el mundo deportivo, así como su posible adaptación al patinaje artístico sobre hielo.
- Uso de aquellas técnicas que se consideren adecuadas para crear diferentes modelos. Pudiéndose comparar sus resultados posteriormente con la idea de seleccionar un modelo óptimo.
- Selección de variables que puedan impactar de manera significativa a la puntuación de las patinadoras, sin tener en cuenta la dificultad técnica de los elementos del programa.
- Recopilación y filtración necesaria para crear una base de datos que resulte útil a la hora de diseñar el modelo y comprobar su eficacia.
- Creación de un modelo preciso y con capacidad de adaptación a nuevas bases de datos, para así poderse aplicar a numerosas competiciones y aprender sobre su funcionamiento.
- Verificación de hipótesis basándonos tanto en las relaciones mostradas por las variables como los resultados de los distintos modelos creados.
- Comprobación del modelo poniéndolo a prueba con datos de diferentes patinadoras. Pudiéndose comprobar así la viabilidad económica de su uso para las apuestas deportivas.

Para cumplir estos objetivos se llevan a cabo numerosos procesos, los cuales se explican en los siguientes capítulos, permitiéndonos crear el modelo óptimo. En primer lugar, es necesaria una revisión de la bibliografía existente, analizando los métodos empleados en otras predicciones deportivas y su adaptabilidad al patinaje artístico.

Una vez llevadas a cabo estas comparaciones se identifican la regresión lineal múltiple y las máquinas de vectores de soportes como las técnicas que pueden ofrecer mejores resultados. Las cuales se emplean para diseñar el modelo, con la idea de así poder aplicarlo de manera práctica.

Para llevar esto a cabo, es necesario identificar las variables relevantes y recopilar la información necesaria. Creando una base de datos que nos permita tanto entrenar el modelo como ponerlo a prueba.

Los resultados obtenidos nos permiten entender la relación entre las variables explicativas y la variable dependiente, así como la capacidad de adaptación del modelo. A partir de estos se pretenden comprobar las distintas hipótesis relacionadas con el modelo, así como comparar ambos modelos para poder seleccionar el óptimo.

También se pretende demostrar la utilidad de este modelo aplicándolo de forma práctica. Para ello se utilizarán los datos de diferentes patinadoras durante las olimpiadas de 2022. A través de este ejemplo se pretende mostrar al lector de manera sencilla el funcionamiento del modelo diseñado durante este trabajo.

Por último, cabe destacar que todo lo explicado anteriormente da lugar a un trabajo el cual presenta una clara alineación con los Objetivos de Desarrollo Sostenible43. Los ODS son una serie de objetivos propuestos en 2015 con la intención de alcanzar un futuro más sostenible para todos. La alineación de este trabajo con dichos objetivos es clara, relacionándose con la salud y bienestar ya que se pretende concienciar sobre la importancia de un deporte en general.

[43] ODS por sus siglas.

Capítulo 2: Estado del arte

El objetivo de este capítulo es llevar a cabo un estudio de algunos de los proyectos que han tratado de aplicar diferentes herramientas disponibles hoy en día para la creación de modelos predictivos. Para ello, se analizará tanto el proceso de creación de los distintos modelos como los resultados obtenidos a partir de estos.

En la actualidad existen un gran número de técnicas que pueden resultar útiles tanto a la hora de diseñar un modelo como al seleccionar las variables óptimas involucradas en este. Al existir tal variedad es necesario conocer correctamente el método que se está empleando y su adecuación al problema en el que se está trabajando.

Para alcanzar este conocimiento, se pretende analizar y comparar distintos estudios disponibles que han resultado de interés. A partir de este análisis se obtendrá la información necesaria para evaluar y discutir la posible aplicación de las herramientas existentes a las competiciones de patinaje artístico sobre hielo.

2.1 Modelo de regresión lineal múltiple

Estudiando las posibles técnicas a utilizar en este trabajo se ha encontrado un método que parece adaptarse al ejemplo con el que trabajamos. Este consiste en la creación de una regresión lineal múltiple para la predicción de precios de las obras de arte utilizando el método de los mínimos[44] cuadrados[45].

El trabajo trata el precio de obras de arte, un tema que puede llegar a ser muy subjetivo[46]. Este uso de la subjetividad para la predicción de un valor numérico nos ha llamado la atención ya que, como se ha mencionado en el capítulo anterior, este proyecto se centra en entender la puntuación de las patinadoras más allá de la complejidad técnica de sus programas.

Sin embargo, los resultados obtenidos durante la predicción de obras de arte no son los esperados. Se plantea que, pese a tratarse de una técnica que en otros ejemplos puede dar unos excelentes resultados, hay casos en los que puede llegar a ser una aproximación demasiado simple[47].

En resumen, por lo general la regresión lineal múltiple parece ser una técnica apropiada para el tipo de modelos predictivos con los que se pretende trabajar, pero a veces puede resultar demasiado simple para situaciones que se ven afectadas por varios factores[48].

[44] Método utilizado para predecir los parámetros de la recta de regresión mediante la minimización de los residuos.
[45] López-Silvarrey, G. (2021)
[46] Ibid.
[47] Ibid.
[48] Ibid.

Para poder confirmar su utilidad se estudiará en los próximos capítulos su uso en el caso de las competiciones de patinaje artístico.

2.2 Modelos predictivos en el mundo del deporte

Desde el comienzo de la humanidad el deporte se ha considerado una parte fundamental de la sociedad, beneficiándola cultural y pedagógicamente[49]. Tal es la importancia que hoy en día grandes cantidades de dinero y tiempo son invertidas en competiciones y patrocinios.[50]

Otra actividad relacionada con este entorno que mueve grandes sumas de dinero son las apuestas deportivas[51]. Estas se remontan a la antigua Grecia, donde la predicción de los resultados de las Olimpiadas levantó el interés de los espectadores. Sin embargo, no fue hasta el siglo XIX cuando las apuestas de carreras de caballos y partidos de béisbol profesional alcanzaron una increíble popularidad, coincidiendo con el surgimiento de las conocidas como casas de apuestas. Desde entonces el mundo de las apuestas ha evolucionado y perfeccionado su modo de operar, habiéndose convertido en los últimos años en una industria multimillonaria.[52]

En países como España, estos beneficios suponen un ingreso extra para entidades y organismos como el gobierno o las propias organizaciones deportivas, llegándose a financiar eventos como el Mundial de Fútbol de 1982 o los juegos Olímpicos de 1992. Las apuestas deportivas también resultan beneficiosas tanto para los operadores de juego como para los apostadores. Estos últimos llegando a invertir cantidades millonarias en ellas.[53]

Debido a todo esto y mucho más, se ha vuelto fundamental tanto para los deportistas como para los inversores intentar averiguar con la mayor precisión posible los resultados de las competiciones[54]. Lo que, sumado a la gran cantidad de datos que son recopilados hoy en día en el mundo del deporte, ha llevado al desarrollo de modelos predictivos con los que obtener un resultado fiable[55]. Estos pueden ser aplicados a varios aspectos, los cuales se discuten a continuación.

Un ejemplo de una de estas posibilidades es su uso en la predicción de lesiones deportivas, información que puede afectar la carrera de algunas personas de manera permanente. La capacidad de conocer cómo o cuándo pueden darse estas lesiones proporciona una ventaja fundamental tanto a deportistas como a entrenadores, permitiéndoles prevenirlas[56].

[49] Arias, F. G. (2017)
[50] Ibid.
[51] Valera, F. (2013)
[52] Important Notice | 888.com™, s. f. (2022)
[53] Carcedo, L. P. (2010)
[54] Bunker, R. P. et. al. (2019)
[55] Valera, F. (2013)
[56] Llamas, M. D. C. J. (2021)

Este tipo de estudios comenzaron con el desarrollo de un modelo de regresión logística sobre la lesión de jugadores de baloncesto.[57] Tras años de investigación y desarrollo se han alcanzado los modelos predictivos actuales, en los que las ecuaciones de regresión logística se consideran un método válido. Sin embargo, aún existe una clara necesidad de mejoría y desarrollo[58].

Otra aplicación destacable de estos modelos es la predicción de resultados en competiciones deportivas. Este trabajo se centrará en su aplicación, desarrollando el tema en profundidad a continuación.

2.3 Antecedentes

Durante la primera parte del capítulo se ha explicado la importancia de la predicción en distintos ámbitos deportivos, tanto para aquellos directamente involucrados en la competición como para los que se ven afectados por sus resultados. El resto del capítulo se centrará en el estudio de diversos modelos predictivos orientados al área de la predicción de resultados en competiciones deportivas.

En 2003 se lleva a cabo uno de los primeros proyectos dedicados a la predicción de las competiciones de patinaje sobre hielo[59]. El objetivo del autor es crear un modelo estadístico capaz de predecir la probabilidad de que un patinador finalice en una determinada posición al final del campeonato tras la obtenida en el programa corto[60].

Este trabajo se centra en recalcar la gran cantidad de elementos que involucran las competiciones de patinaje como un complejo sistema de evaluación o el hecho de que se trate de un evento que involucra gran cantidad de factores que pueden alterar el resultado. Este enfoque tan detallado puede resultar de gran utilidad a la hora de desarrollar el trabajo que se está llevando a cabo.

Lo datos necesarios para el diseño del modelo son obtenidos a través de páginas webs relacionadas con el mundo del patinaje artístico como Icecalc.com o la Asociación Estadounidense de Patinaje Artístico. El autor se limita al uso de datos obtenidos de torneos profesionales a nivel nacional e internacional de patinaje artístico con divisiones de hombres, mujeres y parejas. En total se obtiene información de 107 competiciones diferentes durante una duración de 5 años y con una media de 12 participantes.

A continuación, comienza el segundo paso en la creación del modelo, el cual consiste en la manipulación y filtración de datos. Al utilizarse información procedente de distintas fuentes a veces los datos pueden estar incompletos o erróneos, por lo que es imprescindible verificarlos y resolver los posibles errores existentes.

[57] Shambaugh et al. (1991)
[58] Llamas, M. D. C. J. (2021)
[59] Lemons, M. Q. (2003)
[60] Short program (SP) se trata de la primera prueba realizada en las competiciones de patinaje artístico sobre hielo.

Una vez llevada a cabo esta acción se obtiene un conjunto de datos fiable con el que el autor crea dos matrices distintas[61]. La primera tabla, de tamaño 7x7, es la encargada de mostrar las distintas veces que un patinador ocupa la misma posición en ambos programas. Por otro lado, la matriz de 6x6 al número de veces que un patinador en un lugar particular tras el SP terminó por delante de otro patinador con otra posición tras el programa corto.

Posteriormente, se hace uso de estos datos con la intención de desarrollar un modelo estocástico capaz de predecir de forma precisa los resultados en las competiciones. Para ello el autor menciona distintas características del patinaje artístico a tener en cuenta. Entre ellas destaca el hecho de que el programa largo tiene una duración de prácticamente el doble que el programa corto, duplicando las probabilidades de cometer un fallo.

Inicialmente se considera utilizar modelos que resultan insuficientes como el exponencial o el normal[62]. Tras una serie de ensayos se alcanza la conclusión de que la mejor opción es el uso de un modelo normal mixto[63], el cual tiene en cuenta el hecho de que el patinador actúe por debajo de su rendimiento habitual. Los parámetros involucrados en este se calculan mediante el conocido método de máxima verosimilitud, encargado de calcular los valores óptimos para este modelo.

A continuación, se hace uso de dos herramientas diferentes para evaluar el modelo. Por un lado, la distribución chi-cuadrado se encarga de evaluar la bondad de ajuste del modelo. Es decir, su capacidad para ajustarse adecuadamente a los datos observados. En el caso de la precisión del modelo se utiliza el método Monte Carlo. Se realizan 100 pruebas, cada una con 107.000 simulaciones, mediante el programa FORTRAN para así validar el modelo correctamente.

A partir de este trabajo de pueden obtener diversas conclusiones. La primera es la alta precisión que muestra el modelo a la hora de predecir los resultados deseados, demostrando ser de gran utilidad[64]. El modelo no solo es capaz de predecir competiciones futuras, sino que también es capaz de explicar los resultados obtenidos en el pasado. Otra conclusión que se puede obtener de este proyecto es la eficacia de herramientas como la distribución chi-cuadrado a la hora de evaluar el modelo[65].

Sin embargo, el cambio en el sistema de evaluación de la Unión Internacional de Patinaje sobre Hielo[66], el sistema de puntos oficial, en 2004 puede suponer un inconveniente a la hora de adaptar este modelo al trabajo que se está llevando a cabo. Este problema sumado al hecho de que el objetivo del ejemplo es predecir el puesto final basado en la

[61] Lemons, M. Q. (2003)
[62] Ibid.
[63] Ibid.
[64] Ibid.
[65] Ibid.
[66] International Skater Union (ISU)

posición tras la primera prueba, mientras que este trabajo no se basa en clasificaciones previas en el mismo campeonato a la hora de predecir, puede suponer una gran limitación a la hora de utilizar este modelo.

Por otro lado, en las últimas décadas el desarrollo en el ámbito de las predicciones deportivas se ha visto fomentado por el avance de distintas herramientas entre las que destaca el surgimiento de la inteligencia artificial. El aprendizaje automático se define como un campo de estudio que permite a las máquinas mejorar mediante el uso de información y datos históricos[67]. Esto revolucionó el mundo de los modelos predictivos, permitiendo a los expertos contar con la información necesaria para crear modelos más precisos y sofisticados.

A pesar de estos avances tan significativos, las predicciones deportivas no son tarea fácil ya que la manipulación y análisis de los datos recopilados debe ser exhaustiva y precisa. La dificultad aumenta en el mundo de la alta competición, donde en un mismo evento deportivo hay una enorme cantidad de posibles resultados y de situaciones que pueden alterar estos.

Esta dificultad aumenta al tratarse de un deporte minoritario y de unas herramientas tan novedosas ya que pocos recursos han sido invertidos en el mundo del patinaje. Esto ha supuesto un problema, debido a que se ha encontrado poca información disponible directamente relacionada con el proyecto. Sin embargo, durante los últimos años se ha tratado de desarrollar modelos precisos y útiles en otros ámbitos, incluyendo deportes más populares. Estos han resultado en interesantes métodos con una posible capacidad de adaptación al proyecto, los cuales se desarrollan a continuación.

Entre estos novedosos métodos relacionados con la inteligencia artificial destacan especialmente los conocidos como redes neuronales[68] y máquinas de vectores de soporte[69]. En un intento de comprender mejor ambas técnicas y de perfeccionar la predicción de resultados en el mundo de la natación se publica un estudio en el que a través de ambos métodos se pretende predecir qué nadadores alcanzarán el top 25% a lo largo del tiempo[70].

El proyecto comienza con la recopilación de datos a través de federación de natación estadounidense, obteniéndose toda la información disponible acerca de los tiempos de 7730 nadadores, tanto hombres como mujeres, desde los 10 años. La forma de representar los datos de cada nadador es un modelo vectorial con cinco elementos:

[67] Samuel, A. (1959)

[68] Neural Networks (NN) es una técnica empleada en problemas complejos o poco estructurados, la cual se basa en el uso de unidades de procesamiento inspirada en el sistema nervioso humano.

[69] Support Vector Machine (SVM) es un algoritmo empleado en los problemas de clasificación y regresión, el cual consiste en la búsqueda del hiperplano de separación óptimo entre dos clases maximizando el margen entre ellas.

[70] Xie, J. et. al. (2016)

estilo de nado, distancia de la piscina, edad, tiempo y su tabla de puntos. Posteriormente, los nadadores se dividen en cuatro grupos dependiendo de los tiempos almacenados, de menor a mayor, a la edad de 13 y a la de 18.

Una vez estos datos son representados y procesados, se dividen de la siguiente forma: 80% se destinan al conjunto de entrenamiento y los restantes al de prueba. El primer grupo es el responsable de diseñar correctamente el modelo, mientras que el segundo se encarga de evaluar el funcionamiento y la precisión de este. A continuación, se aplican los dos algoritmos a la misma base de datos, teniendo como entrada los tiempos previamente mencionados, para así crear un modelo capaz de predecir el nivel de un nadador a los 18 años.

Por un lado, se utilizan las previamente mencionadas redes neuronales. Estas no solo son empleadas en el ámbito deportivo, otros proyectos hacen uso de ellas como en el caso del trabajo centrado en la tasación de las obras de arte.[71] El autor hace uso de este algoritmo y de la técnica de regresión hedónica en un intento de diseñar un modelo predictivo óptimo para la tasación de estas obras. Analizando los resultados de ambas herramientas y sus conclusiones se puede comprobar la alta precisión que aportan ambos métodos en campos más subjetivos como puede ser el arte[72].

Se trata de una herramienta compleja, el requerir un entrenamiento y ajuste tan exhaustivo puede resultar en un proceso largo y costoso[73]. En el caso del modelo aplicado en el ejemplo que se está estudiando, el algoritmo es ajustado para alcanzar un total de 2000 iteraciones durante el entrenamiento y una tasa de aprendizaje de 0,001.

En este proyecto también se hace uso del algoritmo SVM a la hora de crear el modelo. En caso de tratarse de clases no linealmente separables[74], una parte fundamental de este método es la conocida función matemática *Kernel*[75], la cual facilita la separación entre las distintas clases. Esta puede ser de distintos tipos como la Kernel lineal o la gaussiana, siendo esta última la utilizada en el proyecto.

Tras aplicar ambos métodos se puede comprobar como ambos tienen resultados similares. Es decir, en general ambos algoritmos presentan un nivel de precisión elevado, pudiéndose hacer uso cualquiera de los dos, incluso para respaldar los resultados entre ellos[76]. Sin embargo, algo que destaca a la hora de analizar este estudio

[71] López-Silvarrey, G. (2021)
[72] Ibid.
[73] Mijwel, M. (2018)
[74] Heras, J. M. (2019, 28 mayo)
[75] Se trata de una función matemática utilizada para transformar los datos en un espacio de mayor dimensión. Estas se emplean a la hora de mapear los datos.
[76] Xie, J. et. al. (2016)

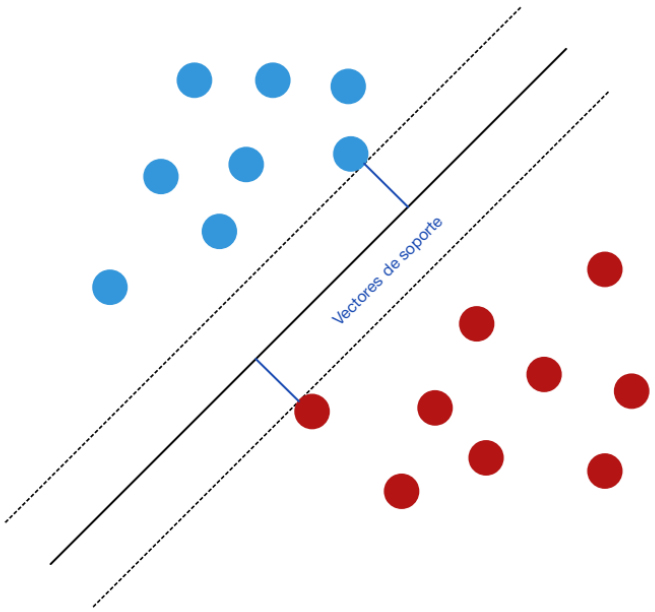
es que en algunas categorías la precisión de estos modelos disminuye, como en el caso de la categoría de braza femenina respecto a la masculina. Estas variaciones son algo a lo que se prestará especial atención a la hora de desarrollar el trabajo, eligiendo las herramientas que mejor se adapten a la categoría estudiada.

Por otro lado, en un intento de mejorar el previamente mencionado algoritmo SVM se plantea la creación un modelo híbrido[77], algo que puede resultar interesante a la hora de trabajar con aprendizaje automático. Este enfoque lleva al desarrollo del estudio que se plantea a continuación[78].

Se plantea esta actualización de las Máquinas de Vectores de Soporte, utilizadas principalmente para problemas de clasificación y regresión, en un intento de perfeccionar la predicción de partidos de baloncesto. Esta mejora consiste en la combinación del mencionado algoritmo con la técnica conocida como árboles de decisión[79], concretamente con el algoritmo C4.5.

El primer elemento de esta combinación es el SVM, el cual pretende hallar el hiperplano de separación óptimo entre dos clases maximizando el margen entre ellas, ver figura 8.

Figura 8: Clasificación mediante SVM



Fuente: Elaboración propia (2023)

Se pretende hallar el margen máximo de separación entre el hiperplano y las clases. Para ello se hace uso de los vectores de soporte, los puntos que definen esta separación

[77] Hyperbolic Support Vector Machines and Decision Tree (HSVMDT)
[78] Pai, P. F., et. al. (2017)
[79] Técnica estadística utilizada para generar reglas de decisión.

máxima. Esto se consigue resolviendo el problema de programación cuadrática que se muestra en la siguiente ecuacuión:

$$Minimizar\ \frac{1}{2}w^2 + C\left(\sum_{j=1}^{n} \Psi_j\right) \qquad [1]$$

Esta ecuación se puede dividir en dos partes: una primera encargada de definir y orientar el hiperplano responsable de la separación de categorías; y una segunda encargada de la regulación de la cantidad de errores permitidos durante la clasificación. Esta se resuelve utilizando la técnica de multiplicadores de Lagrange.

En cuanto a los parámetros de entrada al tratarse del método SVM, vuelve a ser necesario el uso de la función *Kernel*. En este caso, al igual que en el explicado previamente, se utiliza una funcion de *Kernel* gaussiana.

El segundo elemento es el conocido como árboles de decisión, los cuales tienen como objetivo la creación de una estructura que permita conocer las distintas decisiones existentes y sus posibles consecuencias. En este caso su creación se lleva a cabo a través del algoritmo C4.5, el cual es capaz de generar la estructura a partir de un grupo de datos, conocidos como datos de entrenamiento.

El árbol de decisión se forma mediante particiones, a partir de las cuales surge un nuevo nodo. El algoritmo evalúa todas las posibles particiones para así elegir aquella que proporcione mayor cantidad de información. Es decir, a partir de la conocida técnica de proporción de ganancia se selecciona el atributo de división.

Además, se utiliza el método de selección de características basado en la correlación[80] un método de clasificación y evaluación. A partir de medidas como la correlación entre variables clasifica las funciones en subconjuntos, seleccionándose el que mayor relevancia y menos redundancia presente.

Una vez entendidas las distintas herramientas a utilizar es más sencillo de entender el proceso de creación del modelo. El trabajo comienza con la recopilación de información necesaria, la cual se obtiene de fuentes oficiales como “NBA.com” o “ESPN.go.com” que permiten la obtención de información de 400 partidos diferentes, aumentando así la precisión del modelo.

A continuación, comienza una parte fundamental del proyecto. La selección de variables relevantes que han de ser incluidas es un paso fundamental ya que se eliminan las características menos importantes o aquellas que resultan redundantes para así evitar cualquier interferencia en el modelo, consiguiendo así mayor precisión. Este se lleva a cabo mediante la herramienta CFS, la cual identifica un subconjunto óptimo de 7 variables. Durante la preparación de este capítulo es un paso que destaca notoriamente, pues el autor demuestra mediante una serie de tablas comparativas que el uso de

[80] Correlation-based Filter Selection (CFS)

técnicas como CFS puede incrementar la precisión[81] del modelo en un 18,25%. Esta información se tendrá en cuenta a la hora de realizar el trabajo y de decidir qué herramientas utilizar.

El tercer paso en la creación del modelo es la clasificación. Este paso sirve para llevar a cabo la división de los datos previamente ajustados en el conjunto de dato de entrenamiento, utilizado para diseñar el modelo, y el de prueba, utilizado para evaluar este. Esta información es utilizada posteriormente como datos de entrada en el algoritmo SVM para así crear el modelo y evaluarlo.

Por último, tras la creación y evaluación es necesario crear las reglas de decisión. Utilizando los resultados del modelo SVM como datos de entrada para el árbol de decisión es posible generar las normas necesarias. Una vez se finaliza este trámite se puede dar por finalizado el trabajo y comenzar su análisis.

Este modelo resulta ser aceptable a la hora de predecir partidos de baloncestos y ayudar a los entrenadores a la hora de diseñar estrategias, utilizando una herramienta muy común entre los modelos deportivos como es el algoritmo SVM y mostrando predicciones precisas[82]. Otro motivo por el que destaca es que aporta un interesante enfoque a la hora de diseñar modelos predictivos, mejorando una técnica ya existente mediante la combinación de algoritmos.

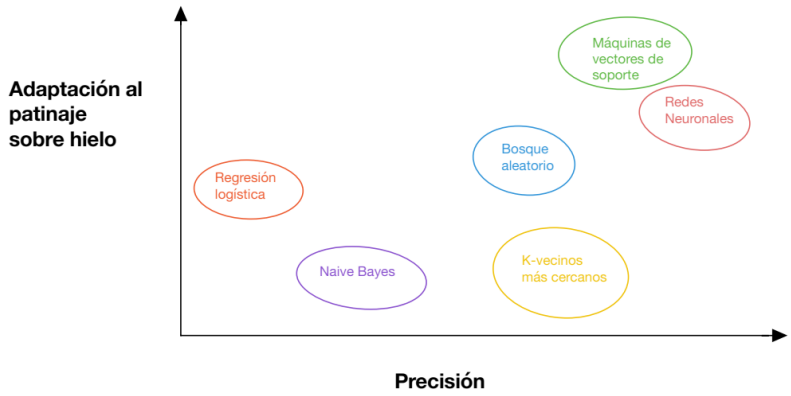
Sin embargo, este tipo de modelo solamente es capaz de diferenciar entre "partido ganado" y "partido perdido", siendo esto una gran limitación a la hora de aplicarlo al patinaje artístico. Esto significa que aún es necesario mejorar el modelo antes de aplicarse al proyecto en el que se está trabajando.

Finalmente, se decidió enfocar las herramientas existentes gracias a la inteligencia artificial a deportes relacionados con el patinaje. Se propuso la creación de modelos para la predicción de resultados en las competiciones de patinaje de velocidad femenino mediante el uso de distintos métodos de aprendizaje automático.[83]

El patinaje de velocidad se trata de un deporte con un sistema de puntuación notoriamente complejo; la competición consiste en cuatro pruebas distintas y solo las mejores patinadoras resultantes de las tres primeras pueden clasificarse para la última. A lo que se añade la compleja evaluación que se lleva a cabo para la organización del ranking. Esta complejidad ha llevado a expertos al estudio a fondo de nuevos modelos predictivos mediante seis algoritmos distintos: máquinas de vectores de soporte, regresión logística (RL), bosque aleatorio (RF), K-vecinos más cercanos (KNN), Naive Bayes (NB) y redes neuronales, ver figura 9.

[81] Pai, P. F., et. al. (2017)
[82] Ibid.
[83] Liu, M., et. al. (2022)

Figura 9: Métodos analizados



Fuente: Elaboración propia (2023)

La base de datos necesaria para la elaboración de estos modelos se obtiene, una vez más, a partir de fuentes oficiales como la ISU. Esta está formada por un total de 71 variables procedentes de los resultados a lo largo de 16 temporadas de 64 mujeres distintas.

Una vez más, los autores toman la decisión de filtrar las características existentes en un intento de disminuir la complejidad del problema y aumentar el rendimiento del modelo. En este caso, se utiliza la regresión de Lasso, encargada de seleccionar las variables importantes al mismo tiempo que reduce el efecto de las redundantes e irrelevantes, representada de la siguiente forma:

$$(\alpha, \beta) = \text{argmin} \sum i = \mathbf{1} n(yi - \alpha i - Xi\beta)\mathbf{2} + \lambda \parallel \beta \parallel_1 \quad [2]$$

Donde Xi es el i-ésimo grupo de variables independientes; α y β son coeficientes de regresión funcionalmente necesarios; Yi es el valor de la variable dependiente Xi; n es el tamaño del conjunto de datos utilizado para el modelo de regresión; y λ y t son parámetros en varias formas de regresión de Lasso. Se hace uso de todos estos elementos en un intento de minimizar la suma de los errores de predicción eligiendo las características óptimas.

A continuación, los datos se dividen en cinco grupos diferentes: cuatro encargados de diseñar el modelo y el último encargado de la verificación de este. Posteriormente se aplicarán los distintos algoritmos a estos subgrupos con el objetivo de crear un modelo capaz de predecir si una patinadora se clasificará para la competición de 5000m y otro

centrado en el medallero, para el cual no es posible aplicar el método NN debido a una incompatibilidad en el tamaño de los datos.

Una vez aplicados los diferentes algoritmos a las variables restantes se puede determinar la calidad de estos a través de los conocidos como indicadores de evaluación. Estos son herramientas utilizadas para medir la precisión de los modelos e identificar posibles mejoras. Existen distintos tipos, pero en este caso se hace uso de algunos de los más reconocidos por los profesionales:

$$Exactitud = \frac{TP + TN}{TP + TN + FP + FN} \quad [3]$$

$$Precisión = \frac{TP}{TP + FP} \quad [4]$$

$$Sensibilidad = \frac{TP}{TP + FN} \quad [5]$$

$$F1 = \frac{2TP}{2TP + FP + FN} \quad [6]$$

Donde TP representa Verdaderos Positivos[84] (True Positives), TN representa Verdaderos Negativos[85] (True Negatives), FP representa Falsos Positivos[86] (False Positives) y FN representa Falsos Negativos[87] (False Negatives).

Además de estos factores, también se ha hecho uso AUC[88], el cual demuestra mayor efectividad por parte del modelo a medida que su valor se acerca a la unidad.

En cuanto al modelo encargado de predecir los resultados de la carrera de 5000m, las medidas expresadas anteriormente muestran que SVM, RF, LR y NN son los algoritmos más efectivos. Todos estos modelos parecen ser útiles y confiables[89]. Sin embargo, entre ellos destacan las máquinas de vectores de soporte con una exactitud, precisión y F1 mayores, demostrando ser la mejor opción para este problema.

Teniendo en cuenta que en el caso del modelo predictivo de medallas no es posible usar el método NN, las medidas de evaluación se aplicarán al resto. Una vez más destaca el

[84] Se conocen como TP los casos que son correctamente identificados por el modelo como positivos.
[85] Se conocen como TN los casos que son correctamente identificados por el modelo como negativos.
[86] Se conocen como FP los casos que son incorrectamente identificados por el modelo como positivos.
[87] Se conocen como FN los casos que son incorrectamente identificados por el modelo como negativos.
[88] El área bajo la curva ROC
[89] Liu, M., et. al. (2022)

algoritmo SVM, mostrando el valor de AUC más alto además de una mayor exactitud, sensibilidad y precisión. Es decir, este modelo demuestra ser el más efectivo y estable de los 5 disponibles.

Estos resultados no suponen una gran sorpresa, ya que el uso del algoritmo SVM es muy común en el ámbito deportivo[90]. Al tratarse de un método capaz de trabajar eficazmente a partir de un conjunto de datos limitado, parece ser ideal para este tipo de proyectos a pequeña escala[91]. Sin embargo, al tratarse de un deporte centrado en la velocidad no se puede confirmar su eficacia en un deporte con un factor subjetivo tan predominante como el patinaje artístico sobre hielo sin antes comprobarlo.

2.5 Conclusiones

Tras el análisis de diversos estudios se comprueba que a día de hoy no es posible asegurar que los métodos existentes sean totalmente eficaces una vez se aplican al mundo del patinaje artístico. Por ello se ha considerado necesario comprobar la eficacia de modelos ya existentes al aplicaros a este caso en particular.

El caso de la regresión lineal múltiple es el primer ejemplo de los diferentes métodos que más han llamado la atención. Se trata de un trabajo con resultados muy interesantes y se cree que tiene gran capacidad de adaptación a la situación actual. En los próximos capítulos se discutirá la posibilidad de aprovechar este método a la hora de predecir competiciones de patinaje artístico sobre hielo.

Por otro lado, ante los resultados obtenidos en otros proyectos se ha decidido investigar la posibilidad de utilizar otra herramienta utilizada con anterioridad conocida como máquinas de vectores de soporte. Este algoritmo ha demostrado ser de gran utilidad en el mundo de las predicciones deportivas debido a su gran capacidad de adaptación y a su efectividad a la hora de trabajar con un número relativamente pequeño de datos[92]. Esto puede resultar beneficioso para el trabajo ya que se trata de un estudio a pequeña escala.

Los beneficios previamente mencionados sumados a su capacidad de generalización, es decir su capacidad para evitar el sobreajuste, convierten a las máquinas de vectores de soporte en una herramienta innovadora y eficaz. En los siguientes capítulos se comprobará si estos famosos beneficios se obtienen a la hora de aplicar el modelo a las competiciones de patinaje artístico.

Otra característica observada a través de estos estudios previos es la importancia de seleccionar los subconjuntos de variables óptimos. Dos métodos muy comunes a la hora de realizar esta tarea son la regresión de Lasso y el método de selección de características basado en la correlación, ambos utilizados en los ejemplos dados. En este

[90] Ibid.
[91] Suárez, E. J. C. (2014)
[92] Ibid.

caso se utilizará el segundo método, al tratarse de una técnica manejable y sencilla de entender a través de Rstudio.

Los próximos capítulos del trabajo se centrarán en la aplicación de estas herramientas con el objetivo de contrastar ambas y seleccionar la que mejor resultados presente. Es decir, se pretende comprobar que técnica ofrece un modelo óptimo para la predicción de competiciones de patinaje artístico sobre hielo.

CAPÍTULO 3: El papel de las variables en los modelos predictivos

A la hora de crear un modelo es imprescindible seleccionar correctamente las variables a utilizar en su diseño. Este paso resulta fundamental pues la correcta selección de variables lleva a un modelo suficientemente explicativo evitando el sobreajuste[93].

Este capítulo se centra en la selección de este conjunto de variables óptimas, así como en la obtención de una base de datos fiable y completa de la cual se pueda extraer información sobre las variables seleccionadas previamente. También se analiza la formulación de hipótesis sobre las relaciones entre las variables que guían el proceso de diseño.

3.1 Definición de variables

El primer paso a la hora de crear un modelo es entender el posible efecto de determinadas variables sobre la variable dependiente, ya sea de forma directa o indirecta. En un ejemplo tan complejo como es una competición deportiva hay que prestar especial atención a este proceso ya que existe una gran cantidad de factores que pueden alterar sus resultados[94].

El objetivo de este trabajo es la creación de un modelo de regresión que sea capaz de predecir de manera eficaz y precisa la puntuación de una patinadora tras la realización de un determinado programa. Teniendo esto en cuenta, se han seleccionado las siguientes variables, las cuales se consideran que pueden llegar a influir en este resultado:

- ***Edad de la patinadora en el momento de la competición***: El patinaje sobre hielo es un deporte que castiga duramente a aquellos que lo practican[95]. Tanto la dificultad de los elementos involucrados como el nivel de perfección demandado suponen un gran desgaste físico y mental para los patinadores[96]. Por ello mantener durante muchos años el ritmo necesario para destacar en un deporte tan exigente es complicado. Esta situación puede beneficiar a las patinadoras jóvenes, cuyos cuerpos son capaces de aguantar esta exigencia[97]. Esto se puede comprobar fácilmente con las patinadoras rusas quienes son capaces de realizar elementos de alta dificultad a una temprana edad, pero que prácticamente cada temporada son reemplazadas[98].

  Por otro lado, durante los últimos años el mundo del patinaje artístico ha comprobado que la inexperiencia puede jugar malas pasadas a sus patinadoras[99].

[93] Carrasco, M. (2016).
[94] Lemons, M. Q. (2003)
[95] Castilla, J. (2018).
[96] Ibid.
[97] Martín, A.(2022).
[98] Ibid.
[99] Efe. (2022)

Por ello, se considera que las patinadoras más adultas pueden jugar con la ventaja de una mejor perspectiva tanto dentro como fuera del hielo, así como un mejor manejo del éxito y el fracaso[100].

Esta fuerza mental puede ser algo decisivo, ya que la presión psicológica ligada a este deporte puede ser fatal a la hora de competir, situación que se pudo comprobar con la joven promesa del patinaje ruso Kamila Valieva[101]. Con tan solo 15 años y un primer puesto en la primera jornada de los Juegos Olímpicos de Beijin se vio involucrada en un escándalo de dopaje ese mismo día, patinar tras esta situación y con el peso de su país en sus hombros la llevó a romperse bajo la presión y quedarse fuera del medallero[102].

Ante estas dos posturas tan diferentes se plantea que la edad de una patinadora puede influir notablemente en la calidad de su programa. Por tanto, esta variable, medida en años, se propone como variable explicativa.

- ***Mejor marca personal de la patinadora***: Esta recoge la puntuación más alta conseguida por una patinadora hasta la fecha. Se espera que la máxima capacidad que puede alcanzar un deportista influya a la hora de predecir su puntuación. Por ello, se propone la mejor marca personal, medida en puntos, como variable de entrada en el modelo.

- ***Tiempo desde la mejor marca personal:*** Al tratarse del momento en el que un deportista ha alcanzado su máximo rendimiento, es importante tener en cuenta si esto se logró en un momento cercano a la competición analizada o si este puede no encontrarse en el momento álgido de su carrera. Por ello se plantea el tiempo transcurrido desde que una patinadora alcanza su mejor puntuación como variable independiente, medida en años.

- ***Edad de la patinadora en el momento de la mejor marca personal:*** Como se ha explicado previamente, el factor de la edad puede ser decisivo para la vida de un deportista. Por esto, conocer la edad de una patinadora en el momento el que ha logrado su máxima puntuación para conocer en que etapa de su carrera se encuentra un deportista. Es decir, puede llegar a determinar si esa marca aún puede ser mejorada o si se considera ese el momento álgido de su carrera. Por ello, se opta por incluir esta como variable explicativa, medida en años.

La perfección exigida a las patinadoras en el hielo se traduce en unos requisitos que sus cuerpos deben cumplir desde una temprana edad y a lo largo de su carrera deportiva[103].

[100] Ibid.
[101] Martín, A.(2022).
[102] Ibid.
[103] Martín, A.(2022).

A continuación, se plantean un par de variables relacionadas con el físico de la patinadora:

- ***Masa de la patinadora:*** A la hora de realizar algunos de los elementos más notorios en el patinaje sobre hielo es necesario un cuerpo muy delgado y fuerte[104]. Teniendo esto en cuenta se ha optado por el uso de esta variable, medida en kilogramos, como un factor clave a la hora de determinar la puntuación final de la patinadora.

- ***Altura de la patinadora***: Cuanto más bajo sea el centro de gravedad más fácil resulta a las deportistas realizar los elementos de manera ágil y elegante[105]. Por ello, se ha seleccionado la altura, la cual se mide en centímetros, como una variable explicativa a la hora de determinar la puntuación.

A lo largo de la historia del patinaje artístico han tenido lugar diferentes escándalos relacionados con el equilibrio entre la complejidad de los elementos y la elegancia femenina de las patinadoras, como es el ejemplo Surya Bonaly[106]. Mientras que la parte técnica está cuidadosamente regulada por la Unión Internacional de patinadores[107], existe una parte más enfocada a la armonía del programa[108]. La subjetividad de esta segunda parte la hace más propensa a la ambigüedad, pudiéndose llegar a utilizar para beneficiar o perjudicar a una patinadora[109]. Por ello se han considerado las siguientes variables relacionadas con esta segunda parte:

- ***Habilidades de la patinadora*** sobre el hielo, incluyendo elementos como su fluidez sobre el hielo, control de los filos[110] o control de velocidad[111].

- ***La actuación*** está relacionada con la capacidad de la patinadora de presentar los elementos técnicos al público[112]. Es decir, su forma de comunicarse con el público y transmitir sus emociones[113].

- ***Las transiciones*** se refieren a la capacidad de la patinadora de enlazar los elementos de un programa con distintos pasos y movimientos[114].

[104] Ibid.
[105] Sauces, I. A. (2020).
[106] Trula, E. M. (2017).
[107] ISU por sus siglas en inglés International Skaters Union
[108] Carreño, F. (2018).
[109] El financiero (2022).
[110] Eje de la cuchilla sobre el que se desliza la patinadora.
[111] U.S. Figure skating
[112] Ibid.
[113] Ibid.
[114] Ibid.

- ***La composición*** está relacionada con la organización y estructura del programa, teniendo en cuenta el espacio utilizado por la patinadora, así como la sintonía entre el programa y la música[115].

- ***Interpretación de la música:*** Esta variable se refiere a la capacidad de la patinadora de traducir la música en un programa[116].

Estas cinco variables encargadas de representar la parte más subjetiva de un programa se consideran decisivas a la hora de predecir la puntuación final de una patinadora. Todas ellas se puntúan sobre 10 y son decididas por un panel de jueces especializado[117].

Otro factor que se considera influyente en la puntuación de una patinadora es su nacionalidad, ya que se piensa que en determinadas ocasiones puede beneficiar a sus resultados. Esto lleva a plantearse las siguientes variables:

- ***Patinadora compite en casa si o no:*** Se emplea una variable binaria que toma el valor 1 si la competición tiene lugar en el país de la patinadora y 0 si es en otro sitio.

- ***Patinadora tiene misma nacionalidad que alguno de los jueces involucrados en la competición:*** Se emplea una variable binaria que toma el valor 1 si la patinadora y el juez tienen misma nacionalidad y 0 si son de distintos sitios.

Nuestra experiencia en distintas competiciones a lo largo de los de los años nos ha llevado a considerar las siguientes variables relevantes a la hora de plantear el modelo:

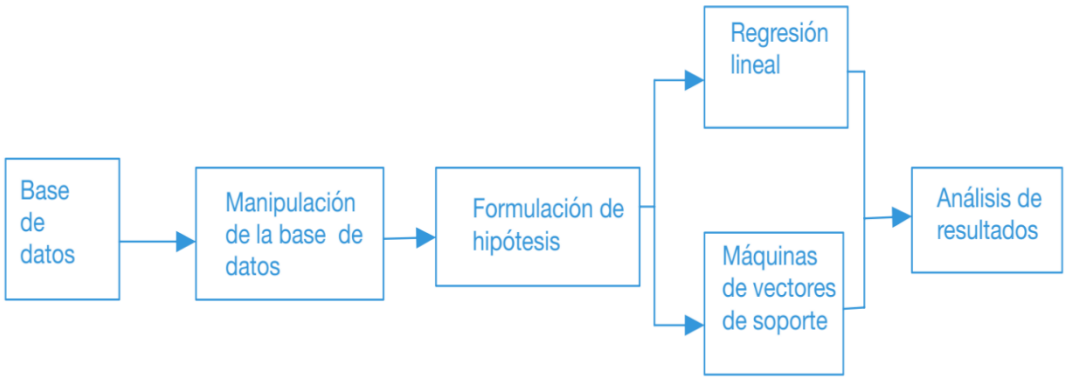
- ***Horas de entrenamiento:*** Cuando se practica este deporte a nivel profesional es necesario disponer de unas instalaciones y unos materiales específicos, como una pista de hielo, una zamboni[118] y unos patines en condiciones. El problema viene cuando algunos de estos recursos no están disponibles, obligando a las patinadoras a reducir su número de horas de entrenamiento.
  Tras años compitiendo a nivel internacional en este deporte, hemos podido comprobar como el problema mencionado previamente marca una diferencia en los resultados. Impidiendo a países como España, en los que el patinaje sigue siendo un deporte minoritario, alcanzar los mismos logros que aquellos en los que los recursos son notablemente mayores.
  Esto lleva a plantera la variable relacionada con las horas de entrenamiento de hielo a la semana de una patinadora con su puntuación final.

[115] Ibid.
[116] Ibid.
[117] Ibid.
[118] Máquinas encargadas de borrar las marcas creadas por los patinadores en el hielo.

- ***Orden en la competición:*** En una competición se ven involucradas un gran número de patinadoras, entre 30 y 40 en el programa corto y unas 24 en el programa largo[119]. Durante estos años compitiendo hemos podido comprobar como esto provoca que determinados puestos a la hora de salir a competir pueden tanto beneficiar como perjudicar a una patinadora, ya que entre unos y otros varía el nivel de exigencia de los jueces.
  Esto nos lleva a calificar la variable relacionada con el orden como influyente a la hora de definir la puntuación final de la patinadora. Esta se mide sobre el número total de patinadoras en la categoría.

Una vez definidas las variables que se consideran importantes a la hora de diseñar el modelo de regresión es necesario encontrar una base de datos que obtenga la información de dichas variables, la manipulación de estas y la posterior formulación de hipótesis basadas en esta información.

Figura 10: Esquema de la creación del modelo



Fuente: elaboración propia (2023)

3.2 Base de datos

A la hora de crear una base de datos que contenga la información necesaria para emplearla en el modelo es imprescindible el uso de una fuente fiable. En este caso se ha

[119] Solo mejores las 24 participantes en el programa corto se clasifican para la segunda jornada de la competición.

optado por la página web oficial de la ISU, la cual regula todo tipo de eventos internacionales relacionados con el patinaje sobre hielo de velocidad y artístico.

La organización guarda información no solo sobre resultados desglosados de diferentes categorías y competiciones desde el año 2013, sino también detalles adicionales de dichas competiciones como pueden ser los nombres de los nueve jueces involucrados y sus nacionalidades. Entre toda esta información se han utilizado los datos disponibles sobre los mundiales senior femenino entre los años 2022 y 2018, ya que se consideran competiciones lo suficientemente recientes y relevantes en el mundo del patinaje como para aportar información representativa[120].

Junto a esta información sobre los resultados históricos, la ISU proporciona una ficha de cada una de las participantes de dicha competición. En estos documentos es posible encontrar distintos datos sobre las patinadoras como su año de nacimiento, su altura o su mejor marca personal.

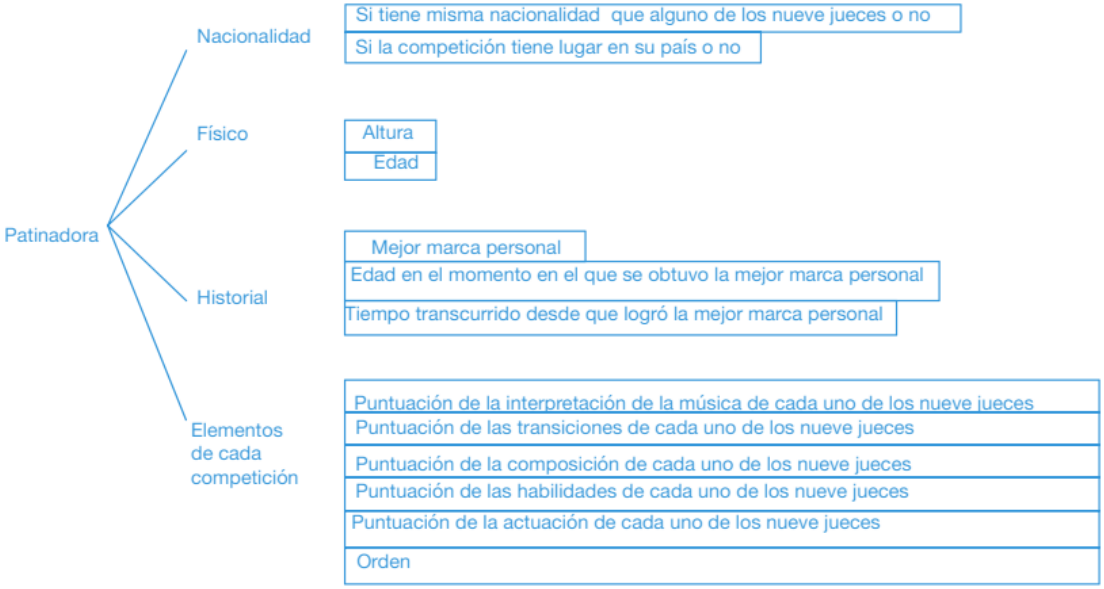
Una característica que en un principio se consideró relevante a la hora de desarrollar el modelo, pero no es proporcionada por la organización es la masa de las patinadoras. Se cree que esta falta de información se debe a los escándalos relacionados con los trastornos de conducta alimentaria que han tenido lugar en este mundo durante años, casos como el de Yulia Lipnitskaya quien con tan solo 19 años y siendo la patinadora artística más joven en convertirse campeona olímpica anunció su retirada en 2017 debido a esta enfermedad[121]. Debido a la controversia relacionada con esta variable no es posible contar con ella en el diseño del modelo.

Otro problema que se encuentra a la hora de recopilar los datos es el hecho de que la ficha de todas las patinadoras no está completa. Un gran número de las participantes no facilita el número de horas que entrena a la semana, impidiendo el uso de esta variable que en un principio se consideró influyente.

Una vez recopilada todos los datos necesarios se obtiene la siguiente información para cada patinadora:

[120] La competición en el año 2020 se canceló debido al COVID-19.
[121] Olivas, M. (2017).

Figura 11: Esquema de las variables que se tienen por patinadora



Fuente: Elaboración propia (2023)

Como se ha mencionado previamente cada una de estas variables está medida en una escala diferente: años, puntos o centímetros, lo que puede provocar que algunas variables tengan más influencia que otras[122]. Una forma de resolver este problema es normalizar la base de datos para que las variables tengan una magnitud comparable[123].

Este proceso de normalización se ha llevado a cabo utilizando la herramienta Rstudio[124]. Utilizando la función "scale()" se escala la base de datos de tal forma que su media es 0 y su desviación típica 1 a partir de la siguiente ecuación:

$$\frac{(x_i - \bar{x})}{s} \qquad [7]$$

Donde:

- x_i **es el valor original de la variable.**
- $\bar{x}$ **es la media de la muestra.**
- **s es la desviación** típica **de la muestra.**

Una vez llevada a cabo esta operación se dividen los datos de manera aleatoria en dos clases: datos de entrenamiento (70 % de la base de datos) y datos de prueba (30% restantes). El primer subconjunto es aquel cuyos datos que se utilizan para entrenar desarrollar y mejorar el modelo, mientras que los datos de prueba son los encargados

[122] García, J. A. (2020).
[123] Ibid.
[124] Un entorno de desarrollo integrado empleado en el lenguaje de programación R.

de comprobar el funcionamiento diseñado por los anteriores[125]. Este último es una manera de comprobar que no se ha producido lo que se conoce como sobreajuste, lo cual ocurre cuando el modelo se ajusta tan bien a los datos de entrenamiento que no es capaz de generalizar a otros conjuntos de datos[126]. Para ello se utiliza una vez más la herramienta Rstudio.

Una vez estructurada la base de datos a utilizar para la creación de este modelo se puede comenzar a formular hipótesis sobre su comportamiento.

3.3 Formulación de hipótesis

Una vez estudiada la información disponible y el funcionamiento de las competiciones de patinaje se pueden formular las siguientes hipótesis:

- Generalmente las patinadoras que no han alcanzado la pubertad se ven beneficiadas por un cuerpo delgado y sin desarrollar[127]. Se espera que las patinadoras de menor edad alcancen una puntuación más alta.

- Como se ha mencionado anteriormente, un cuerpo menudo beneficia a las patinadoras a la hora de realizar los distintos elementos involucrados en un programa[128]. Se espera que las patinadoras con menor altura obtengan mejores resultados.

- La mejor marca personal indica la máxima puntuación que es capaz de obtener una patinadora cuando las condiciones son óptimas, una mejor marca personal se asocia con un mayor nivel por parte de la patinadora. Se espera que aquellas patinadoras con mejores resultados históricos logren mejores puestos.

- El tiempo transcurrido desde que se logra la mejor marca personal puede indicar si esa patinadora se encuentra en su mejor época, si esta ya ha pasado (altos valores mayores que 0) o si está por llegar (altos valores menores que 0). Se supone que cuanto menor sean estos datos (en valor absoluto) más cerca se encuentra la patinadora de la cima de su carrera y mejor son sus resultados en las competiciones

- Se prevé que las patinadoras que obtienen mejor puntuación final son aquellas que trabajan en perfeccionar la parte artística de su programa. Es decir, aquellas con mejores resultados en sus habilidades, actuación, transiciones, composición e interpretación.

[125] De Los Santos, P. R. (2022).
[126] Ibid.
[127] Martín, A. (2022).
[128] Ibid.

- Se considera que el hecho de que una patinadora y un juez tengan la misma nacionalidad hace que consciente o inconscientemente este beneficie a la participante. Por ello se espera que aquellas patinadoras que compartan nacionalidad con al menos uno de los jueces obtienen mejores resultados.

- La nacionalidad de la patinadora también puede beneficiarla a la hora de competir si el evento se celebra en su país. La familiaridad del ambiente o un mayor número de aficionados capaces de trasladarse a la pista de hielo para apoyarla puede marcar una diferencia en los resultados. Por ello, se cree que aquellas patinadoras que estén compitiendo en su país logran mayores puntuaciones.

- Se espera que aquellas patinadoras que ocupen los últimos puestos en el orden de la competición se vean beneficiadas ante aquellas que ocupen los primeros.

Todas estas afirmaciones sirven de orientación para una investigación adicional. Los próximos capítulos se centran en plantear el modelo de forma teórica para su posterior aplicación práctica a la base de datos, pudiendo así corroborar o desmentir las hipótesis formuladas.

Cabe destacar que el objetivo de este trabajo es la creación de modelos regresivos mediante dos métodos distintos, para así poder compararlos y elegir la mejor opción. Por ello, cuando se menciona el planteamiento del modelo y su aplicación práctica que tienen lugar en los próximos capítulos se refiere tanto a la creación mediante una regresión lineal múltiple y las máquinas de vectores de soporte.

En definitiva, el modelo de regresión creado en este trabajo se centra en la predicción de puntuaciones en competiciones de patinaje sobre hielo y la comprobación de las hipótesis definidas previamente.

Por otro lado, la hora de diseñar este modelo hay una serie de pasos previos a su aplicación teórica que resultan imprescindibles. Estos pasos son la definición de variables y la creación de la base de datos, los cuales son de suma importancia ya que la calidad del modelo es directamente proporcional a la de estos[129].

Una vez tienen lugar estos pasos de definición de variables y recopilación de datos es necesario tanto el estudio de las técnicas a utilizar como su aplicación. Esto se lleva a cabo en los próximos capítulos, encargados de la explicación de la regresión lineal múltiple y las máquinas de vectores de soporte, así como su adaptación a la base de datos obtenida en este capítulo.

[129] De Los Santos, P. R. (2022).

Capítulo 4: Regresión lineal múltiple

Este capítulo se centra en la regresión lineal múltiple, el primer método que se pretende comparar en este trabajo. El objetivo principal de este capítulo es proporcionar una explicación detallada del modelo de regresión lineal y su aplicación teórica en el contexto del patinaje artístico.

A lo largo de este capítulo, se proporciona una descripción clara y concisa del modelo de regresión lineal múltiple, adaptándolo a las consideraciones especiales relacionadas con su aplicación en el patinaje artístico. Esto sentará las bases para la posterior comparación con el segundo método de análisis.

4.1 Fundamento teórico

Como ya se ha explicado en el capítulo previo, en las competiciones de patinaje artístico se ven involucrados distintos elementos que pueden afectar a la puntuación de cada patinadora de diversas formas[130]. Estos elementos pueden incluso llegar a escaparse del control de la patinadora, como pueden ser características físicas o intereses entre países, dificultando aún más el estudio de estos resultados[131].

Por lo tanto, a la hora de estudiar las puntuaciones de las patinadoras en las competiciones es importante tener en cuenta esta dependencia[132]. Esto lleva a la consideración del uso de la regresión lineal múltiple, un modelo estadístico encargado de explicar el comportamiento de esta variable a partir de las variables independientes, asumiendo una relación estadística lineal[133]. Es decir, para todo el fenómeno observado, cada variable participa de forma aditiva y constante[134]. Obteniéndose un modelo con la siguiente expresión general:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + U \quad [8]$$

Donde:

- Yi es la variable explicada
- β_0 es el valor de la variable cuando el resto tienen valor 0, un término constante del modelo
- $\beta 1 \ldots \beta n$ son el efecto que las variables explicativas tienen sobre la dependiente.

[130] Lemons, M. Q. (2003).
[131] Ibid.
[132] Ibid.
[133] Rodríguez-Jaume, J., et. at. (2001).
[134] Baños, V., et. al. (2019).

- U es el término relativo al error, o residuo

U se trata de la variable aleatoria que representa la influencia de otras variables explicativas que influyen sobre la variable dependiente y no se han tienen en cuenta[135].

Para definir el modelo es necesario estimar los parámetros presentes en la ecuación 8. Esto logra a partir de los datos disponibles[136]. Es decir, utilizando las observaciones disponibles se obtiene un sistema de n ecuaciones con n+k+1 incógnitas.

$$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \cdots + \beta_k X_{k1} + U_1$$
$$Y_2 = \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + \cdots + \beta_k X_{k2} + U_2 \quad [9]$$
$$Y_n = \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + \cdots + \beta_k X_{kn} + U_n$$

A la hora de calcular los elementos explicados previamente, es importante tener en cuenta una serie de supuestos simplificadores sobre la perturbación, las variables explicativas y las variables explicadas, así como los parámetros del modelo[137].

1. Los errores son variables aleatorias de media nula[138].
2. Todos los errores tienen la misma varianza[139].
3. Todos los errores están correlacionados entre sí[140].
4. El error tiene una distribución conjunta normal. Junto a la hipótesis anterior se concluye que los errores son independientes entre sí[141].
5. El error no depende de las variables explicativas X[142].
6. Linealidad: El modelo definido previamente genera los valores de la variable dependiente[143].

$$Y = X * B + U \quad [10]$$

7. Homocedasticidad: se asume que todas las perturbaciones (errores) tienen la misma varianza, es decir, la varianza de cada perturbación es constante y se denota como[144]:

[135] Chirivella, V. (2015).
[136] Ibid.
[137] Abuín, J. (2007).
[138] Ibid.
[139] Ibid.
[140] Ibid.
[141] Ibid.
[142] Ibid.
[143] Abuín, J. (2007).
[144] Ibid.

$$V(ui) = \sigma 2 \qquad [11]$$

1

8. Independencia: Las perturbaciones aleatorias no están relacionadas entre sí[145].

$$E(ui \cdot u\, j) = ,0\ \forall i \neq j \quad [12]$$

9. Normalidad: se postula que la distribución de los errores sigue una distribución normal, es decir, los errores U se aproximan a la siguiente distribución normal[146]:

$$U \approx N,0\ (\sigma 2) \quad [13]$$

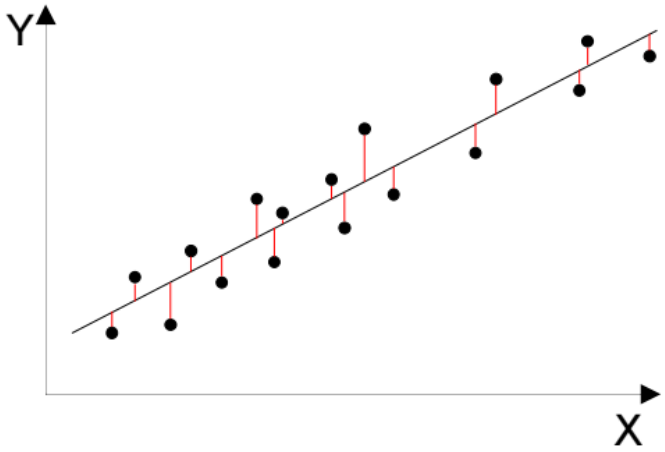
10. Precisión de las variables explicativas: las variables explicativas se consideran medidas con precisión perfecta y no tienen incertidumbre en sus valores[147].

Si se cumplen las hipótesis mencionadas anteriormente, el método de estimación de mínimos cuadrados ordinarios[148] proporciona estimadores óptimos[149]. Es decir, se obtienen estimadores eficaces y de alta calidad, centrados alrededor de los verdaderos valores poblacionales[150].

Mediante el MCO se pretende llevar a cabo la creación de un hiperplano que minimice la suma de los cuadrados de las distancias entre cada una de las observaciones de la variable y dicho hiperplano, también conocido como residuos[151]. Es decir, se pretende minimizar la distancia, en vertical, entre los distintos puntos representados en la figura 12 (observaciones) y la recta de regresión lineal.

[145] Ibid.
[146] Ibid.
[147] Ibid.
[148] MCO
[149] Abuín, J. (2007).
[150] Teorema de Gauss-Markov
[151] Abuín, J. (2007).

Figura 12: Recta ajustada y error cometido



Fuente: Chirivella, V. (2015).

En otras palabras, como se ha mencionado, U es un error, por lo que lo más conveniente sería que fuera lo más pequeño posible[152]. Sin embargo, dado que el error es una variable aleatoria significa que el objetivo es que su valor medio sea cero y su varianza lo más pequeña posible[153]. La forma más común de asegurar que se cumpla esto es la creación de un hiperplano óptimo, cuya idoneidad se comprueba al acumular los errores sumándolos[154]. Sin embargo, esta aproximación presenta limitaciones, ya que la suma de residuos positivos y negativos puede anularse mutuamente, dando lugar a un valor que no refleje realmente las distancias entre las observaciones y la recta[155]. Una solución a este problema es elevar los errores al cuadrado, eliminando así la distinción entre residuos positivos y negativos, y amplificando aún más los valores grandes, lo que penaliza los puntos distantes de la recta[156]. De esta forma se obtiene la expresión conocida como la suma de cuadrados de residuos[157]:

$$SCR = \sum_{j=1}^{n} e_j^2 = \sum_{j=1}^{n} (Y_j - \hat{Y}_j)^2 = \sum_{j=1}^{n} (Y_j - (b_0 + b_1 X_{1j} + b_2 X_{2j} + \cdots + b_k X_{kj}))^2 \quad [14]$$

Donde:

- e_j representa el residuo de cada observación[158]
- $\hat{Y}_j$ representa los valores de la variable dependiente estimados por el modelo de regresión

[152] Chirivella, V. (2015).
[153] Ibid.
[154] Ibid.
[155] Ibid.
[156] Ibid.
[157] Conocida por el acrónimo SCR
[158] Diferencia entre el valor observado y el valor estimado

- b representa la estimación de los parámetros β previamente mencionados.

La SCR proporciona una medida cuantitativa de la distancia entre la recta y los puntos, representantes de las observaciones. Un valor alto de SCR indica que los puntos están muy lejos de la recta, mientras que un valor bajo indica que están ajustados. Es decir, un mejor ajuste a las observaciones.

Una vez agrupadas las variables independientes, sus coeficientes y la variable explicada en sus respectivas matrices $\mathbf{X}[x_{ij}]$, $\mathbf{B}[b_j]$ e Y $[y_j]$ se deriva la ecuación 14 con respecto a B, obteniéndose la siguiente expresión[159]:

$$\frac{dRSS}{dB} = -2X^T(Y - XB) \quad [15]$$

Dado que lo que estamos buscando es el mínimo de la función RSS, es necesario igualar la ecuación 15 a 0 para calcularlo[160]. De esta forma es posible obtener:

$$\hat{B} = (X^T X)^{-1} X^T Y \quad [16]$$

$$\hat{Y} = X\,\hat{B} = (X^T X)^{-1} X^T Y \quad [17]$$

Siendo $\hat{B}$ la matriz correspondiente a los coeficientes estimados.

Una vez calculados estos parámetros es necesario medir la adecuación del modelo. Para ello se usan los siguientes métodos:

Por un lado, es necesario calcular lo que se conoce como coeficiente de determinación o R^2. Se trata de una medida que indica la cantidad de comportamiento o fluctuaciones de la variable de respuesta que se pueden atribuir al modelo, así como su capacidad para determinar cómo se relaciona con las variables predictoras[161]. Es decir, R2 muestra cómo el modelo se ajusta a los datos y cuánto de la variación de la variable de respuesta puede ser explicada por las variables predictoras del modelo[162].

$$R^2 = 1 - \frac{S_{RES}^2}{S_Y^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad [18]$$

R^2, varía entre 0 y 1. Un valor de 0 indica que el modelo no explica ninguna variabilidad de la variable de respuesta, mientras que un valor de 1 indica que el modelo explica toda la variabilidad, es decir se trata de un buen ajuste.

[159] López-Silvarrey, G. (2021).
[160] Ibid.
[161] Ibid.
[162] Ibid.

Por otro lado, a partir de la raíz del error cuadrático medio es posible medir la precisión del modelo. Siendo mejor modelo cuanto menor sea el valor de este.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad [19]$$

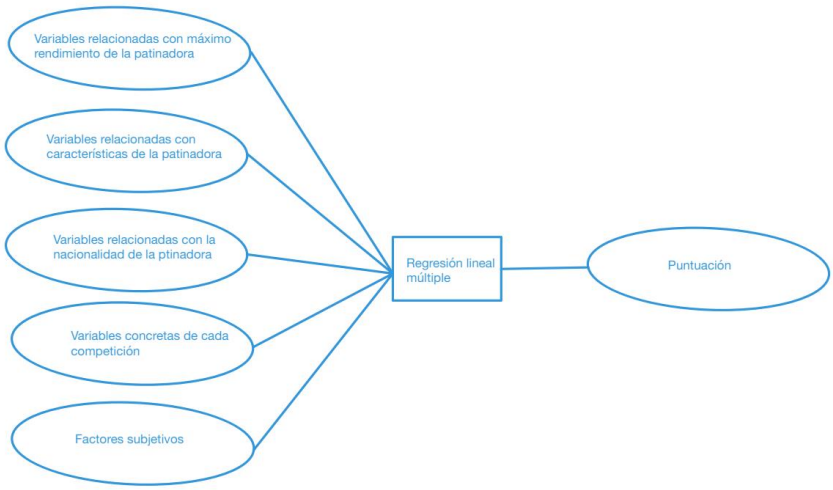
4.2 Creación del modelo

Una vez explicada la técnica a utilizar el capítulo continuará con una explicación del proceso de creación del modelo.

Por un lado, se toman como variables de entrada las diferentes variables que se consideran que pueden llegar a tener influencia sobre la variable explicada. Este proceso de selección de variables de entrada es conocido como *Enter* y consiste en una selección de variables manual[163]. Es decir, se crea un modelo inicial teniendo en cuenta todas las variables que se consideran participativas, para posteriormente ir descartando aquellas que resultan ser menos influyentes[164].

A continuación, la figura 13 muestra un esquema de la estructura del modelo. En esta se puede apreciar como las variables de entrada son divididas en distintas categorías según con qué aspecto de la competición estén relacionadas.

Figura 13: Estructura del modelo



Fuente: Elaboración propia

[163] Peláez, I. (2016).
[164] Ibid.

Las distintas categorías contienen las siguientes variables, las cuales han sido explicadas detalladamente en el capítulo anterior:

- Variables relacionadas con máximo rendimiento de la patinadora: Estas variables recogen datos sobre el momento en el que la patinadora alcanzó el pico de su carrera. Es decir, la competición en el que obtuvo su puntuación más alta hasta la fecha.

    1. x_1: Edad en la que la patinadora alcanza la puntuación más alta de su carrera. Esta variable es medida en años.
    2. x_2: Mejor marca personal hasta la fecha. Esta variable es medida en puntos.

- Características de la patinadora: Estas variables recogen datos concretos de cada patinadora a nivel más físico.

    1. x_4: Edad de la patinadora en el momento de la competición. Esta variable es medida en años.
    2. x_5: Altura de la patinadora. Esta variable es medida en centímetros.

- Variables relacionadas con la nacionalidad de la patinadora: Estas variables tienen en cuenta la nacionalidad de la patinadora, ya que esto puede afectar tanto a su puntuación como a su actuación.

    1. x_6: Variable binaria que determina si la patinadora compite en su país (x_6=1) o en el extranjero (x_6=0).
    2. x_7: Variable binaria que determina si la patinadora tiene la misma nacionalidad que un juez determinado (x_6=1) o no (x_6=0).

- Variables concretas de cada competición: Estas variables recogen datos sobre factores que no solo pueden cambiar en cada competición, sino que también pueden cambiar dentro de cada programa.
    1. x_8: Posición en la que compite la patinadora. Esta variable se mide sobre el número total de patinadoras en una determinada categoría.

- Factores subjetivos: Variables que recogen datos sobre factores que están en manos de los jueces. Estos pueden variar de un juez a otro al tratarse de distintos criterios.
    1. x_9: Habilidades sobre el hielo demostradas por la patinadora en el programa. Esta variable se mide sobre 10.
    2. x_{10}: Calidad de la presentación general de la patinadora Esta variable se mide sobre 10.

3. x_{11}: Estructura y diseño general del programa de patinaje. Esta variable se mide sobre 10.
4. x12: Transiciones entre los distintos elementos del programa. Esta variable se mide sobre 10.
5. x13: Capacidad de la patinadora para interpretar la música utilizada durante el programa. Esta variable se mide sobre 10.

Una vez importados estos datos es necesario normalizarlos para así evitar problemas con las escalas[165]. Para ello se utiliza la función "scale()", la cual se encarga de transformar los valores de las variables para que tengan una escala común y se centren alrededor de cero[166].

Cabe destacar que en este caso no se incluye la variable **tiempo** empleada en la base de datos original. En un principio se tuvo en cuenta dicha variable, pero una vez realizadas las estimaciones pertinentes se obtuvo como resultado un modelo el cual prescindía del tiempo que había pasado desde que la patinadora había logrado su mejor marca personal. Analizando esta situación y el modo de trabajo del software utilizado[167], se llega a la conclusión de que esto se debe a una dependencia entre esta variable y las conocidas como: x_1 y x_4.

A la hora de seleccionar las variables dependientes se tiene en cuenta otra técnica existente denominada *Stepwise*[168]*.* Esta técnica consiste en comenzar el modelo sin variables para después ir incluyendo o eliminando variables basándose en los criterios de entrada y salida[169]. En cada paso, se introduce una variable que cumpla los criterios de entrada y se evalúa si alguna variable cumple los criterios de salida, realizándose una valoración global del modelo en cada paso[170]. El proceso continúa iterativamente hasta que no haya más variables que cumplan los criterios de selección o eliminación, obteniéndose así un modelo final con las variables más relevantes y significativas[171]. Sin embargo, este método puede provocar problemas a la hora de aplicar el modelo a distintos datos.

El uso de ambos métodos nos permite comparar sus resultados, eligiendo así aquel que se adapte mejor tanto a los datos de entrenamiento como a los de prueba. De esta forma nos aseguramos de estar diseñando el modelo con la combinación óptima de variables.

[165] Quintanilla, L. (2023).
[166] Ibid.
[167] Rstudio
[168] Rodríguez-Jaume, J., et. al (2001).
[169] Ibid.
[170] Ibid.
[171] Ibid.

Por otro lado, tras aplicar el modelo de regresión lineal se obtiene una variable de salida explicada por aquellas previamente mencionadas. Esta variable dependiente indica la puntuación obtenida por la patinadora.

En cuanto al output, dependiendo de los datos introducidos en el modelo se puede obtener la puntuación de la patinadora para el programa corto o para el programa largo. En este caso se ha creado un modelo diferente para cada una de las situaciones ya que el número de participantes se reduce notoriamente en el segundo caso.

Por último, para valorar la precisión de este método se utilizan las medidas nombradas en el apartado anterior: el coeficiente de determinación y la raíz del error cuadrático medio. Junto los que se calcula una tercera medida conocida como gráfica de dispersión, la cual compara los valores reales con los previstos por el modelo[172]. De esta forma se puede comprobar si las predicciones se ajustan bien a los valores reales, ya que cuanto mejor es este ajuste más se acerca a una línea diagonal[173].

En resumen, al analizar la teoría implicada en los modelos de regresión lineal múltiple estos parecen adecuarse correctamente al ejemplo de las competiciones de patinaje artístico sobre hielo. Sin embargo, esto no puede darse por hecho simplemente basándonos en la teoría, por lo que más adelante se aplica este modelo de forma práctica para así poder comprobar esto.

En un intento de seleccionar el modelo que mejor se ajuste a esta situación se pretende estudiar la aplicación de otro método conocido en los problemas de regresión, las máquinas de vectores de soporte[174]. De esta forma, una vez aplicadas de manera práctica ambas técnicas es posible comparar la eficacia y precisión de ambos modelos para así seleccionar la mejor opción.

[172] Landajuela, I. (2019).
[173] Ibid.
[174] Suárez, E. (2014).

Capítulo 5: Máquinas de vectores de soporte

La segunda técnica para crear modelos predictivos que se emplea en este trabajo es la conocida como máquinas de vectores de soporte. El objetivo principal de este capítulo es proporcionar una explicación detallada del modelo de regresión lineal y su aplicación teórica.

A lo largo de este capítulo, se pretende proporcionar una explicación clara tanto del modelo como de su adaptación al mundo del patinaje artístico sobre hielo. De esta forma se sentará las bases para la posterior comparación con el método explicado previamente.

5.1 Fundamento teórico

Las SVMs son una técnica conocida por su gran capacidad de adaptación, ya que pueden ser utilizadas para resolver tanto problemas de clasificación binaria, su uso original, como problemas de regresión, agrupamiento o multiclasificación, entre otros[175]. Esta adaptabilidad junto con sus fundamentos teóricos ha permitido el reconocimiento de esta técnica durante los últimos años[176].

En este caso se utilizan las máquinas de vectores de soporte para un problema de regresión, por lo que se las designa como Regresión de vectores de soporte[177]. El objetivo de esta técnica es la creación del hiperplano que mejor se ajuste a los datos de entrenamiento[178].

$$f(x) = (w_1 x_1 + \cdots + w_d x_d) + b \leq w, x > +b \qquad [20]$$

Donde:

- w representa los pesos de la función lineal, parámetros a estimar.
- x representa las variables de entrada.
- b representa el sesgo asociado a la función.

El objetivo es que esta función esté lo más próxima posible a los puntos de los datos de entrenamiento. Sin embargo, en muchos casos, los datos no pueden ser perfectamente ajustados a la función de regresión, por lo que se permite cierto grado de error o tolerancia alrededor de esta[179]. Esta tolerancia se controla mediante la función de pérdida ε -insensible, la cual penaliza los errores de estimación por encima del umbral[180].

[175] Martín Guareño, J. J. (2016).
[176] Suárez, E. (2014).
[177] Support Vector Regression (SVR)
[178] Suárez, E. (2014).
[179] Suárez, E. (2014).
[180] Martín Guareño, J. (2016).

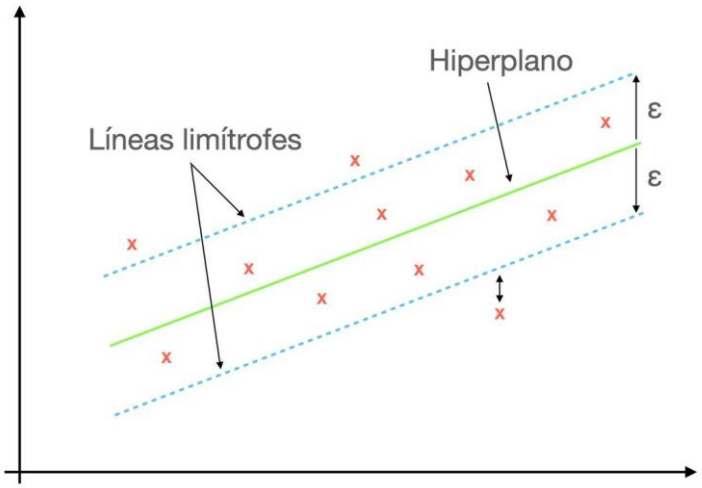
$$L_\varepsilon(y, f(x)) = \begin{cases} 0 & si\ |y - f(x)| \leq \varepsilon \\ |y - f(x)| - \varepsilon & en\ otro\ caso \end{cases} \qquad [21]$$

Donde:

- $f(x)$ representa la predicción de la función.
- y representa el valor real de la variable.

Esta función considera que los errores menores o iguales a ε son aceptables[181], por lo que aquellos datos pertenecientes a esta zona definida por $\pm\varepsilon$ no se consideran vectores de soporte[182]. Es decir, aquellos situados en los límites de este margen creado por ε alrededor de la función de regresión son los más influyentes a la hora de definir la función de regresión[183].

Figura 14: Funcionamiento básico de las SVRs

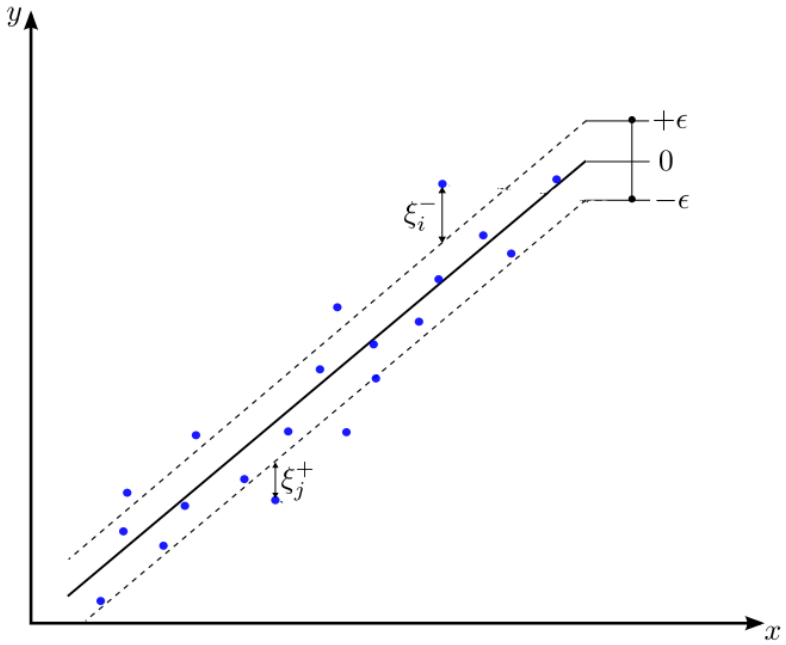


Fuente: Rodríguez, D., & Rodríguez, D. (2021)

Una manera de cuantificar aquellos errores que no se consideran aceptables es el uso de las variables de holgura[184] ε_i^+ y ε_i^-. Estas variables representan la medida de desviación o distancia de un punto de datos con respecto al margen o al hiperplano de regresión[185]. Es decir, si un punto de datos se ajusta adecuadamente al margen o cumple con los criterios épsilon-insensibles, las variables de holgura correspondientes son cero, pero si un punto de datos se desvía del margen las variables de holgura reflejan la magnitud de esta desviación[186].

[181] Ibid.
[182] Se trata de los puntos más cercanos al margen y que tienen una mayor influencia en la construcción del modelo.
[183] Sanabria-Castro, A., et. al. (2023).
[184] Suárez, E. (2014).
[185] Ibid.
[186] Ibid.

Figura 15: Variables de holgura en SVR



Fuente: Suárez, E. (2014)

La suma de estas variables proporciona una métrica para evaluar el costo relacionado con la cantidad de ejemplos que presentan errores de predicción significativos[187]. Por lo tanto, la minimización de la suma de las variables de holgura tiene como objetivo reducir el impacto de los ejemplos con errores de predicción significativos en la calidad de la regresión [188].

Esta condición se compagina[189] con el objetivo de reducir el riesgo estructural[190]. Este se consigue mediante el margen más plano. Es decir, reduciendo el valor de la norma cuadrada del vector de pesos w^T. Esta combinación de requisitos lleva a la siguiente expresión[191]:

$$\min \frac{1}{2}||\vec{w}||^2 + C\sum_{i=1}^{n}(\varepsilon_i^+ + \varepsilon_i^-) \qquad [22]$$

[187] Ibid.
[188] Ibid.
[189] Sanabria-Castro, A., et. al. (2023).
[190] Equilibrio entre el ajuste a los datos de entrenamiento y la capacidad de generalización del modelo.
[191] Sanabria-Castro, A., et. al. (2023).

$$s.a. \qquad y_i - (w^T x_i + b) \leq \varepsilon + \varepsilon_i^- \qquad i = 1 \dots n$$

$$(w^T x_i + b) - y_i \leq \varepsilon + \varepsilon_i^+ \qquad i = 1 \dots . n \qquad [23]$$

$$\varepsilon_i^+, \varepsilon_i^- \geq 0 \quad i = 1 \dots n$$

Donde:

- y_i es el valor real de la variable
- C es el parámetro de costo, el cual controla la relación entre la regularidad de función y la tolerancia permitida para los errores[192].

En esta ecuación se puede distinguir una primera parte centrada en la reducción del riesgo estructural y una segunda cuyo objetivo es reducir la suma de las variables de holgura. Combinando ambas condiciones en una misma ecuación.

Mediante la utilización de la teoría de optimización lineal y la teoría de la dualidad, es posible obtener un estimador relacionado para la función lineal en SVR[193]:

$$f(x) = \sum_{i=1}^{n} (\alpha_i^- - \alpha_i^+) < x, x_i > + b^* \qquad [24]$$

Donde α_i^- y α_i^+ son las soluciones del problema dual, los cuales sirven para evaluar la importancia de cada punto de datos en la estimación de la función lineal en SVR[194]. Una vez obtenida esta expresión es posible resolver el problema dual empleando la función 'svm' en Rstudio, encargada de obtener los valores óptimos del parámetro w.

A partir de este problema dual, junto con las condiciones de complementariedad de holgura de Karush-Kuhn-Tucker[195], también es posible la obtención del valor del sesgo[196] b^*, ver ecuaciones 25 y 26.

$$b^* = y_i + \varepsilon - < w^*, x_{i>} \quad si\ 0 < \alpha_i^+ < C \qquad [25]$$

$$b^* = y_i + \varepsilon - < w^*, x_{i>} \quad si\ 0 < \alpha_i^- < C \qquad [26]$$

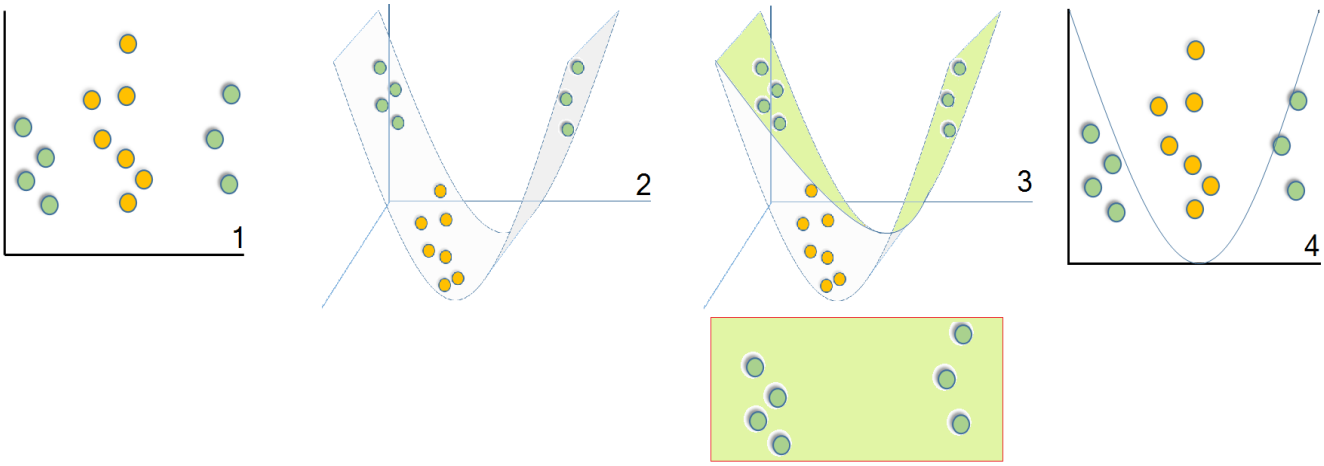
Todo lo explicado previamente corresponde al fundamento teórico de las SVR. Sin embargo, al tratarse de un caso en el que no se puede dar por hecho que los datos

[192] Martín Guareño, J. (2016).
[193] Ibid.
[194] Ibid.
[195] Conjunto de condiciones esenciales que deben cumplirse para que una solución sea considerada óptima en problemas de optimización convexa con restricciones.
[196] Martín Guareño, J. (2016).

puedan ajustarse por una función lineal en el espacio original, es necesario una ligera modificación.

Para solucionar el problema previamente mencionado se deben transformar las variables de entrada en un nuevo espacio que permita ajustarlos mediante un regresor lineal[197]. Para lo que es necesario el uso de la función Kernel.

Figura 16: Ejemplo sencillo de utilización de función Kernel



Fuente: Numerentur.org. *(s. f.).*

Este método resulta ser especialmente útil cuando se trabaja con una base de datos de alta dimensionalidad, adaptando así el estimador explicado en la ecuación[198] 24.

$$f(x) = \sum_{i=1}^{n} (\alpha_i^- - \alpha_i^+) K(x, x_i) \qquad [27]$$

A pesar de existir una amplia variedad de funciones Kernel, en este ejemplo se ha optado por el uso de un Kernel radial. Este se considera uno de los más comunes y sencillos de aplicar al modelo al tratarse de una técnica que resulta útil en una amplia cantidad de casos[199]. Lo cual se debe a su capacidad de adaptarse a diversas formas de distribución de datos y su gran capacidad de manejo de datos 200.

Esta función se encarga de calcular la similitud entre dos puntos, X_1 y X_2 , determinados o la distancia entre dichos puntos. Puede ser expresada de la siguiente manera[201]:

[197] Suárez, E. (2014).
[198] Ibid.
[199] Sreenivasa, S. (2021, 16 diciembre).
[200] Ibid.
[201] Rodrigo, J. (s.f.)

$$K(X_1, X_2) = e^{(-\gamma||X_1 X_2||^2)} \qquad [28]$$

Donde:

- γ es el hiperparámetro de varianza en el Kernel, encargado de controlar la influencia de la distancia entre los puntos.
- $||X_1 - X_2||$ es la medida de distancia Euclidiana entre ambos puntos.

5.2 Creación del modelo

Una vez explicada la técnica a utilizar el capítulo continuará con una explicación del proceso de creación del modelo.

Para ello se utilizan las siguientes variables[202]:

- x_1: Edad en la que la patinadora alcanza la puntuación más alta de su carrera. Esta variable es medida en años.
- x_2: Mejor marca personal hasta la fecha. Esta variable es medida en puntos.
- x_3: Cantidad de tiempo transcurrido desde que se obtuvo la mejor marca personal hasta el momento de la competición. Esta variable es medida en años.
- x_4: Edad de la patinadora en el momento de la competición. Esta variable es medida en años.
- x_5: Altura de la patinadora. Esta variable es medida en centímetros.
- x_6: Variable binaria que determina si la patinadora compite en su país (x_6=1) o en el extranjero (x_6=0).
- x_7: Variable binaria que determina si la patinadora tiene la misma nacionalidad que un juez determinado (x_6=1) o no (x_6=0).
- x_8: Posición en la que compite la patinadora. Esta variable se mide sobre el número total de patinadoras en una determinada categoría.
- x_9: Habilidades sobre el hielo demostradas por la patinadora en el programa. Esta variable se mide sobre 10.
- x_{10}: Calidad de la presentación general de la patinadora Esta variable se mide sobre 10.
- x_{11}: Estructura y diseño general del programa de patinaje. Esta variable se mide sobre 10.
- x12: Transiciones entre los distintos elementos del programa. Esta variable se mide sobre 10.
- x13: Capacidad de la patinadora para interpretar la música utilizada durante el programa. Esta variable se mide sobre 10.

Una vez importados estos datos es necesario normalizarlos para así evitar problemas con las escalas[203]. Para ello se utiliza la función “scale()”, la cual se encarga de transformar

[202] Todas las variables que han sido explicadas en el capítulo 2.
[203] Quintanilla, L. (2023)

los valores de las variables para que tengan una escala común y se centren alrededor de cero[204].

Después de la manipulación y división de datos en los conjuntos de datos de entrenamiento y prueba[205] se puede comenzar a crear el modelo. Para ello, se utiliza la librería "e1071" en Rstudio la cual proporciona funciones y métodos útiles a la hora de entrenar y realizar predicciones con modelos de SVM y SVR[206].

Como ya se ha mencionado en el apartado anterior, la creación del modelo se lleva a cabo utilizando una función de Kernel radial. A la hora de usar esta función es necesario definir el hiperparámetro gamma, el cual puede variar ampliamente dependiendo de los datos y del modelo descrito[207]. En este caso, tras comprobar el efecto de diferentes valores de gamma, se llega a la conclusión de que un valor válido es $\gamma = 0.1$, ya que cuanto menor sea este parámetro más suavizadas están las predicciones del modelo[208].

Por otro lado, también es necesario definir los parámetros C y ε, lo cual se ha llevado a cabo a través de una búsqueda de cuadrícula. Esta técnica implica explorar diversas combinaciones de valores para los parámetros seleccionados y evaluar la precisión del modelo para cada una de estas combinaciones. Una vez realizada esta operación se obtienen C= 256 y ε=0, los valores óptimos de estos parámetros[209].

Una vez definidos los valores de todos los parámetros necesarios es posible entrenar el modelo a partir de los datos de entrenamiento. Se utilizan todas las variables explicadas en el capítulo 2 para predecir de la manera más precisa posible la puntuación de una patinadora en un programa determinado.

A continuación, es necesario medir la eficacia del modelo para así compararlo con el resto de los modelos y poder elegir el que mejor se adapte a este ejemplo. Para ello, se realiza la evaluación del rendimiento del modelo en los datos de entrenamiento y prueba, utilizando las medidas explicadas en el capítulo anterior:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad [29]$$

[204] Ibid.
[205] Conjuntos necesarios para el diseño del modelo explicados en el capítulo 2 de este trabajo.
[206] Rodrigo, J. (s.f.)
[207]Vaquerizo. R. (2020).
[208]Ibid.
[209] Se espera que las predicciones sean exactas.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad [30]$$

Junto a estas medidas se calcula la conocida como gráfica de dispersión, la cual compara los valores reales con los previstos por el modelo[210]. De esta forma se puede comprobar si las predicciones se ajustan bien a los valores reales, ya que cuanto mejor es este ajuste más se acerca a una línea diagonal[211].

En resumen, según la teoría estudiada esta segunda técnica parece poder adaptarse correctamente a una base de datos compleja para así crear un modelo de regresión óptimo. Sin embargo, igual que en el caso anterior, esto no puede asegurarse sin una comprobación previa. Por ello es necesario plantear el modelo de forma práctica y analizar sus resultados.

Una vez aplicado este segundo método es posible estudiar su precisión y eficacia, permitiéndonos comparar ambos modelos. De esta forma, es posible comparar ambas técnicas permitiéndonos así elegir la mejor opción de las dos.

[210] Landajuela, I. (2019
[211] Ibid.

CAPÍTULO 6: ANÁLISIS DE RESULTADOS

Este capítulo se centra en el análisis de los resultados obtenidos tras la creación y aplicación de los modelos estudiados en este trabajo. En primer lugar, se comprueban una a una algunas de las hipótesis formuladas en el capítulo 3 para luego discutir los resultados obtenidos por ambos métodos.

6.1 Validación de hipótesis

Durante el capítulo 3 se ha llevado a cabo la formulación de varias hipótesis relacionadas con las distintas variables involucradas en el modelo. Estas hipótesis son corroboradas durante esta parte del trabajo, pudiéndose así verificar su importancia.

En primer lugar, se analiza la dependencia entre algunas variables mediante diagramas de dispersión, una forma de analizar el efecto de estas sin necesidad de conocer los resultados de los modelos. A continuación, se hace un análisis exhaustivo de los resultados obtenidos por los distintos métodos para poder complementar la información extraída de los diagramas de dispersión.

6.1.1 Efecto de la edad

La primera hipótesis que se plantea durante el diseño de este modelo está relacionada con la edad: Se espera que las patinadoras de menor edad alcancen una puntuación más alta.

La dificultad de los paso, piruetas y giros que un patinador realiza durante su carrera suponen un desgaste físico descomunal[212]. Esto, sumado al impacto del cuerpo contra una superficie tan complicada como el hielo lleva a numerosas lesiones como la tendinitis sobre todo en zonas como pies, tobillos, rodillas, caderas y espalda[213]. Todo ello lleva a que las carreras deportivas en este deporte sean relativamente cortas[214].

El efecto de este desgaste se puede comprobar en casos como el de Javier Fernández, quien tras haber conseguido ser campeón de Europa siete veces consecutivas y dos veces campeón del mundo, tuvo que retirarse en 2019 con tan solo 28 años[215]. Esta situación es aún más marcada en el patinaje femenino, donde las carreras de las deportistas son más cortas aún[216].

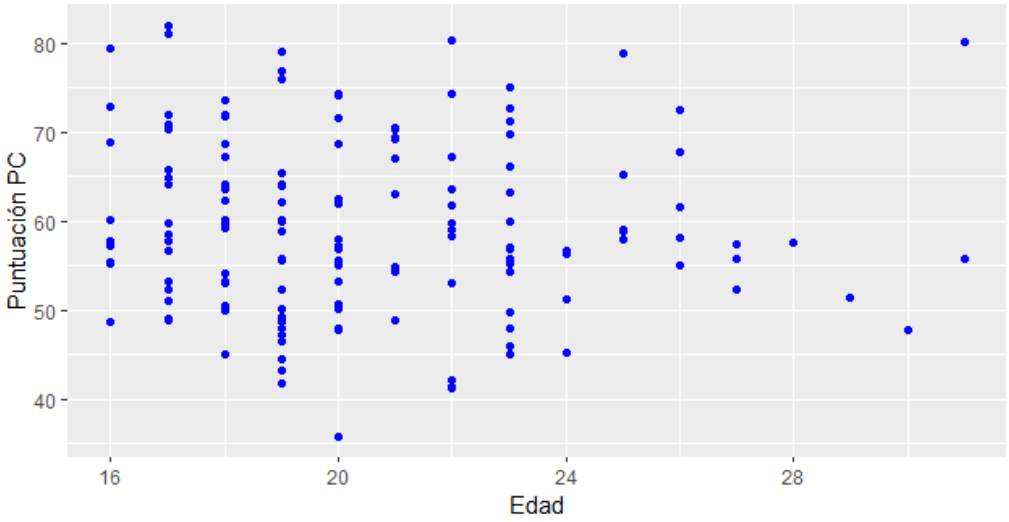
Los cambios físicos que sufre una mujer al alcanzar la pubertad pueden provocar problemas a la patinadora a la hora de rotar con agilidad o de saltar con altura[217]. Esto, sumado al hecho de que los huesos de estas son más frágiles que los de sus compañeros masculinos provocan que sus cuerpos no puedan realizar rutinas de alta dificultad

[212] Castilla, J. (2018).
[213] Ibid.
[214] Martín, A. (2022).
[215] Castilla, J. (2018).
[216] Martín, A. (2022).
[217] Ibid.

durante un largo periodo de tiempo[218]. Por ello se cree que cuanto menor es la edad mejor es la puntuación.

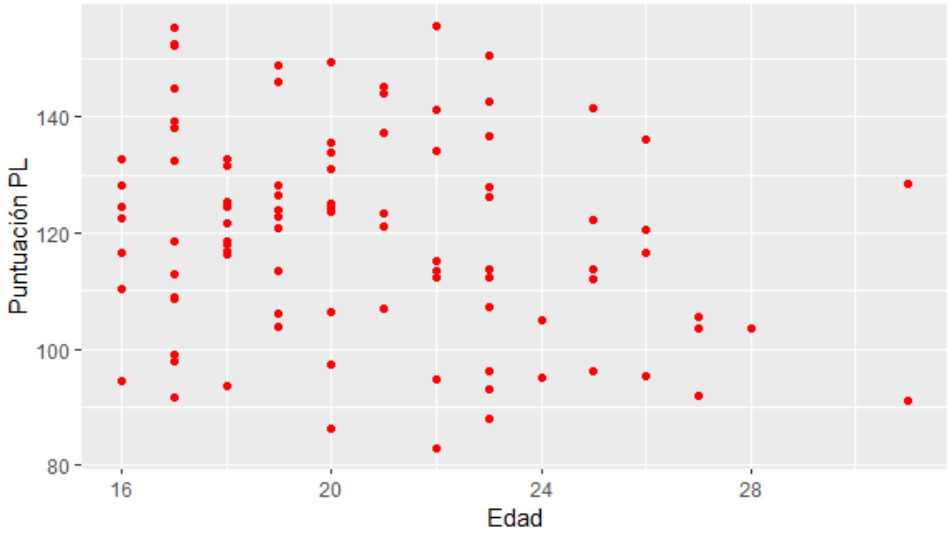
A la hora de demostrar estos factores se han empleado diagramas de dispersión. Esta herramienta permite analizar la correlación existente entre un par de determinadas variables, pudiéndose analizar así las tendencias relacionadas con la edad de las patinadoras[219].

Figura 17: Diagrama de dispersión- Puntuación programa corto vs. Edad



Fuente: Elaboración propia (2023)

Figura 18: Diagrama de dispersión-Puntuación programa largo vs. Edad



Fuente: Elaboración propia (2023)

[218] Ibid.
[219] Pértega, S., et. al. (2001).

A partir de estas dos gráficas es posible analizar la influencia de la edad sobre la puntuación de las patinadoras. Pudiéndose observar que no existe una relación inversamente proporcional entre la edad y la puntuación de las deportistas como se había teorizado en un principio. Sin embargo, sí que existe un número significativamente mayor de patinadoras jóvenes, quienes de forma general parecen obtener mejores resultados que las mayores de 23 años.

En la primera gráfica se observa que, pese a haber una distribución de edades de entre 16 y 31 años, existe un pico en el rango de edad entre 17 y 23 años. A partir de esta última, a medida que aumenta la edad disminuye drásticamente el número de patinadoras.

Por otro lado, analizando los datos del programa largo se puede observar como la mayoría de las patinadoras están concentradas también en un rango de edad de entre 17 y 23 años. Disminuyendo una vez más el número de estas una vez alcanzan esa edad.

Tabla 1: Edades de las patinadoras en el programa corto

Edad	Número
16	81
17	171
18	153
19	198
20	162
21	81
22	108
23	162
24	36
25	45
26	45
27	27
28	9
29	9
30	9
31	18

Fuente: Elaboración propia

Tabla 2:Edades de las patinadoras en el programa largo

Edad	Número
16	63
17	135
18	99
19	90
20	90

21	54
22	72
23	99
24	18
25	45
26	36
27	27
28	9
29	0
30	0
31	18

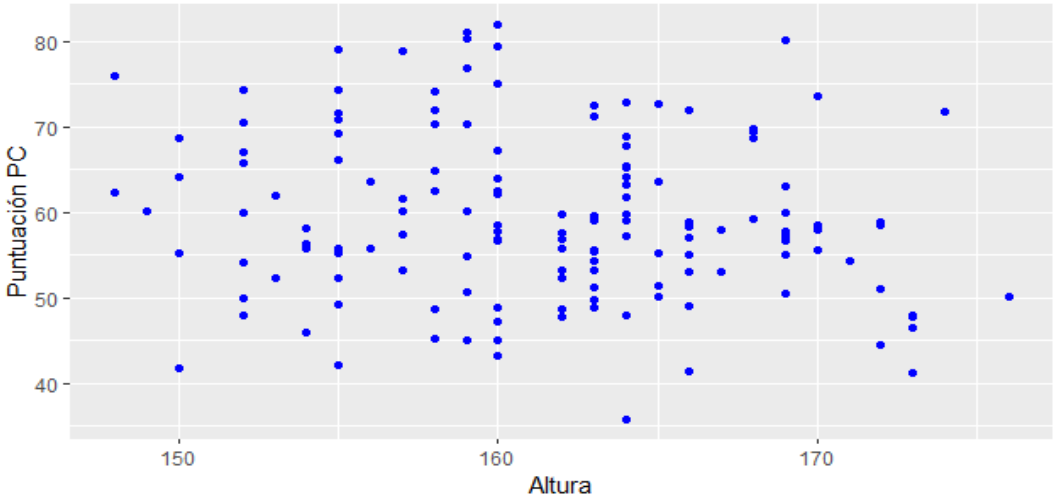
Fuente: Elaboración propia

En resumen, no se puede afirmar que cuanto menor sea una patinadora mayor es su puntuación, ya que las gráficas de dispersión representadas en las figuras 17 y 18 muestran una gran variedad en su distribución. Aun así, se puede observar que el patinaje es un deporte donde predominan las competidoras jóvenes y que las mejores puntuaciones son obtenidas en su mayoría por aquellas que no superan los 23 años, lo que apoya los testimonios de otros autores en los que aseguran la importancia de esta variable[220].

6.1.2 Efecto de la altura

La siguiente hipótesis formulada en capítulos anteriores es la relación inversamente proporcional entre de la altura y los resultados finales. Es decir, se cree que cuanto más menuda es la patinadora más alta es su puntuación. Esto se puede comprobar a partir de las siguientes gráficas:

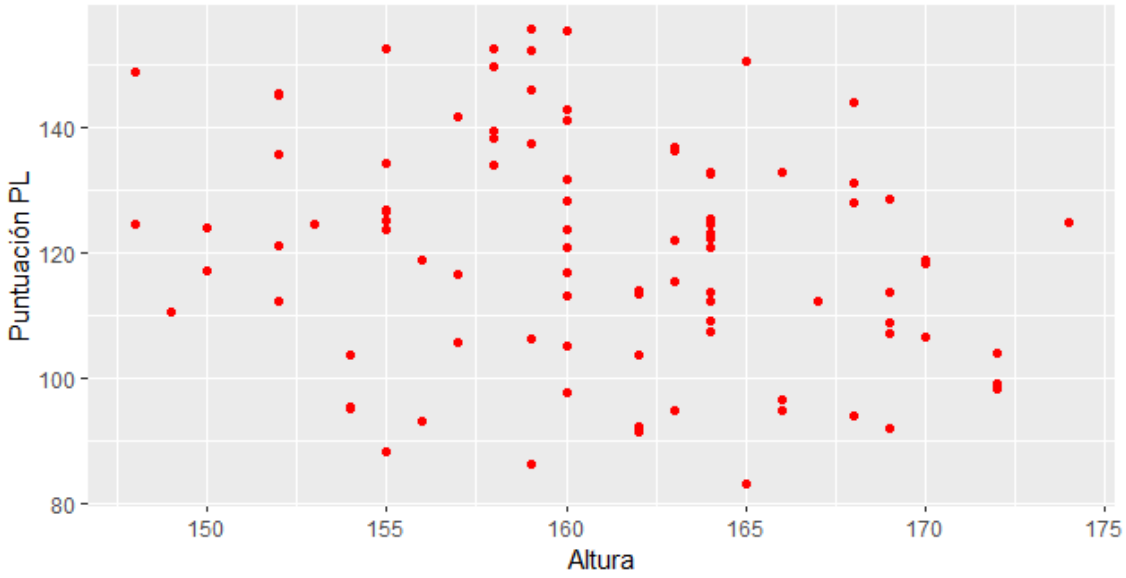
Figura 19: Diagrama de dispersión- Puntuación programa corto vs. altura



Fuente: Elaboración propia (2023)

[220] Autores como Castilla, J. (2018) o Martín, A.

Figura 20: Diagrama de dispersión- Puntuación programa largo vs. altura



Fuente: Elaboración propia (2023)

A partir de estas dos gráficas se puede comprobar que no se cumple la suposición inicial. No se puede apreciar una relación inversamente proporcional entre ambas variables, ya que los datos siguen una dispersión aleatoria.

Sin embargo, ambas gráficas muestran una mayor concentración de datos en el rango menor de 164 centímetros, disminuyendo notablemente el número de patinadoras una vez superada esta altura. Es decir, aunque no se pueda confirmar la hipótesis inicial, si que se puede comprobar que las afirmaciones aportadas por otros estudios[221] en los que se asegura que el patinaje artístico es un deporte practicado en su mayoría por deportistas menudas.

Otro factor que se puede apreciar en estas gráficas es que las puntuaciones más altas pertenecen a los rangos de altura menores de 170 centímetros. Pudiéndose comprobar que, pese a no existir una relación perfectamente definida entre ambas variables, el hecho de no exceder una determinada altura sí que puede ser decisivo.

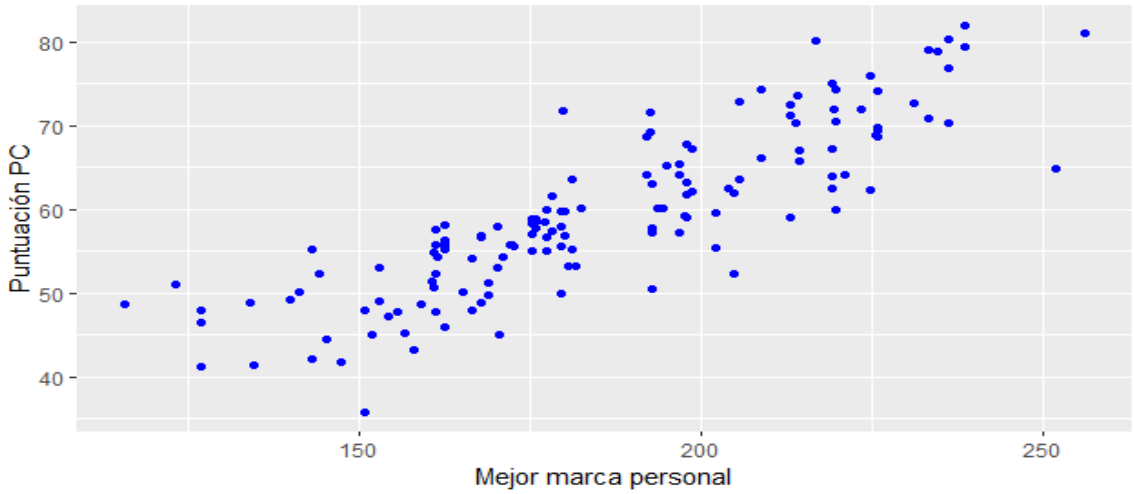
En definitiva, pese a no existir una relación lineal entre ambas variables, si se puede asegurar la importancia de la altura a la hora de dedicarse profesionalmente a este deporte.

6.1.3 Efecto de la mejor marca personal

A continuación, se va a comprobar la siguiente hipótesis formulada durante la creación del modelo: Se espera que aquellas patinadoras con mejores resultados históricos logren mejores puestos.

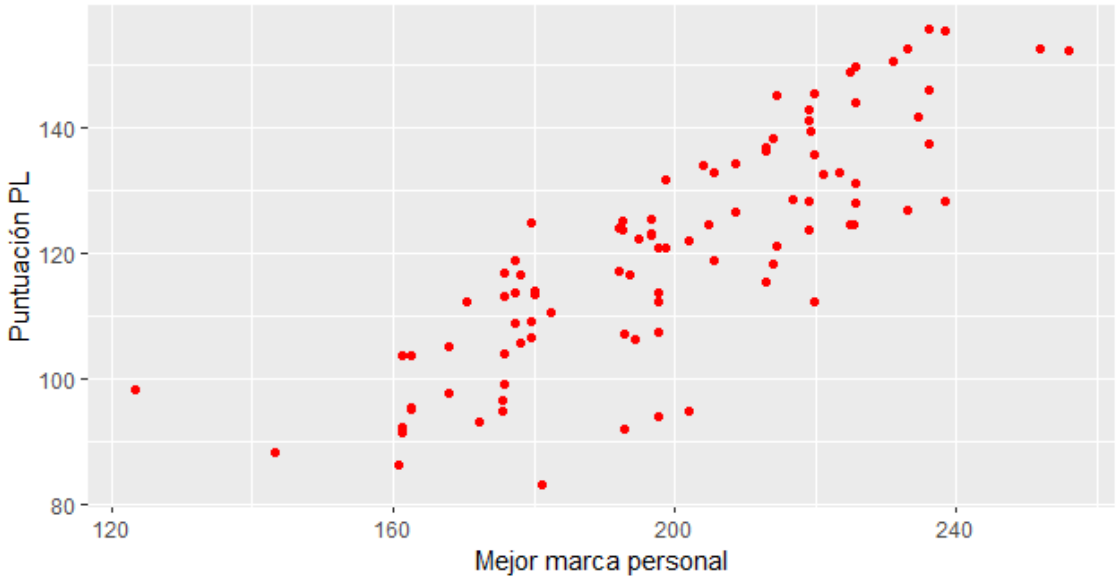
[221] Autores como Castilla, J. (2018) o Martín, A.

Figura 21: Diagrama de dispersión- Puntuación programa corto vs. Mejor marca personal



Fuente: Elaboración propia (2023)

Figura 22: Diagrama de dispersión- Puntuación programa largo vs. Mejor marca personal



Fuente: Elaboración propia (2023)

Gracias a ambas gráficas es posible observar cierta relación entre las dos variables estudiadas. Tanto en el programa corto como en el programa largo se aprecia una relación lineal positiva entre la mejor marca personal y la puntuación de dicho programa, pudiéndose confirmar así lo teorizado inicialmente.

Como es normal, en estas gráficas también se pueden observar datos que no encajan en esta relación lineal. Estas excepciones se deben a que durante la ejecución de un

programa siempre existe la posibilidad de que una patinadora cometa errores y no actúe de acuerdo con su nivel usual[222].

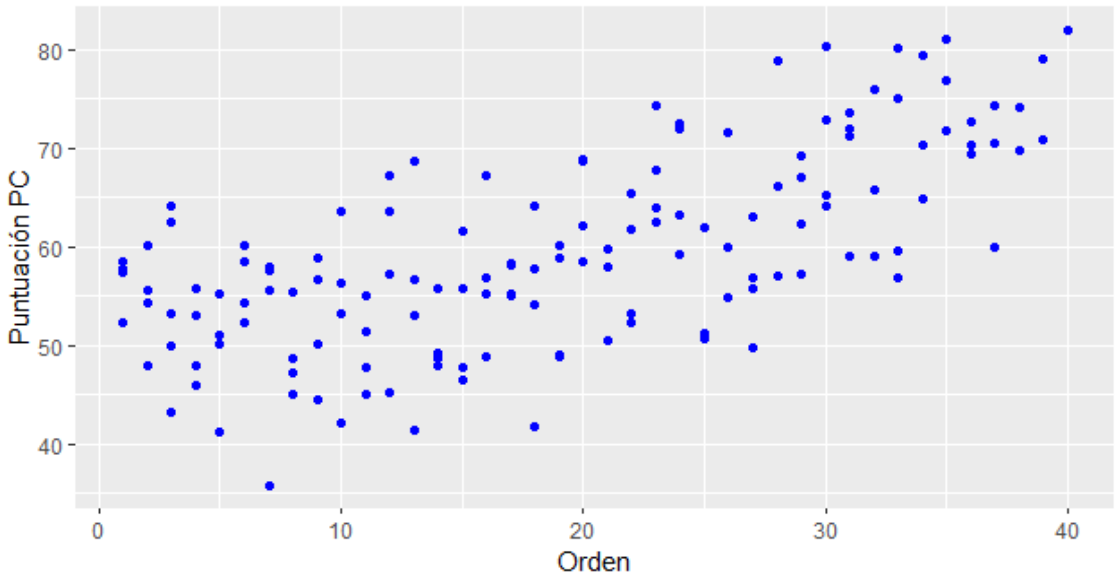
En conclusión, pese a existir ciertas excepciones en las que patinadoras con mejores marcas personales elevadas consiguen una puntuación más baja, sí que se puede afirmar una relación directamente proporcional entre las variables.

6.1.4 Efecto del orden

Otro factor que se teoriza en un principio es el efecto del orden de competición sobre la puntuación de una patinadora. Se espera que existan ciertos puestos que beneficien o perjudiquen a las deportistas. Es decir, se cree que aquellas patinadoras que ocupan los últimos puestos en el orden de la competición se ven beneficiadas ante aquellas que ocupan los primeros.

Las siguientes gráficas muestran la relación entre ambas variables, permitiéndonos analizar y entender la relación entre las dos variables mencionadas previamente:

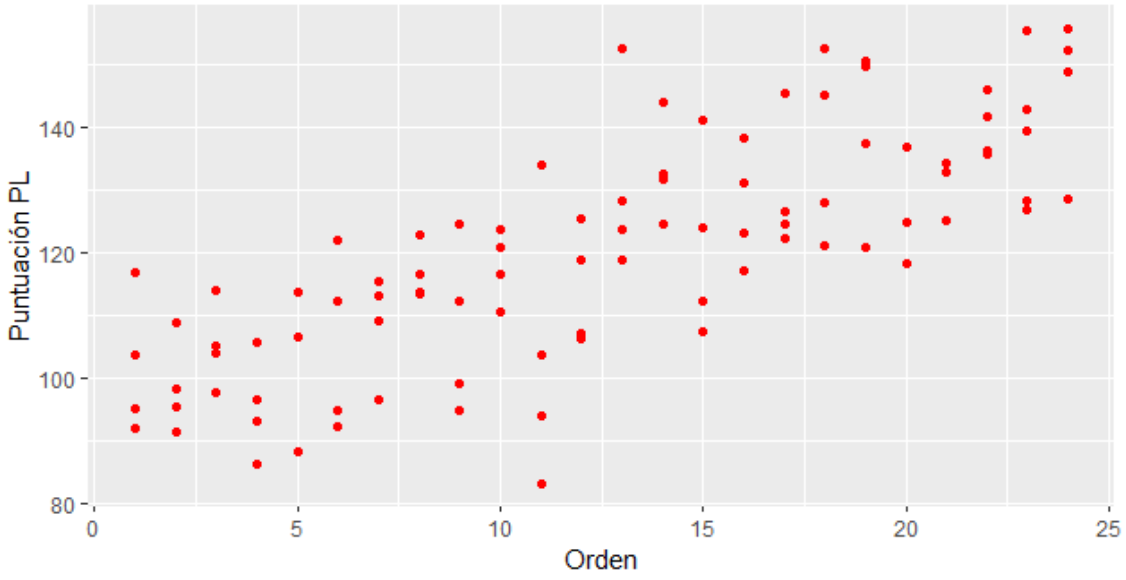
Figura 23: Diagrama de dispersión- Puntuación programa corto vs. orden



Fuente: Elaboración propia (2023)

[222] Lemons, M. Q. (2003).

Figura 24: Diagrama de dispersión- Puntuación programa corto vs. orden



Fuente: Elaboración propia (2023)

Ambas gráficas muestran una relación lineal positiva entre las variables estudiadas. Pese a no ser una relación excesivamente fuerte, es posible confirmar la hipótesis formulada en un principio. Es decir, aquellas patinadoras que compiten en los últimos grupos se ven beneficiadas ante aquellas que compiten primero.

Por otro lado, gracias a las gráficas se puede observar una relación que no es nombrada en la hipótesis inicial, ya que esta se centra en los primeros y últimos puestos de la competición. Los puestos más intermedios en el orden de la competición también parecen ajustarse a esta relación directamente proporcional mencionada previamente. Demostrando que ningún puesto es indiferente y que según a que grupo pertenezca una patinadora pertenece a un grupo u otro se puede ver beneficiada.

En este caso, igual que en el apartado explicado anteriormente, se pueden distinguir una serie de excepciones. Esto se debe una vez más a la posibilidad de que una patinadora cometa un error crítico durante su actuación[223]. Esta posibilidad siempre está ahí y es independiente de la posición en la que compita la deportista[224].

En definitiva, gracias a estas gráficas se puede concluir que el orden en el que se patina puede llegar afectar a la posición final obtenida por la patinadora.

[223] Ibid.
[224] Ibid.

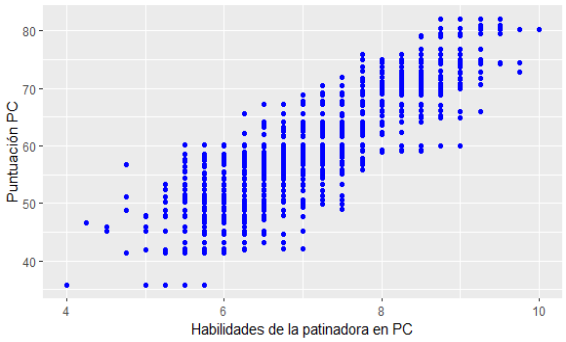
6.1.5 Efecto de la parte artística

En un principio se formula la siguiente hipótesis: Se prevé que las patinadoras que obtienen mejor puntuación final son aquellas que trabajan en perfeccionar la parte artística de su programa.

A continuación, se encuentran distintas gráficas que representan la relación entre los distintos elementos relacionados con la parte subjetiva y la puntuación de la patinadora para los distintos programas de la patinadora.

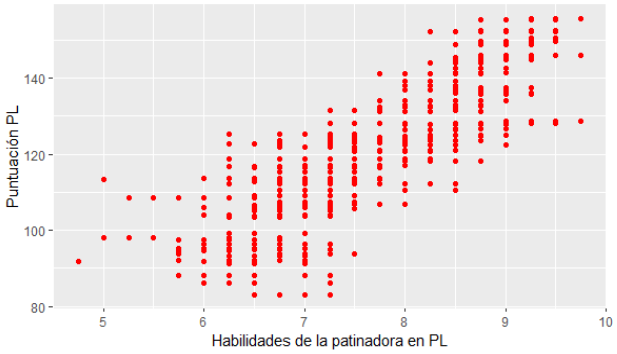
Habilidades de la patinadora:

Figura 25: Diagrama de dispersión- Puntuación programa corto vs. habilidades



Fuente: Elaboración propia

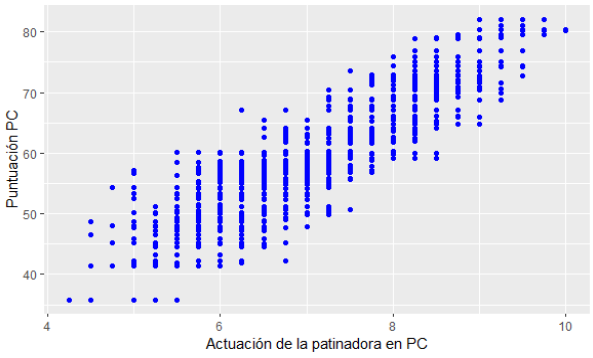
Figura 26: Diagrama de dispersión- Puntuación programa largo vs. habilidades



Fuente: Elaboración propia

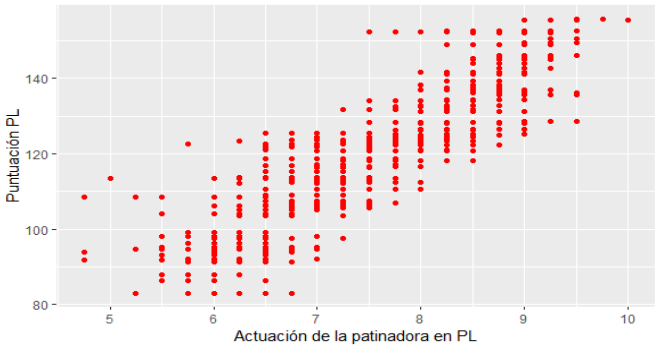
Actuación de la patinadora:

Figura 27: Diagrama de dispersión- Puntuación programa corto vs. actuación



Fuente: Elaboración propia (2023)

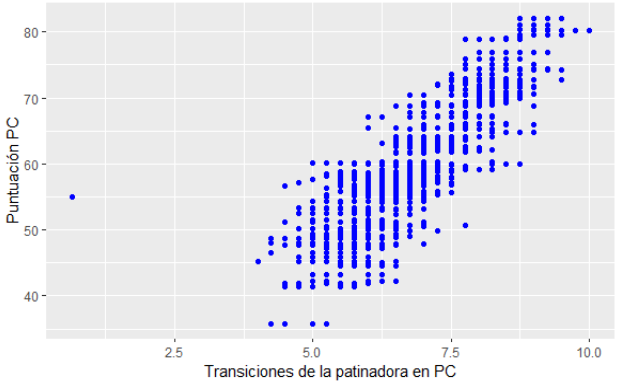
Figura 28: Diagrama de dispersión- Puntuación programa largo vs. actuación



Fuente: Elaboración propia (2023)

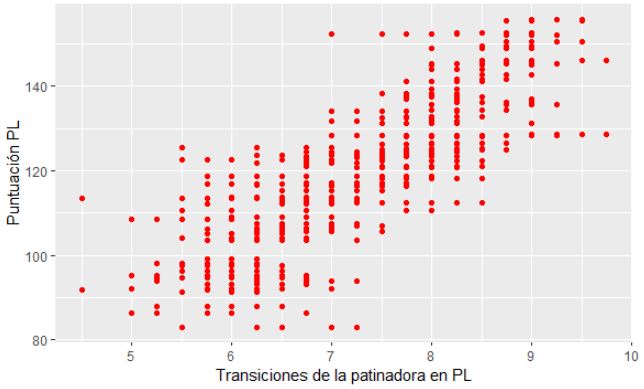
Transiciones de la patinadora:

Figura 29: Diagrama de dispersión- Puntuación programa corto vs. transiciones



Fuente: Elaboración propia (2023)

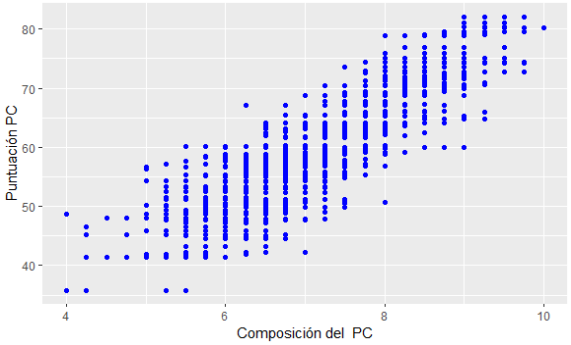
Figura 30: Diagrama de dispersión- Puntuación programa largo vs. transiciones



Fuente: Elaboración propia (2023)

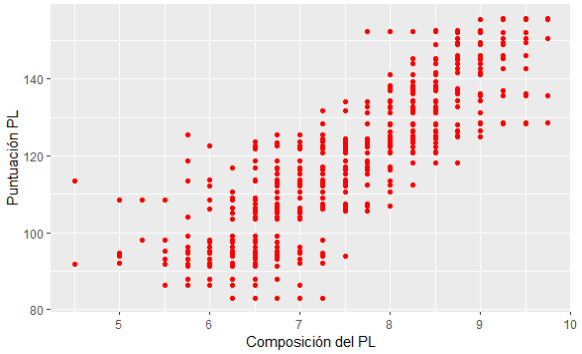
Composición del programa:

Figura 31: Diagrama de dispersión- Puntuación programa corto vs. composición



Fuente: Elaboración propia (2023)

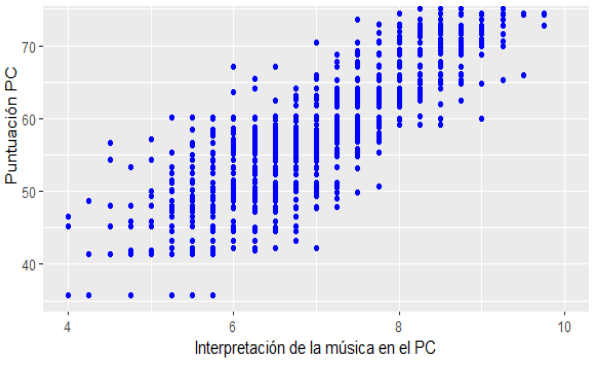
Figura 32 Diagrama de dispersión- Puntuación programa largo vs. composición



Fuente: Elaboración propia (2023)

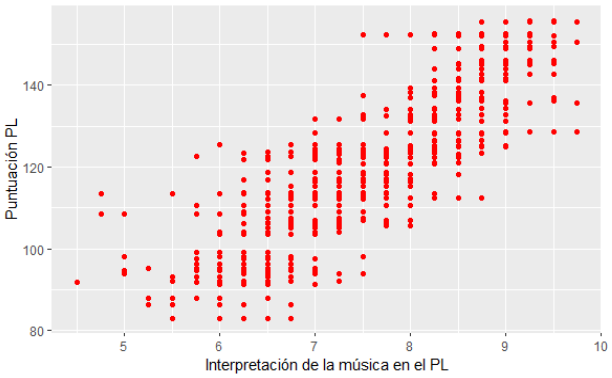
Interpretación de la música:

Figura 33: Diagrama de dispersión- Puntuación programa corto vs. Interpretación de la música



Fuente: Elaboración propia (2023)

Figura 34: Diagrama de dispersión- Puntuación programa largo vs. Interpretación de la música



Fuente: Elaboración propia (2023)

A partir de estas graficas se puede apreciar una clara relación lineal positiva entre la puntuación de una patinadora y sus resultados en la parte subjetiva del programa. Es decir, por mucho que la parte técnica sea importante y las patinadoras deban centrarse en ejecutar elementos difíciles, la perfección y limpieza tanto a la hora de ejecutar estos elementos como los pasos existentes entre ellos son iguales o incluso más decisivos.

En definitiva, independientemente de si se cumplen las hipótesis formuladas en un principio o no, se puede observar la relevancia de diferentes variables seleccionadas en el capítulo 3. A continuación, se sigue el estudio tanto del efecto de estas variables como sus resultados en los diferentes modelos creados.

6.2 Regresión lineal múltiple

A la hora de crear el primer modelo se opta por el uso de la regresión lineal múltiple, obteniéndose el modelo mostrado en el Anexo I. Como se puede observar en este, la primera estrategia a la hora de crear este modelo es el uso de todas las variables para la posterior eliminación de las menos relevantes.

Utilizando todas las variables en su diseño, se obtiene un modelo cuyo coeficiente de determinación[225] es 0,8551 para el programa corto y 0,857 para el largo. Es decir, el 85,51% y 85,7 de las varianzas de las muestras puedes ser explicados. Teniendo en cuenta que cuanto más se acerque a la unidad esta medida más preciso se considera el modelo, estos valores muestran una gran eficacia por parte de los modelos diseñados.

A pesar de considerarse valores aceptables, es necesario estudiar otros factores involucrados en el análisis de resultados como la significancia estadística de las variables independientes. Esta significancia estadística se mide a partir del p-valor, la cual concluye que una variable tiene un efecto significativo si esta tiene un p-valor asociado menor a 0,05[226]. En este caso, se han obtenido los siguientes resultados:

Tabla 3: Significancia estadística en modelo 1 por regresión lineal múltiple para el programa corto

Variable	**p-valor**
Edad	0.16622
Mejor marca personal	< 2e-16
Edad cuando se logró la mejor marca personal	0.01011
Altura	9.75e-05
Orden	0.92650
Misma nacionalidad que el juez	0.77592
Competición en su país	7.65e-05
Habilidades de la patinadora	0.66956

[225] Su explicación se incluye en los capítulos 4 y 5.
[226] Suárez, N. R. (2012).

Actuación de la patinadora	< 2e-16
Composición del programa	0.00935
Transiciones de la patinadora	0.61339
Interpretación de la música	0.97651

Fuente: Elaboración propia

Tabla 4: Significancia estadística en modelo 1 por regresión lineal múltiple para el programa largo

Variable	**p-valor**
Edad	0.08445
Mejor marca personal	< 2e-16
Edad cuando se logró la mejor marca personal	0.10424
Altura	0.03843
Orden	0.40415
Misma nacionalidad que el juez	0.99270
Competición en su país	0.00228
Habilidades de la patinadora	0.63749
Actuación de la patinadora	< 2e-16
Composición del programa	0.65330
Transiciones de la patinadora	1.11e-05
Interpretación de la música	0.37820

Fuente: Elaboración propria

Observando el p-valor de los distintos inputs y aplicando la teoría se han alterado las variables explicativas hasta obtener un modelo en el que se tienen en cuenta las solamente aquellas variables cuyo p-valor no supera el 0,05:

Tabla 5: Significancia estadística en modelo 2 por regresión lineal múltiple para el programa corto

Variable	**p-valor**
Mejor marca personal	< 2e-16
Altura de la patinadora	0.000290
Competición en su país	0.000245
Actuación de la patinadora	< 2e-16
Composición del programa	0.000419
Edad	< 2e-16

Fuente: Elaboración propia

Tabla 6: Significancia estadística en modelo 2 por regresión lineal múltiple para el programa largo

Variable	**p-valor**
Mejor marca personal	< 2e-16
Altura de la patinadora	0.1505

Competición en su país	0.0109
Actuación de la patinadora	< 2e-16
Transiciones de la patinadora	1.28e-10

Fuente: Elaboración propia

Estos nuevos modelos muestran un coeficiente de determinación de 0,8447 y 0,8421 respectivamente. Estas nuevas medidas muestran que el hecho de eliminar las variables con mayores p-valor no mejora el modelo. De hecho, esto parece incluso empeorarlo ligeramente.

Esto último apunta a una ligera ventaja del primer modelo. Teniendo esto en cuenta, junto con las relaciones que se observan a partir de los diagramas de dispersión en los que todas las variables muestran tener cierta influencia en la puntuación de la patinadora, el modelo 1 parece ser la mejor opción.

Por otro lado, otro proceso que pareció interesante a la hora de crear el modelo fue el stepwise[227]. Utilizando este método se ha creado el modelo incluido en el Anexo I, obteniéndose los siguientes resultados:

Tabla 7: Resumen de las medidas del modelo 3 por regresión lineal múltiple

Medida	**Programa corto**	**Programa largo**
R^2	0.835	0.850

Fuente: Elaboración propia

Una vez más, en general se observan mejores resultados con la primera opción. Teniendo en cuenta una vez más esto, junto a las conclusiones obtenidas en los diagramas de dispersión al comienzo de este capítulo, el primer modelo parece seguir siendo la mejor opción.

En definitiva, basándome en los resultados de la tabla 8 y en las observaciones explicadas durante este capítulo puedo afirmar que el modelo 1 se ajusta bien a los datos, tiene buen poder predictivo y alta significación estadística.

Tabla 8: Resumen de las medidas del modelo 1 por regresión lineal múltiple

Medida	**Programa corto**	**Programa largo**
R^2 entrenamiento	0,8551	0,857
RMSE entrenamiento	0.1507803	0.1371814

Fuente: Elaboración propia

Una vez decidido el modelo a utilizar, es necesario definir la ecuación de la recta de regresión. Obteniéndose las siguientes expresiones a partir de los parámetros estimados mediante el método de mínimos cuadrados[228]:

[227] Explicado en el capítulo 4.
[228] Método explicado en el capítulo 4.

*Puntuación programa corto =0.011 - 0.085*edad cuando obtuvo la mejor marca personal + 0.419*mejor marca personal + 0.037*altura - 0.010*habilidades del PC +0.808*actuación del PC - 0.269*composición del PC+0.0407*si compite en su país + 0.025*orden del PC 0.053*edad - 0.013*si tiene la misma nacionalidad que algún juez -0.017*transiciones del PC + 0.015*interpretación de la música del PC*

Puntuación programa largo =0.019 - 0.063 edad cuando obtuvo la mejor marca personal + 0.398*mejor marca personal - 0.032*altura - 0.038* habilidades del PL +0.946* actuación del PC -0.215* composición del PC+0.061* si compite en su país + 0.022* orden del PL -0.048*Edad-0.009*si tiene la misma nacionalidad que algún juez - 0.214* transiciones del PL+ 0.030*interpretación de la música del PL*

Una vez se obtiene la expresión final del modelo, es posible aplicarlo a los datos de prueba. De esta manera se comprueba la capacidad de adaptación y precisión de este a distintos conjuntos de datos. Haciendo esto se obtienen los siguientes resultados:

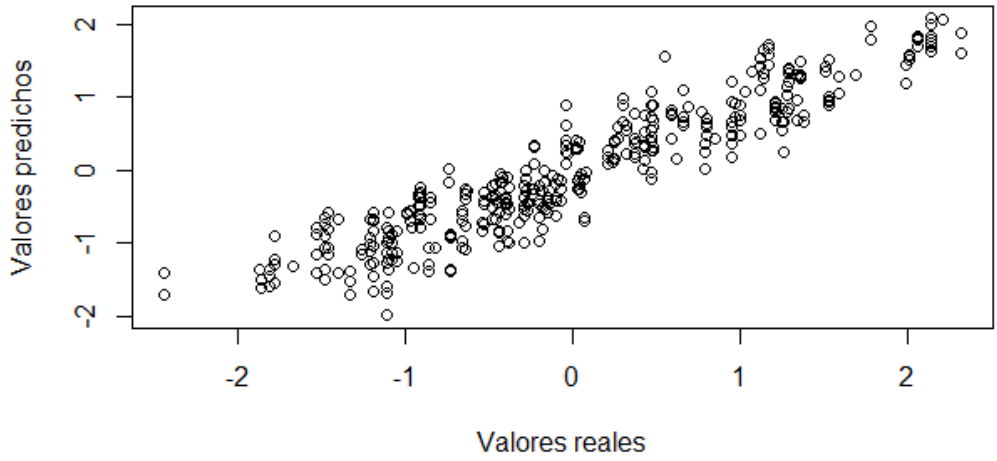
Tabla 9: Resumen de resultados obtenidos a partir de los datos de prueba

Medida	Programa corto	Programa largo
R^2 prueba	0.8431	0.8488
RMSE prueba	0.1516	0.1387

Fuente: Elaboración propia

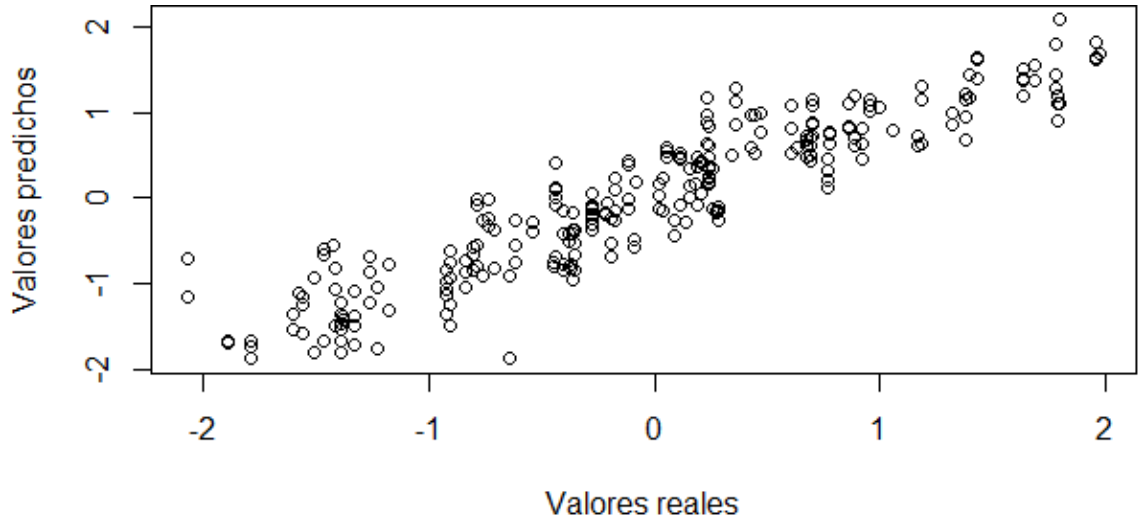
Con unos coeficientes de determinación tan cercanos a la unidad y unos errores cuadráticos medios bajos, se puede afirmar que el modelo no solo es capaz de predecir con bastante precisión la variable explicada, sino que también presenta una gran capacidad de generalización ante nuevos datos.

Figura 35: Valores reales vs. predichos en el PC (Prueba)



Fuente: Elaboración propia (2023)

Figura 36: Valores reales vs. predichos en el PL (Prueba)



Fuente: Elaboración propia (2023)

Por otro lado, a partir de las figuras 35 y 36 se puede comprobar de manera visual la capacidad del modelo para capturar la relación entre las variables elegidas y predecir los valores deseados. Cabe destacar, la cercanía de los puntos a una recta diagonal en ambos casos, mostrando así la coherencia del modelo y su buen ajuste. Analizando estos resultados se puede comprobar el buen rendimiento de los modelos creados.

6.2.1 Efecto de competir en su país

Una vez diseñados los modelos y analizado sus resultados es posible comprobar otra de las hipótesis formuladas inicialmente. En un principio se considera que: aquellas patinadoras que están compitiendo en su país logran mayores puntuaciones.

Basándonos en el coeficiente que acompaña a esta variable en el modelo diseñado y su p-valor podemos comprobar dicha hpótesis:

Tabla 10: coeficiente y p-valor de la variable: Competir en su país

Medida	Programa corto	Programa largo
Coeficiente	0.0407	0.061
p-valor	7.65e-05	0.00228

Fuente: Elaboración propia

Los bajos resultados obtenidos para el p-valor en ambos programas indican que el hecho de competir en su país tiene un efecto significativo sobre la variable dependiente. Por otro lado, a partir de los coeficientes se puede observar que el efecto es el que se sospechó en un principio. Es decir, existe una relación positiva entre ambas variables.

6.2.2 Efecto de compartir nacionalidad con algún juez

Los resultados obtenidos a la hora de diseñar el modelo nos permiten comprobar también la siguiente hipótesis: Se espera que aquellas patinadoras que comparten nacionalidad con al menos uno de los jueces obtienen mejores resultados.

Tabla 11: coeficiente y p-valor de la variable: Compartir nacionalidad con algún juez

Medida	Programa corto	Programa largo
Coeficiente	- 0.013	-0.009
p-valor	0.77592	0.992

Fuente: Elaboración propia

En este caso se ha obtenido un p-valor muy por encima de 0,05, el cual se considera el necesario para considerar la variable significativa. Por lo que no se puede asegurar que el hecho de que una patinadora tenga la misma nacionalidad que alguno de los jueces afecte a sus resultados.

Sin embargo, a la hora de diseñar el modelo se comprueba que al eliminar esta variable se obtienen peores resultados, por lo que dicha variable sí que influye en la dependiente. Aun así, esto no nos permite confirmar la hipótesis inicial ya que los coeficientes que la acompañan son negativos, lo que significa que el hecho de que compartan nacionalidad disminuye su puntuación final.

En resumen, tras la creación de diversos modelos utilizando diferentes métodos, se selecciona aquel que parece adaptarse mejor a los datos de los que se dispone y que mayor poder predictivo presenta. Su gran rendimiento se demuestra posteriormente comprobando su funcionamiento ante nuevos datos, aportando una vez más resultados satisfactorios.

Una vez se selecciona el modelo definitivo, es posible definir la ecuación de la recta de regresión, la cual representa la relación entre la variable dependiente y las explicativas. Para ello es necesario estimar el valor de los coeficientes relacionados con las distintas variables a utilizar, los cuales nos indican que influencia tiene cada una de ellas sobre la explicada.

Dado que el objetivo de este trabajo es la comparación de distintos métodos y la selección de un modelo óptimo, es necesario utilizar una segunda herramienta para así comparar sus resultados. Por ello, a continuación, se explican y analizan los resultados obtenidos a partir de las máquinas de vectores de soporte.

6.3 Máquinas de vectores de soporte

Para el siguiente ejemplo se emplean las máquinas de vectores de soporte[229]. Mediante este método se pretende crear, igual que en el caso de la regresión lineal múltiple, dos

[229] Técnica explicada en el capítulo 5.

modelos capaces de adaptarse a los datos tanto del programa corto como el programa largo. Esto ha llevado a la creación de los dos modelos incluidos en el Anexo I.

Por un lado, se tiene un modelo orientado a la predicción de puntuaciones del programa corto, el cual se recuerda que es la primera de las dos pruebas en una competición de patinaje. El diseño de este se lleva a cabo utilizando los parámetros[230]: C=256, γ= 0.1 y ε=0, y cuenta con 906 vectores de soporte[231].

En segundo lugar, se tiene un modelo orientado a la predicción de puntuaciones del programa largo, segunda parte de estos eventos deportivos. El diseño de este se lleva a cabo utilizando los parámetros: C=512, γ= 0.1 y ε=0, y cuenta con 595 vectores de soporte.

Como se ha explicado en capítulos anteriores, ambos han sido diseñados y ajustados hasta encontrar el diseño óptimo. Es decir, hasta encontrar un modelo capaz de adaptarse a diferentes bases de datos y predecir las puntuaciones de las patinadoras de forma precisa. Para poder afirmar esto es necesario llevar a cabo una serie de medidas que muestren su calidad. En este caso se utilizan los siguientes métodos:

En primer lugar, se calcula el coeficiente de determinación[232], cuya ecuación se recuerda que es:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad [31]$$

A partir de esta expresión se obtienen los siguientes resultados:

Tabla 12: Resultados obtenidos a partir del coeficiente de determinación

	Programa corto	Programa largo
R^2 entrenamiento	0.999	0.9999
R^2 prueba	0.979	0.9648

Fuente: Elaboración propia

Teniendo en cuenta que cuanto más cercano sea su coeficiente de determinación a la unidad mejor explica el modelo la variabilidad de los datos, los resultados obtenidos a partir de los datos de entrenamiento muestran un excelente rendimiento por parte de ambos modelos.

Por otro lado, tanto el modelo diseñado para el programa corto, como el del programa largo resultan en un coeficiente de determinación cercano a la unidad también para el

[230] La explicación del valor de estos parámetros se encuentra en el capítulo 5.
[231] El número de puntos de datos de entrenamiento que influyen en la estructura de la superficie de decisión.
[232] Explicación incluida en capítulos anteriores.

conjunto de datos de prueba. Esto muestra la gran capacidad de adaptación que ofrecen ambos modelos antes nuevos conjuntos de datos, siendo capaces de predecir de manera precisa y ajustada a los datos de los que disponen.

La siguiente medida de la que se dispone es la raíz del error cuadrático medio[233], cuya ecuación es la siguiente:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \qquad [32]$$

A partir de esta expresión se obtienen los siguientes resultados:

Tabla 13: Resultados obtenidos a partir de la raíz del error cuadrático medio

	Programa corto	Programa largo
RMSE entrenamiento	0.0009	7.319e-05
RMSE prueba	0.0210	0.0318

Fuente: Elaboración propia

Dado que esta medida se utiliza para estimar la diferencia entre las predicciones del modelo y los valores observados en un conjunto de datos[234], los resultados obtenidos a partir de los datos de entrenamiento muestran un gran rendimiento por parte de ambos modelos en esta fase. Es decir, los modelos son capaces de predecir con gran exactitud las puntuaciones de las patinadoras en el conjunto de entrenamiento.

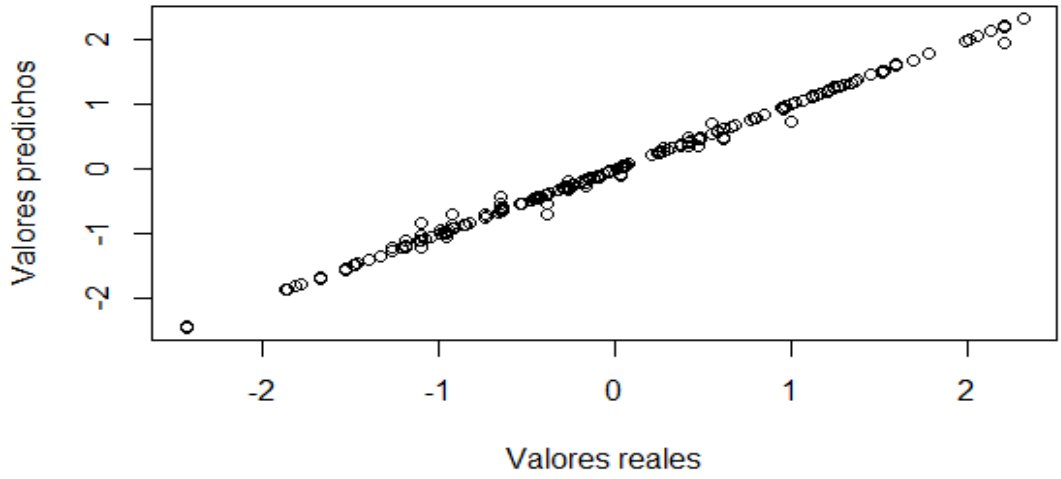
Por otro lado, se obtiene un RMSE mayor en el caso de los datos de prueba, lo que indica un peor rendimiento por parte del modelo. Aun así, estos valores siguen siendo bastante bajos, por lo que se considera que ambos modelos son capaces de predecir con gran exactitud las puntuaciones de las patinadoras en los dos programas con conjuntos de datos nuevos.

Cabe destacar, pese a que las predicciones del programa corto en el entrenamiento muestran una diferencia bastante mayor con los valores reales, en el caso de los datos de prueba parece tener un rendimiento mejor. Es decir, pese mostrar menos precisión en el proceso de entrenamiento, parece tener una mayor capacidad de ajuste de nuevos conjuntos de datos.

[233] Explicada en capítulos anteriores.
[234] Chai, T., et. Al. (2014).

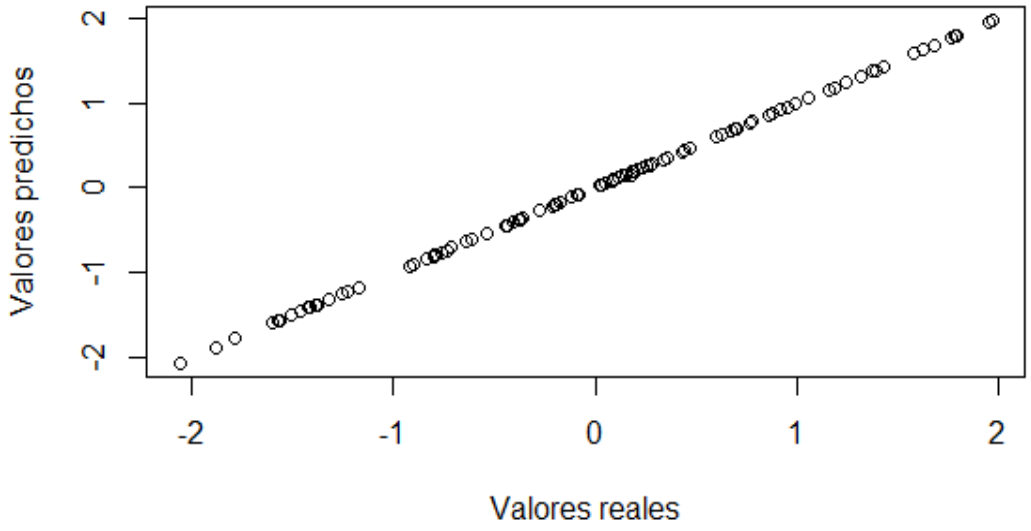
Por último, los diagramas de dispersión son una manera de ver de forma gráfica esta relación entre las puntuaciones predichas y las observadas:

Figura 37: Valores reales vs. predichos en el PC (Entrenamiento)



Fuente: Elaboración propia (2023)

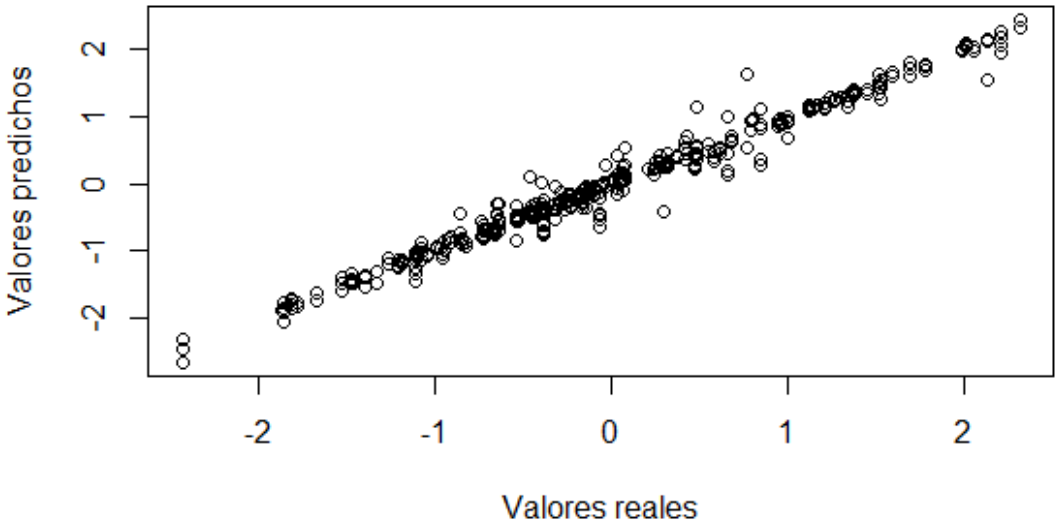
Figura 38: Valores reales vs. predichos en el PL (Entrenamiento)



Fuente: Elaboración propia (2023)

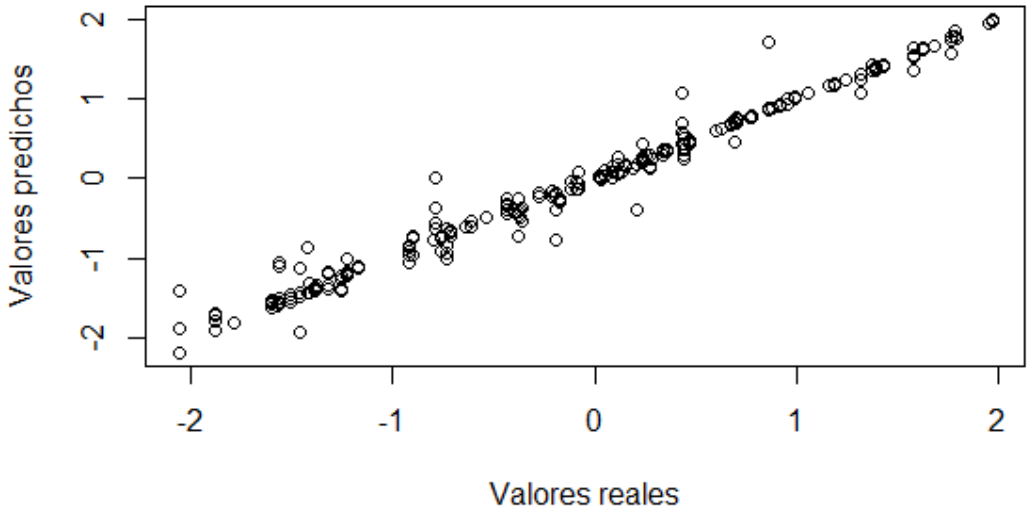
A partir de estas dos gráficas se puede confirmar lo que ya se había observado previamente. El proceso de entrenamiento para ambos modelos presenta un rendimiento muy elevado, con un gran ajuste a los datos observados, destacando especialmente el programa largo con una recta de regresión prácticamente perfecta.

Figura 39: Valores reales vs. predichos en el PC (Prueba)



Fuente: Elaboración propia (2023)

Figura 40: Valores reales vs. predichos en el PL (Prueba)



Fuente: Elaboración propia (2023)

Una vez más, las gráficas confirman lo observado gracias a las medidas previas. Muestran un modelo con un gran rendimiento ante nuevos conjuntos de datos, pudiéndose observar una alta relación entre las puntuaciones predichas y las reales en estos casos.

En resumen, el modelo creado por las máquinas de vectores de soporte muestra una precisión y capacidad de adaptación muy elevados. Obteniendo grandes resultados tanto con los datos de entrenamiento como los de prueba.

Gracias a los resultados obtenidos a partir de ambos métodos es posible compararlos y así elegir el modelo óptimo para esta tarea. Esto se lleva a cabo en la siguiente parte del capítulo.

6.4 Comparación de ambos modelos

A continuación, se comparan los resultados obtenidos a partir de los distintos métodos empleados a lo largo del trabajo para así poder seleccionar el modelo que mejor rendimiento presente a la hora de predecir las competiciones de patinaje. Para ello se utilizan las medidas explicadas previamente:

Tabla 14: Resumen de medidas de precisión para ambos métodos

	Regresión lineal múltiple		Máquinas de vectores de soporte	
	PC	PL	PC	PL
R^2 entrenamiento	0,8551	0,857	0.999	0.9999
RMSE entrenamiento	0.1483	0.1372	0.0009	7.319e-05
R^2 prueba	0.8431	0.8488	0.979	0.9648
RMSE prueba	0.1516	0.1387	0.0210	0.0318

Fuente: Elaboración propia

En la tabla se puede observar que en ambos casos se obtienen mejores resultados con los datos de entrenamiento. Esto es lógico ya que este es el conjunto a partir del cual se crea el modelo.

Aun así, ambos métodos presentan una gran capacidad de predicción tanto para los datos de entrenamiento como conjuntos de datos nuevos. Esto implica que eligiendo cualquiera de las dos opciones se obtiene un modelo que proporciona resultados precisos y coherentes.

Sin embargo, observando las medidas proporcionadas se pueden distinguir mejores resultados por parte de las máquinas de vectores de soporte, mostrando un modelo con una alta precisión tanto con los datos de entrenamiento como con los de prueba. Por lo que podemos afirmar que, aunque la regresión lineal múltiple se considera una buena técnica a la hora de predecir competiciones de patinaje artístico, el modelo óptimo es el creado por las máquinas de vectores de soporte.

CAPITULO 7: EJEMPLO

Una vez diseñado el modelo con los datos de prueba se ha puesto a prueba su capacidad de adaptación con los denominados datos de prueba, ambos procedentes de una misma base de datos y divididos de manera aleatoria[235]. Este método ha resultado útil tanto para el diseño del modelo como para su estudio, pero se ha considerado que puede ser poco claro para los lectores a la hora de evaluar los beneficios de usar este modelo.

Este capítulo se centra en la aplicación práctica del modelo óptimo[236] elegido en el capítulo anterior a un ejemplo determinado, pudiéndose así relacionar las predicciones realizadas con una patinadora en concreto. Esto nos permite comparar la puntuación obtenida según el modelo con la real de manera sencilla.

Para llevar a cabo este paso se puede elegir cualquier base de datos que recoja las variables incluidas en el modelo, ya que se ha demostrado previamente su alta capacidad de adaptación a nuevos conjuntos de información. En este proyecto se ha decidido utilizar como ejemplo a las diez mejores patinadoras de los Juegos Olímpicos de Pekín 2022, ya que se considera una competición lo suficientemente reciente e importante como para ser representativa[237].

Tabla 15: Comparación entre puntuaciones predichas y reales

PATINADORA	PUNTUACIÓN PC	PREDICCIÓN PC	PUNTUACIÓN PL	PREDICCIÓN PL
Anna Scherbakova	80,20	68,34	175,50	157,14
Alexandra Trusova	74,60	57,16	177,13	154,37
Kaori Sakamoto	79,84	65,54	153,29	135,69
Kamila Valieva	82,16	75.31	141,93	126,95
Wakaba Higuchi	73,51	46,72	140,93	141,47
Young You	70,34	59,31	142,75	121,12
Alysa Liu	69,50	65,49	139,45	111,84
Loena Hendrickx	70,09	53,32	136,7	112,53
Yelim Kim	67,78	45,94	134,85	96,12
Mariah Bell	65,38	40,52	136,92	100,34

Fuente: Elaboración propia

[235] Se ha dividido la base de datos original en un 70% datos de entrenamiento y 30% de prueba.
[236] El modelo creado por las máquinas de vectores de soporte.
[237] Los datos necesarios para crear la base de datos utilizada en el ejemplo se obtienen una vez más a través de la página web oficial de la ISU.

Pese a no obtenerse los resultados esperados, predicciones precisas de las puntuaciones de las patinadoras, a partir de esta tabla se puede obtener otro tipo de información. Resumida en las siguientes tablas:

Tabla 16: Comparación de puestos predichos y reales en el programa corto

Patinadora	Puesto real	Puesto predicho	Puesto final
Anna Scherbakova	2	2	1
Alexandra Trusova	3	6	2
Kaori Sakamoto	4	3	3
Kamila Valieva	1	1	4
Wakaba Higuchi	5	8	5
Young You	6	5	6
Alysa Liu	8	4	7
Loena Hendrickx	7	7	8
Yelim Kim	9	9	9
Mariah Bell	10	10	10

Fuente: Elaboración propia

Tabla 17: Comparación de puestos predichos y reales en el programa largo

Patinadora	Puesto real	Puesto predicho	Puesto final
Anna Scherbakova	2	1	1
Alexandra Trusova	1	2	2
Kaori Sakamoto	3	4	3
Kamila Valieva	5	5	4
Wakaba Higuchi	6	3	5
Young You	4	6	6
Alysa Liu	7	8	7
Loena Hendrickx	9	7	8

Yelim Kim	10	10	9
Mariah Bell	8	9	10

Fuente: Elaboración propia

Analizando ambas tablas se puede entender la importancia de los elementos artísticos a la hora de competir a nivel profesional. En ellas se compara el puesto que obtendría una patinadora si solo se tuvieran en cuenta las variables que se han considerado explicativas en el diseño del modelo y el obtenido en la vida real.

Algunas patinadoras como Kamila Valieva o Yelim Kim obtienen el mismo puesto tanto en la competición real como en las predicciones de nuestro modelo. Teniendo en cuenta que en la vida existen un mayor número de factores afectan a la puntuación final, entre los que destaca la dificultad de los elementos técnicos, se puede concluir que estas patinadoras presentan un equilibrio tanto en la parte artística[238] como la parte técnica[239].

Por otro lado, existen casos en los que los resultados predichos y los reales no concuerdan. En este caso se puede llegar a la conclusión de que, al no tenerse en cuenta la parte técnica en el modelo, esta diferencia representa lo perjudicada o beneficiada que se ve una patinadora por las distintas partes de un programa.

Entre estos casos destaca Alexandra Trusova a quien se atribuyen puestos inferiores a los reales tanto en el programa corto como el largo. Esto nos indica la clara desventaja que sufre la deportista en la rama artística.

Esta desventaja parece ser compensada por lo que debe ser una gran superioridad técnica, ya que consigue mantenerse en los primeros puestos durante toda la competición. Sin embargo, esto no parece ser suficiente para ganar la competición debido la importancia de los elementos más subjetivos, confirmando lo dicho previamente por profesionales[240].

La importancia de estos elementos se puede comprobar una vez más en esta tabla a partir de Anna Scherbakova quien, aun no siendo capaz de alcanzar el primer puesto en ninguna de las dos jornadas, se proclama ganadora de la competición de todas formas. Tras analizar el funcionamiento del modelo y sus resultados se ha llegado a la conclusión de que la consistencia que muestra la joven patinadora entre la puntuación real y la

[238] Parte que se ha tenido en cuenta a la hora de diseñar el modelo.
[239] Parte que no se ha tenido en cuenta a la hora de diseñar el modelo y que puede tanto beneficiar como perjudicar notablemente a una patinadora.
[240] Profesionales como Pedro Lamelas, director de Hielo Español.

dependiente de los elementos subjetivos, incluso cierta superioridad de la segunda en algunas ocasiones es decisiva[241].

Por último, a partir de las diferentes tablas mostradas a lo largo de este capítulo se puede observar que a la hora de predecir las puntuaciones de las participantes en las competiciones de patinaje artístico sobre hielo de forma precisa es necesario tener en cuenta los elementos técnicos. Pese a haber comprobado la importancia de las cualidades artísticas, el modelo parece no ser capaz de predecir con precisión sin tener en cuenta los distintos elementos ejecutados y sus niveles. Esta mejora es imprescindible si se quiere dar uso al modelo en campos como las apuestas deportivas.

En conclusión, a partir de este ejemplo hemos sido capaces de entender la influencia que puede llegar a tener la presentación artística de un programa cuando nos encontramos en competiciones de alto nivel. Hemos comprobado que en puede llegar a ser decisiva, como el caso de Anna y Alexandra.

Sin embargo, este ejemplo ha servido para demostrar que, a la hora de predecir puntuaciones en vez de puestos, lo cual se recuerda que es el objetivo de este trabajo, el modelo no es preciso. Esto supone un problema a la hora de aplicar el modelo en campos como las apuestas deportivas. Este se considera, una vez mejorado el modelo, un campo que tiene mucho potencial, por lo que el siguiente capítulo se dedica a explicar su posible aplicación.

[241] Corroborando una vez más las conclusiones de expertos como Pedro Lamelas.

CAPÍTULO 8: MEMORIA ECONÓMICA Y POTENCIAL DEL MODELO

Como se ha mencionado previamente se considera que el modelo creado en este trabajo tiene mucho potencial económico en un campo como las apuestas deportivas. Este capítulo se centra en el estudio y comprensión de las ganancias económicas tanto de este mundo como de la posible aplicación del modelo.

8.1 Apuestas deportivas

El surgimiento de las apuestas deportivas se remonta a la antigua Grecia, donde los ciudadanos apostaban por el deportista favorito de los dioses[242]. Durante todos estos años ha continuado creciendo, alcanzando cada vez mayor popularidad[243]. En las últimas décadas, novedades como la llegada de internet y la liberación de las regulaciones gubernamentales han permitido la evolución de este mundo hasta alcanzar una notable importancia en la actualidad[244].

Hoy en día las apuestas deportivas constituyen un negocio billonario, alcanzando en 2020 ventas de más de $8,3 billones[245]. Estos números aumentan cada año, ya que cada vez existen más plataformas que ofrezcan estos servicios y más oferta deportiva[246].

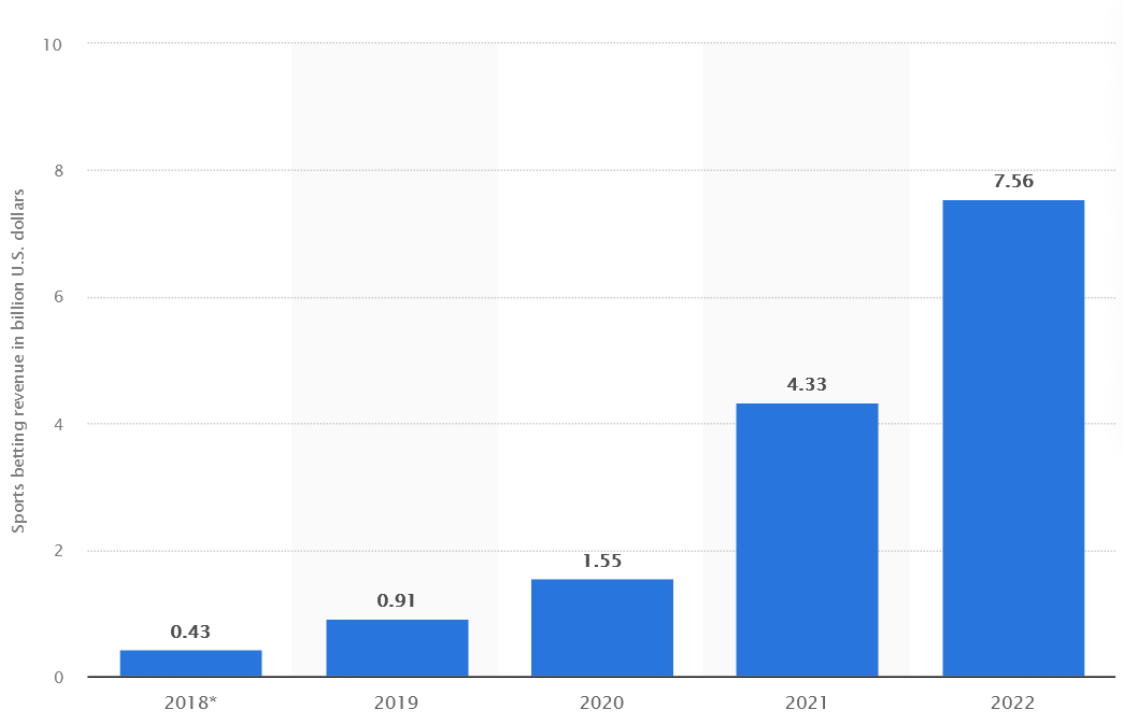
Estas grandes cantidades de dinero invertidas no solo benefician a las casas de apuestas, sino que también pueden influir en la economía de un país[247]. Llegando a beneficiarlo significativamente, pues el mundo de las apuestas deportivas supone una fuente de financiación alternativa a los gobiernos[248].

Hablando de países en el mundo de las apuestas deportivas, cabe destacar que no en todos se mueve la misma cantidad de dinero. Entre los países que más dinero invierten en esta actividad se encuentran algunos como China, Finlandia o Canadá [249]. Destacando este último donde se estima que, una vez se finalice el proceso de legalizar las apuestas deportivas en todas las jurisdicciones, se puede llegar a generar hasta 25 mil millones de dólares a partir de este negocio[250].

Por otro lado, tras la caída de la ley PASPA en Estados Unidos en 2018 el mundo de las apuestas deportivas ha crecido significativamente. Llegando a colocarse también entre los países líderes en este negocio[251].

[242] Redacción. (2022).
[243] Bohórquez, K. S. (2021).
[244] Concepción, M. (2022).
[245] Bohórquez, K. S. (2021).
[246] Ibid.
[247] Carcedo, L. P. (2010).
[248] Ibid.
[249] Rodríguez, A. (2023).
[250] PlayCanada.com. (2021).
[251] Ibid.

Figura 41: Ganancias por apuestas deportivas en E.E.U.U. (2018-2022)



Fuente: Statista. (2023)

Estos países no solo destacan por su inversión en el mundo de las apuestas, sino que también se encuentran entre los países donde mayor afición por el patinaje artístico hay. Esta pasión por el deporte, sumada a la creciente popularidad de las apuestas deportivas nos hace pensar que puede ser un sector muy lucrativo.

8.2 Apuestas en el mundo del patinaje

El patinaje artístico sobre hielo cuenta con mucha popularidad tanto en los países mencionados previamente como en China, Japón o muchos otros países europeos, donde la inversión en apuestas es estable y elevada[252]. Estos países suelen optar por apostar en determinadas competiciones como los mundiales, el campeonado de Europa o el Gran Prix, todas ellas internacionales y de gran importancia.

Sin embargo, cada cuatro años el número de interesados en este deporte incrementa considerablemente gracias a los Juegos Olímpicos de invierno. Estos producen un aumento significativo en el número de apuestas relacionadas con el patinaje sobre hielo, aumentando aún más la cantidad de dinero generado[253].

A la hora de apostar en este deporte existen diferentes posibilidades, algunas más populares que otras. Lo más común es centrarse en el patinador o pareja que se espera que vaya a ganar la competición. Esta categoría suele quedar reducida a un pequeño

[252] Kateryna. (2022).
[253] Ibid.

grupo de participantes, ya que se trata de un deporte que exige un alto nivel de precisión y técnica, siendo prácticamente imposible que alguien gane por puro azar[254].

Esto nos lleva a la siguiente opción, apostar por que un patinador quede en un determinado "top". Las casas de apuestas ofrecen distintos rangos como puede ser los tres primeros, cinco primeros o incluso diez primeros, y los jugadores deben predecir si un determinado patinador, pareja o equipo conseguirá pertenecer a ese grupo[255].

Por último, está la opción de apostar centrándose en el rendimiento de los patinadores más que en su posición final[256]. Es decir, consiste en predecir si la puntuación final de los deportistas quedará por encima o por debajo de una marca preestablecida[257]. Esta manera de apostar puede resultar interesante si se ha llevado a cabo un estudio del historial y las condiciones físicas de los participantes, pero no suele ser tan común como las dos anteriores[258].

Analizando estas tres opciones se ha llegado a la conclusión de que el modelo desarrollado en este trabajo puede tener mucha utilidad. Este puede ofrecer a sus usuarios una clara ventaja sobre los demás jugadores, permitiéndoles predecir las puntuaciones de las patinadoras, sabiendo a partir de estas los puestos que pueden llegar a alcanzar. Convirtiéndolo en un proyecto con mucho potencial económico.

Las casas de apuestas no suelen publicar las cantidades exactas de dinero invertidas en cada evento, por lo que no ha sido posible obtener números concretos de la cantidad de dinero invertida en el patinaje artístico. Sin embargo, teniendo en cuenta la importancia de las apuestas en los países explicados previamente y la afición por el patinaje artístico en muchos de ellos, hemos llegado a la conclusión de que con un correcto mecanismo de predicción es posible que se tarte de un campo muy lucrativo en el que utilizar nuestro modelo.

Sin embargo, no parece que el modelo trabaje correctamente en todas las categorías mencionadas. A pesar de tratarse de un proyecto cuyo objetivo es la predicción de las puntuaciones de las patinadoras, pudiéndose así cubrir las tres áreas explicadas, se ha podido comprobar en el capítulo anterior la necesidad de algunas mejorías para poder considerar el modelo totalmente fiable.

Una vez mejorado, este modelo puede ser aplicado a diversos campos que incluyan tanto una parte técnica, como una más subjetiva. A continuación, se incluyen otras áreas en las que el modelo diseñado podría ser de gran utilidad.

[254] Legalbet. (s. f.).
[255] Ibid.
[256] Ibid.
[257] Ibid.
[258] Ibid.

8.3 Otros campos de interés

Durante la creación de todo este proyecto se ha podido observar como las predicciones de resultados son muy importantes en las competiciones de patinaje artístico. Sin embargo, esta no es la única aplicación que se le encuentra a este trabajo.

Se considera que con una adaptación previa el modelo se puede aplicar a cualquier evento en los que se involucren elementos subjetivos evaluados por distintos jueces. Esto incluye otros deportes como puede ser el salto de trampolín, la gimnasia o incluso el boxeo.

Figura 42: Gimnasta Simone Biles en Tokio 2020

Fuente: Semana. (2022)

Yendo más allá, se plantea el uso de este modelo en ámbitos no deportivos. Sugiriéndose situaciones como juicios, donde distintos jurados deben tomar una decisión basándose tanto en elementos objetivos como subjetivos. Se trata de una aplicación más ambiciosa, pero posible.

En definitiva, a lo largo de este capítulo se ha podido comprobar la importancia que tienen las apuestas deportivas en distintos países, entre los que se encuentran algunos con una fuerte afición al patinaje. Esto nos hace pensar que las apuestas en el mundo del patinaje sobre hielo son muy lucrativas, por lo que una técnica predictiva como nuestro modelo puede ser de gran utilidad.

Sin embargo, antes de poder ser utilizado por los jugadores, el modelo necesita ciertas mejoras. Pudiéndose asegurar así resultados precisos y fiables. Estos futuros desarrollos se sugieren en el próximo capítulo.

CAPITULO 9: CONCLUSIONES Y DESARROLLOS FUTUROS

Este proyecto se ha centrado en el diseño, evaluación y comparación de dos modelos predictivos para las competiciones de patinaje artístico sobre hielo, tanto para el programa largo como para el programa corto. Para ello se ha recopilado la información facilitada por la Unión Internacional de Patinadores para ambos programas, obteniéndose una base de datos con 182 patinadoras para la primera parte y 95 para la segunda. Todos estos datos se han dividido de forma aleatoria con la intención de utilizar un 70% de ellos para entrenar el modelo y un 30% para evaluarlo.

Por un lado, se ha diseñado ha creado un modelo de regresión lineal múltiple mediante el método de mínimos cuadrados. Este ha demostrado muy buenos resultados tanto con los datos de entrenamiento como para los de prueba:

Tabla 18: Resumen de las medidas del modelo de regresión lineal múltiple

Medida	**Programa corto**	**Programa largo**
R^2 entrenamiento	0,8551	0,857
RMSE entrenamiento	0.1507803	0.1371814
R^2 prueba	0.8431	0.8488
RMSE entrenamiento	0.1507803	0.1371814
RMSE prueba	0.1516	0.1387

Fuente: Elaboración propia

Ante estos resultados hemos podido concluir que se trata de un modelo con un alto nivel de precisión y buena capacidad de adaptación.

Por otro lado, se han empleado las máquinas de vectores de soporte para la creación del segundo modelo. Una vez más, los resultados obtenidos han demostrado su alta calidad:

Tabla 19: Resumen de las medidas del modelo por máquinas de vectores de soporte

Medida	Programa corto	Programa largo
R^2 entrenamiento	0.999	0.9999
R^2 prueba	0.979	0.9648
RMSE entrenamiento	0.0009	7.319e-05
RMSE prueba	0.0210	0.0318

Fuente: Elaboración propia

Analizando ambas tablas y comparando sus resultados hemos podido seleccionar el modelo óptimo. Pese a que ambos modelos parecen ser capaces de predecir las puntuaciones de las patinadoras, las máquinas de vectores de soporte obtienen mejores resultados. Esta técnica no solo presenta un mayor de precisión, sino que parece tener mejor capacidad de adaptación a nuevas bases de datos.

Durante este proceso no solo se ha cumplido el objetivo de seleccionar un modelo óptimo. También hemos sido capaces de corroborar las hipótesis formuladas al comienzo del proyecto, pudiendo confirmar la relevancia de las variables seleccionadas. Ya sea de forma clara, como en el caso de la mejor marca personal, o no, como con la variable compartir nacionalidad con un juez. Esta variable nos ha parecido insignificante en un principio dado su alto p-valor, pero tras una serie de comprobaciones esta idea es descartada, encontrando así una relación entre variables menos obvia que las demás.

A continuación, se ha aplicado el modelo a un ejemplo concreto, pudiéndose así comparar los resultados obtenidos a partir de las predicciones con los obtenidos por las patinadoras en la vida real. Esto nos ha mostrado algunos errores por parte del modelo, ya que no parece posible predecir exactamente las puntuaciones de las patiandoras sin tener en cuenta los elementos técnicos. Aún asi, este paso nos ha ayudado a entender la influencia de la representación artística en los resultados finales de las competiciones, comprobándose que la existencia de desequilibrios entre la calidad técnica y la artística puede llegar a pasar factura como en el caso de Alexandra Trusova.

Visto esto, se han planteado una serie de campos en los que el modelo diseñado puede resultar de gran utilidad, como pueden ser las apuestas deportivas o los juicios que involucren jueces. Llamándonos especialmente la atención la primera, pues no solo parece ser un sector muy lucrativo, sino que se sospecha que el modelo puede ser aplicable a una amplia selección de deportes que involucren jueces encargados de evaluar tanto elementos técnicos como subjetivos.

Sin embargo, para ello el modelo necesita mejorar esos fallos comprobados durante su aplicación al ejemplo y desarrollarse más a fondo. Al tratarse de un tema tan amplio como es el deporte nos ha sido imposible llevar a cabo estos cambios, pero se dejan las puertas abiertas para una continuación.

9.1 Desarrollos futuros

Como ya se ha mencionado, el mundo del deporte y de los elementos que alteran sus resultados es un campo demasiado amplio como para ser cubierto en este trabajo. Por ello, se consideraría interesante la continuación de esta línea de trabajo para futuros proyectos.

Un tema que nos resulta interesante es la incorporación de los elementos técnicos al modelo. Durante la aplicación del modelo a un ejemplo se ha llegado a la conclusión que la falta de calidad de este se debe a la importancia de los elementos más objetivos en sus variables. Se cree que, mejorando esta parte, podríamos llegar a obtener un modelo que fuera capaz de predecir las puntuaciones de las patinadoras de manera mucho más precisa.

Por otro lado, el uso de una base de datos con más información también podría beneficiar a la calidad del modelo. Al comienzo del trabajo se han considerado variables

como el número de horas entrenadas a la semana como imprescindibles a la hora de crear el modelo, pero la falta de información ha impedido que las podamos usar. Por ello, se considera que completar la base de datos original con la información sobre estas variables consideradas en un principio podría ser un gran comienzo para desarrollos futuros. Se cree que estas variables tienen una gran influencia en los resultados finales y que podrían ayudar enormemente a una mejora del modelo actual.

Por último, otra mejoría que se considera pendiente en este trabajo son las variables relacionadas con si la patinadora compite en su país y si es de a misma nacionalidad que un juez. Este planteamiento es útil a la hora de clasificar a las patinadoras y comprobar su efecto sobre la puntuación final, pero puede llegar a ser una estrategia demasiado simple. Se cree que la influencia de los países va más allá de compartir nacionalidad o no, habiendo países que se apoyan entre ellos más que otros. Esto acuerdos se pueden deber a distintos motivos como pueden ser intereses comunes, historia o incluso política, pero sean por lo que sean pueden terminar afectando a los deportistas. Por ello, se cree que un mayor desarrollo de este sistema de clasificación beneficiaría notablemente al modelo.

CAPÍTULO 10: BIBLIOGRAFÍA

Abuín, J. R. (2007). Regresión lineal múltiple. IdEyGdM-Ld Estadística, Editor, 32.

Acevedo, Y., & Loaiza, G. (2023). Identificación de candidatos a primos Mersenne mediante clasificación ova-angular utilizando aprendizaje automático con regresión SVM y Kernel Gaussiano. Revista Politécnica, 19(37), 103-110.

aDmEUr. (2017, 21 diciembre). *EL MERCADO DE LAS APUESTAS DEPORTIVAS - europer.net*. europer.net

Admin. (2021). Regresión lineal y sus fundamentos. frankgalandev.

Afp, A. (2022, 7 febrero). Descalifican a cinco equipos de esquí mixto en Beijing 2022 por polémico detalle en sus trajes. Grupo Milenio.

Allaire, J. (2012). RStudio: integrated development environment for R. Boston, MA, 770(394), 165-171.

Arias, F. G. (2017). ECONOMÍA Y DEPORTE Analogía entre el sistema económico y el deporte de élite. ACTIVIDAD FÍSICA Y CIENCIAS/PHYSICAL ACTIVITY AND SCIENCE, 1(1).

Baños, R. V., Torrado-Fonseca, M., & Álvarez, M. R. (2019). Análisis de regresión lineal múltiple con SPSS: un ejemplo práctico. REIRE Revista d'Innovació i Recerca en Educació, 12(2), 1-10

Bohórquez, K. S. (2021, 23 junio). El millonario ecosistema de las apuestas deportivas. *Forbes Colombia*.

Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. Applied computing and informatics, 15(1), 27-33.

Cabrera, J. D. L., & Pereira-Toledo, A. (2018). Análisis del comportamiento del algoritmo SVM para diferentes kernel en ambientes controlados. Holos, 5, 101-115.

Cano, V. (2018, 17 febrero). Estas son las 10 mejores piruetas de Javier Fernández. Business Insider España

Carrasco Carrasco, M. (2016). Técnicas de regularización en regresión: implementación y aplicaciones.

Carcedo, L. P. (2010). El mercado de apuestas deportivas. Aranzadi. Fundación Codere. Thomson Reuters.

Carreño, F. M. (2018b, febrero 8). Patinaje artístico sobre hielo: reglas, historia y modalidades olímpicas. Marca.com.

Castilla, J. L. A. (2018b, febrero 19). Un 'viejo' con solo 26 años de edad: por qué Javi Fernández se jugaba la última carta. elconfidencial.com.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE). Geoscientific model development discussions, 7(1), 1525-1534.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE). Geoscientific model development discussions, 7(1), 1525-1534.

Chirivella González, V. (2015). Hipótesis en el modelo de regresión lineal por Mínimos Cuadrados Ordinarios.

Competition Strategy Analysis Based on Machine Learning Algorithms. Frontiers in Psychology, 13.

Concepción, M. (2022). Estadísticamente, ¿cuáles países apuestan más? ¿Por qué? | *Centroapuesta*

Dávila, J. (2022, 17 febrero). El drama de Trusova antes del podio: «Odio este deporte».

De Los Santos, P. R. (2022, 24 enero). Datos de entrenamiento vs datos de test - Think Big Empresas. Think Big.

Dupre, E., & Chung, G. (2022, 15 febrero). Russian Figure Skater Kamila Valieva Cleared to Compete in Olympics Amid Scandal. E! Online.

Efe. (2022, 18 octubre). La ISU eleva la edad minima para competir en patinaje tras el 'caso Valieva' MARCA.

El “jueceo” en una competencia. (2016, 9 septiembre). Patinaje Artístico sobre Hielo.

Figure skating | History & Competitions. (2023, 23 junio). Encyclopedia Britannica.

Figure Skating - International Skating Union. (s. f.).

García, J. A. (2020). Comparación de modelación por Inteligencia Artificial y Regresión Multivariable del comportamiento a flexión del UHPFRC. Dyna, 87(214), 258-267.

Gúzeva, A. (2022, 21 febrero). Cómo acabó el drama del patinaje artístico olímpico para Rusia. Russia Beyond ES.

Heras, J. M. (2019). Máquinas de Vectores de Soporte (SVM). IArtificial. net.

Horvat, T., & Job, J. (2020). The use of machine learning in sport outcome prediction: A review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(5), e1380.

Important Notice | 888.comTM. (s. f.).

ISU Judging System - International Skating Union. (2022).

ISU World Synchronized Skating Championships 2022. (s. f.).

Jaimes, K. (2022, 10 febrero). Escándalo en los Juegos de Invierno: Descalifican a cinco esquiadores por su vestimenta. Metro World News.

Kateryna. (2022, 2 febrero). How to Bet on Figure Skating in 2022. *GamingZion*.

Legalbet. (s. f.). *Figure skating betting: short-season analysis and sportsbook offers*. legalbet.com.

Lemons, M. Q. (2003). Predictive modeling of professional figure skating tournament data (Doctoral dissertation, University of Georgia).

Liu, M., Chen, Y., Guo, Z., Zhou, K., Zhou, L., Liu, H., ... & Zhou, J. (2022). Construction of Women's All-Around Speed Skating Event Performance Prediction Model and

Landajuela, I. (2019, 5 febrero). Diagramas de dispersión y modelo de regresión líneal en R.

Llamas, M. D. C. J. (2021). Modelización matemática para la predicción y prevención de lesiones deportivas. Retos: nuevas tendencias en educación física, deporte y recreación, (39), 681-685.

Martín, A., & Martín, A. (2022). La tortura oculta detrás del patinaje artístico. Hipertextual.

Martín Guareño, J. J. (2016). Support vector regression: propiedades y aplicaciones.

McCarvel, N. (2022, 18 febrero). Sui Wenjing y Han Cong, nuevo récord mundial y líderes del programa corto de patinaje por parejas. Olympics.com.

Mijwel, Maad M. "Artificial neural networks advantages and disadvantages." Retrieved from LinkedIn https//www. linkedin. com/pulse/artificial-neuralnet Work (2018).

Núcleo (Kernel) de las SVM – Numerentur.org. (s. f.)

Olivas, M. (2017, 28 agosto). Una campeona olímpica de patinaje se retira por problemas de anorexia. ELMUNDO

Pai, P. F., ChangLiao, L. H., & Lin, K. P. (2017). Analyzing basketball games by a support vector machines with decision tree model. Neural Computing and Applications, 28, 4159-4167.

Peláez, I. M. (2016). Modelos de regresión: lineal simple y regresión logística. Revista Seden, 14, 195-214.

Pértega Díaz, S., & Pita Fernández, S. (2001). Representación gráfica en el análisis de datos. Cad Aten Primaria, 8, 112-117.

PlayCanada.com. (2021, 26 agosto). Canada Could Grow into $2 Billion-a-Year Sports Betting Market, According to White Paper from PlayCanada.

Quintanilla, L. (2023, 5 junio). Preparación de los datos para la compilación de un modelo - ML.NET. Microsoft Learn.

Redacción. (2022c). La historia milenaria de las apuestas deportivas. *Red Historia*.

Redacción. (2022, 17 febrero). Beijing 2022: ¿Por qué Alexandra Trusova ganó plata en patinaje si hizo 5 saltos cuádruples? El Financiero.

Riedell. (2023, 7 junio). What Are the Different Types of Figure Skating? - Riedell Ice. Riedell Ice.

Rodrigo, J. A. (2017). Máquinas de vector soporte (support vector machines, svms). cienciadedatos. net, Abril.

Rodríguez, A. (2023). Top 10 de los países con los más grandes apostadores 🤑. *Blog Strendus - Apuestas deportivas y casino*

Rodríguez, D. (2021). Regresión de Vectores de Soporte (SVR, Support Vector Regression). Analytics Lane.

Rodríguez-Jaume, M. J., & Mora Catalá, R. (2001). Análisis de regresión múltiple.

Rvaquerizo, & Rvaquerizo. (2020, 13 octubre). El parámetro gamma, el coste, la complejidad de un SVM - Análisis y Decisión. Análisis y Decisión - Transformando datos en decisiones

Sanabria-Castro, A., Meneses-Guzmán, M., & Chiné-Polito, B. (2023). Uso de regresión de soporte vectorial para el control de espuma metálica. Revista Tecnología en Marcha, ág-42.

Sánchez, E. (2023, 21 marzo). Calendario de patinaje artístico sobre hielo 2022 | Cómo ver por tv y online streaming - Mundiales de Saitama. Eurosport Espana.

Sauces, I. A. (2020). Alina Zagitova: la dura transición de niña a mujer. Ravelo.

Scoring System | U.S. Figure Skating. (s. f.-b).

Semana. (2022, 14 febrero). ¿Robots como jueces de Gimnasia o Patinaje? No está lejos de ser una realidad. *Semana.com Últimas Noticias de Colombia y el Mundo*.

Staff, B. (2023, 7 febrero). Billboard. Billboard.

Suárez, E. J. C. (2014). Tutorial sobre máquinas de vectores soporte (sVM). Tutorial sobre Máquinas de Vectores Soporte (SVM), 1, 1-12.

Suárez, N. R. (2012). La revolución en la toma de decisiones estadísticas: el p-valor. Telos, 14(3), 439-446.

Sreenivasa, S. (2021, 16 diciembre). Radial Basis Function (RBF) Kernel: The Go-To Kernel. Medium

Statista. (2023, 15 junio). *Sports betting revenue in the U.S. 2018-2022*. The Associated Press. (2022, 18 febrero). “¡Odio este deporte!”: crisis en el patinaje ruso. Primera Hora.

Today, C. B. U. (2022, 5 febrero). How does scoring work in figure skating at the Olympics? Here’s how judges analyze the sport. USA TODAY.

TOP 10 países donde hacen apuestas deportivas. (2021, 11 marzo).

Trula, E. M. (2017). Surya Bonaly, la patinadora inalcanzable a la que acabaron prohibiendo «bailar con la muerte». Xataka.

Valera Guardiola, F. (2013). Sistema de predicción de resultados en eventos deportivos y su aplicación en las apuestas (Master's thesis).

Xie, J., Xu, J., Nie, C., & Nie, Q. (2016). Prediction on Performance of Age Group Swimming Using Machine Learning. In High Performance Computing and Applications: Third International Conference, HPCA 2015, Shanghai, China, July 26-30, 2015, Revised Selected Papers 3 (pp. 178-184). Springer International Publishing.

ANEXO I: CÓDIGO EMPLEADO PARA LA CREACIÓN DEL MODELO

Modelo definitivo: regresión lineal múltiple programa corto

```
# Importar los datos desde el archivo Excel

datosPC <- readxl::read_excel("VarPC.xlsx")

library(dplyr) # Cargar la librería de manipulación de dataframes "dply

NuevosPC<- select(datosPC,puntuación_PC, edad_MMP, mejor_marca_personal, altura, Skating_skills_PC, Performance_PC, Composition_PC, comp_encasa, ordenPC, Edad, Misma_Nac_PC, Transitions_PC, Interpretation_of_music_PC, tiempo)


# Renombrar las columnas del dataset

colnames(NuevosPC) <- c("puntuacion_PC", "edad_MMP", "mejor_marca_personal", "altura", "Skating_skills_PC", "Performance_PC", "Composition_PC", "comp_encasa", "ordenPC", "Edad", "Misma_Nac_PC", "Transitions_PC", "Interpretation_of_music_PC", "tiempo")


#Normalizar los datos

NormalesPC<- scale(NuevosPC)

#Scale devuelve una matriz-> convertirlo en un dataframe necesario para SVR

NormalesPC <- as.data.frame(NormalesPC)


#Renombrar columnas de la base de datos normalizada

colnames(NormalesPC) <- c("puntuacion_PC_N", "edad_MMP_N", "mejor_marca_personal_N", "altura_N", "Skating_skills_PC_N", "Performance_PC_N", "Composition_PC_N", "comp_encasa_N", "ordenPC_N", "Edad_N", "Misma_Nac_PC_N", "Transitions_PC_N", "Interpretation_of_music_PC_N", "tiempo_N")


#Dividir datos en entrenamiento y prueba

datos_conjuntos<- sample(c(TRUE, FALSE), nrow(NormalesPC), replace=T,prob=c(0.7,0.3))

datos_entrenamiento <- NormalesPC[datos_conjuntos,]

datos_prueba <- NormalesPC[!datos_conjuntos,]
```

```
#Uso de regresion lineal simple para modelar

model <- lm(puntuacion_PC_N~., data=datos_entrenamiento)

summary(model)


#Predicción de datos

prediccion_entrenamiento <- predict(model,newdata = datos_entrenamiento)

predicciones_prueba<- predict(model, newdata = datos_prueba)


#COMPROBACION MODELO

# Calcula el error cuadrático medio (MSE)

mse_entrenamiento <- mean(( prediccion_entrenamiento- datos_entrenamiento$puntuacion_PC_N )^2)

mse_prueba<- mean((predicciones_prueba- datos_prueba$puntuacion_PC_N)^2)

# Calcula el coeficiente de determinación (R²)

r2_entrenamiento<- 1 - sum((datos_entrenamiento$puntuacion_PC_N - prediccion_entrenamiento)^2) / sum((datos_entrenamiento$puntuacion_PC_N - mean(datos_entrenamiento$puntuacion_PC_N))^2)

r2_prueba <- 1 - sum((datos_prueba$puntuacion_PC_N - predicciones_prueba)^2) / sum((datos_prueba$puntuacion_PC_N - mean(datos_entrenamiento$puntuacion_PC_N))^2)

cat("Error cuadrático medio (MSE):", mse_entrenamiento, "\n")

cat("Error cuadrático medio (MSE):", mse_prueba, "\n")

cat("Coeficiente de determinación (R²):", r2_entrenamiento, "\n")

cat("Coeficiente de determinación (R²):", r2_prueba, "\n")

plot(datos_prueba$puntuacion_PC_N, predicciones_prueba, main = "Valores reales vs. Valores predichos (Prueba)", xlab = "Valores reales", ylab = "Valores predichos")


#Una vez decido el modelo, calculo coeficientes:

# Obtener los coeficientes del modelo

coeficientes <- coef(model)

# Obtener los nombres de las variables predictoras

variables <- names(coeficientes)[-1]  # Excluimos el coeficiente de intercepción

# Construir la .ión de regresión
```

ecuacion <- paste("puntuacion_PC_N =", coeficientes[1], "+", paste(coeficientes[-1], variables, sep = "", collapse = " + "))*

 # Imprimir la ecuación

print(ecuacion)

Modelo definitivo: regresión lineal múltiple programa largo

 # Importar los datos desde el archivo Excel

datosPL <- readxl::read_excel("VarPL.xlsx")

library(dplyr) # Cargar la librería de manipulación de dataframes "dply

NuevosPL<- select(datosPL,puntuación_PL, edad_MMP, mejor_marca_personal, altura, Skating_Skills_PL, Performance_PL, Composition_PL, comp_encasa, ordenPL, Edad, Misma_Nac_PL, Transitions_PL, Interpretation_of_the_Music_PL, tiempo)

 # Renombrar las columnas del dataset

colnames(NuevosPL) <- c("puntuacion_PL", "edad_MMP", "mejor_marca_personal", "altura", "Skating_Skills_PL", "Performance_PL", "Composition_PL", "comp_encasa", "ordenPL", "Edad", "Misma_Nac_PL", "Transitions_PL", "Interpretation_of_music_PL", "tiempo")

#Normalizar los datos

NormalesPL<- scale(NuevosPL)

#scale devuelve una matriz-> convertirlo en un dataframe necesario para SVR

NormalesPL <- as.data.frame(NormalesPL)

#Renombrar columnas de la base de datos normalizada

colnames(NormalesPL) <- c("puntuacion_PL_N", "edad_MMP_N", "mejor_marca_personal_N", "altura_N", "Skating_skills_PL_N", "Performance_PL_N", "Composition_PL_N", "comp_encasa_N", "ordenPL_N", "Edad_N", "Misma_Nac_PL_N", "Transitions_PL_N", "Interpretation_of_music_PL_N", "tiempo_N")

#Dividir datos en entrenamiento y prueba

datos_conjuntos<- sample(c(TRUE, FALSE), nrow(NormalesPL), replace=T,prob=c(0.7,0.3))

```
datos_entrenamiento <- NormalesPL[datos_conjuntos,]

datos_prueba <- NormalesPL[!datos_conjuntos,]


#Uso de regresion lineal simple para modelar

model <- lm(puntuacion_PL_N~., data=datos_entrenamiento)

summary(model)


#Una vez entrenados los datos es importante descartar las variables menos significativas (menor Pr) aunque que tenga estrellas puede ser importante

#Puede haber una alta correlación o dependencia entre la variable tiempo_N y otras variables en el conjunto de datos, lo que puede llevar a problemas de multicolinealidad. Esto podría hacer que el modelo no pueda estimar un coeficiente específico para tiempo_N.-> tengo edad de ahora y edad de mejor marca personal


#Predicción datos entrenamiento

prediccion_entrenamiento <- predict(model,newdata = datos_entrenamiento)

#Predicción datos prueba

predicciones_prueba<- predict(model, newdata = datos_prueba)


#COMPROBACION MODELO

# Calcula el error cuadrático medio (MSE)

mse_entrenamiento <- mean(( prediccion_entrenamiento- datos_entrenamiento$puntuacion_PL_N )^2)

mse_prueba<- mean((predicciones_prueba- datos_prueba$puntuacion_PL_N)^2)


# Calcula el coeficiente de determinación (R²)

r2_entrenamiento<- 1 - sum((datos_entrenamiento$puntuacion_PL_N - prediccion_entrenamiento)^2) / sum((datos_entrenamiento$puntuacion_PL_N - mean(datos_entrenamiento$puntuacion_PL_N))^2)

r2_prueba <- 1 - sum((datos_prueba$puntuacion_PL_N - predicciones_prueba)^2) / sum((datos_prueba$puntuacion_PL_N - mean(datos_entrenamiento$puntuacion_PL_N))^2)

cat("Error cuadrático medio (MSE):", mse_entrenamiento, "\n")

cat("Error cuadrático medio (MSE):", mse_prueba, "\n")
```

cat("Coeficiente de determinación (R²):", r2_entrenamiento, "\n")

cat("Coeficiente de determinación (R²):", r2_prueba, "\n")

 # Gráfico de dispersión para datos de entrenamiento

plot(datos_entrenamiento$puntuacion_PL_N, prediccion_entrenamiento, main = "Valores reales vs. Valores predichos (Entrenamiento)", xlab = "Valores reales", ylab = "Valores predichos")

 # Gráfico de dispersión para datos de prueba

plot(datos_prueba$puntuacion_PL_N, predicciones_prueba, main = "Valores reales vs. Valores predichos (Prueba)", xlab = "Valores reales", ylab = "Valores predichos")

#Una vez decido el modelo, calculo coeficientes:

 # Obtener los coeficientes del modelo

coeficientes <- coef(model)

 # Obtener los nombres de las variables predictoras

variables <- names(coeficientes)[-1] # Excluimos el coeficiente de intercepción

 # Construir la ecuación de regresión

ecuacion <- paste("puntuacion_PL_N =", coeficientes[1], "+", paste(coeficientes[-1], variables, sep = "", collapse = " + "))*

 # Imprimir la ecuación

print(ecuacion)

Regresión lineal múltiple: otras opciones probadas

 # Importar los datos desde el archivo Excel

datosPC <- readxl::read_excel("VarPC.xlsx")

library(dplyr) # Cargar la librería de manipulación de dataframes "dply

NuevosPC<- select(datosPC,puntuación_PC, edad_MMP, mejor_marca_personal, altura, Skating_skills_PC, Performance_PC, Composition_PC, comp_encasa, ordenPC, Edad, Misma_Nac_PC, Transitions_PC, Interpretation_of_music_PC, tiempo)

 # Renombrar las columnas del dataset

colnames(NuevosPC) <- c("puntuacion_PC", "edad_MMP", "mejor_marca_personal", "altura", "Skating_skills_PC", "Performance_PC", "Composition_PC", "comp_encasa", "ordenPC", "Edad", "Misma_Nac_PC", "Transitions_PC", "Interpretation_of_music_PC", "tiempo")

#Normalizar los datos

NormalesPC<- scale(NuevosPC)

#scale devuelve una matriz-> convertirlo en un dataframe necesario para SVR

NormalesPC <- as.data.frame(NormalesPC)

#Renombrar columnas de la base de datos normalizada

colnames(NormalesPC) <- c("puntuacion_PC_N", "edad_MMP_N", "mejor_marca_personal_N", "altura_N", "Skating_skills_PC_N", "Performance_PC_N", "Composition_PC_N", "comp_encasa_N", "ordenPC_N", "Edad_N", "Misma_Nac_PC_N", "Transitions_PC_N", "Interpretation_of_music_PC_N", "tiempo_N")

#Dividir datos en entrenamiento y prueba

datos_conjuntos<- sample(c(TRUE, FALSE), nrow(NormalesPC), replace=T,prob=c(0.7,0.3))

datos_entrenamiento <- NormalesPC[datos_conjuntos,]

datos_prueba <- NormalesPC[!datos_conjuntos,]

#Una vez entrenados los datos es importante descartar las variables menos significativas (menor Pr) aunque que tenga estrellas puede ser importante

#con todas las variables R^2= 0.846

#Puede haber una alta correlación o dependencia entre la variable tiempo_N y otras variables en el conjunto de datos, lo que puede llevar a problemas de multicolinealidad. Esto podría hacer que el modelo no pueda estimar un coeficiente específico para tiempo_N.-> tengo edad de ahora y edad de mejor marca personal

#Modelo sin variables poco significativas

#model2 <- lm(puntuacion_PC_N~ mejor_marca_personal_N + edad_MMP_N + altura_N + comp_encasa_N + Performance_PC_N + Composition_PC_N, data=datos_entrenamiento)

#summary(model2)

#R^2=0.7727

#Modelo sin variables poco significativas

model3 <- lm(puntuacion_PC_N~ edad_MMP_N+ + mejor_marca_personal_N + altura_N +Skating_skills_PC_N + comp_encasa_N + Performance_PC_N + Composition_PC_N + ordenPC_N + Edad_N + Transitions_PC_N + Interpretation_of_music_PC_N, data=datos_entrenamiento)

summary(model3)

#MODELO3

```
#predicción datos

prediccion_entrenamiento3 <- predict(model3,newdata = datos_entrenamiento)

predicciones_prueba3 <- predict(model3, newdata = datos_prueba)

#COMPROBACION MODEL3

# Calcula el error cuadrático medio (MSE)

mse_entrenamiento3 <- mean(( prediccion_entrenamiento3- datos_entrenamiento$puntuacion_PC_N )^2)

mse_prueba3<- mean((predicciones_prueba3- datos_prueba$puntuacion_PC_N)^2)

# Calcula el coeficiente de determinación (R²)

r2_entrenamiento3<- 1 - sum((datos_entrenamiento$puntuacion_PC_N - prediccion_entrenamiento3)^2) / sum((datos_entrenamiento$puntuacion_PC_N - mean(datos_entrenamiento$puntuacion_PC_N))^2)

r2_prueba3 <- 1 - sum((datos_prueba$puntuacion_PC_N - predicciones_prueba3)^2) / sum((datos_prueba$puntuacion_PC_N - mean(datos_entrenamiento$puntuacion_PC_N))^2)

cat("Error cuadrático medio 3(MSE):", mse_entrenamiento3, "\n")

cat("Error cuadrático medio 3(MSE):", mse_prueba3, "\n")

cat("Coeficiente de determinación 3(R²):", r2_entrenamiento3, "\n")

cat("Coeficiente de determinación 3(R²):", r2_prueba3, "\n")

predichos (Prueba)", xlab = "Valores reales", ylab = "Valores predichos")

#GRAFICO MODEL3

plot(datos_entrenamiento$puntuacion_PC_N, prediccion_entrenamiento3, main = "Valores reales vs. Valores predichos (Entrenamiento)", xlab = "Valores reales", ylab = "Valores predichos")

plot(datos_prueba$puntuacion_PC_N, predicciones_prueba3, main = "Valores reales vs. Valores predichos (Prueba)", xlab = "Valores reales", ylab = "Valores predichos")


#MODELO MEDIANTE STEPWISE

# Ajustar un modelo inicial con todas las variables predictoras

model2 <- lm(puntuacion_PC_N ~ ., data = datos_entrenamiento)

# Aplicar el método stepwise hacia adelante y hacia atrás

step_model <- step(model2, direction = "both")

# Imprimir el resumen del modelo final seleccionado
```

```
summary(step_model)

predicciones_prueba2<- predict(step_model, newdata = datos_prueba)

#COMPROBACIÓN DEL MODELO2

# Calcula el error cuadrático medio (MSE)

mse2 <- mean((predicciones_prueba2- datos_prueba$puntuacion_PC_N)^2)

# Calcula el coeficiente de determinación (R²)

r_squared2 <- 1 - sum((datos_prueba$puntuacion_PC_N - predicciones_prueba2)^2) / sum((datos_prueba$puntuacion_PC_N - mean(datos_entrenamiento$puntuacion_PC_N))^2)

cat("Error cuadrático medio (MSE):", mse2, "\n")

cat("Coeficiente de determinación (R²):", r_squared2, "\n")



#PROGRAMA LARGO

#MODELO2

#predicción datos entrenamiento

prediccion_entrenamiento2 <- predict(model2,newdata = datos_entrenamiento)

#predicción datos prueba

predicciones_prueba2<- predict(model2, newdata = datos_prueba)

#COMPROBACION MODELO

# Calcula el error cuadrático medio (MSE)

mse_entrenamiento2 <- mean(( prediccion_entrenamiento2- datos_entrenamiento$puntuacion_PL_N )^2)

mse_prueba2<- mean((predicciones_prueba2- datos_prueba$puntuacion_PL_N)^2)

# Calcula el coeficiente de determinación (R²)

r2_entrenamiento2<- 1 - sum((datos_entrenamiento$puntuacion_PL_N - prediccion_entrenamiento2)^2) / sum((datos_entrenamiento$puntuacion_PL_N - mean(datos_entrenamiento$puntuacion_PL_N))^2)

r2_prueba2 <- 1 - sum((datos_prueba$puntuacion_PL_N - predicciones_prueba2)^2) / sum((datos_prueba$puntuacion_PL_N - mean(datos_entrenamiento$puntuacion_PL_N))^2)

cat("Error cuadrático medio (MSE):", mse_entrenamiento2, "\n")

cat("Error cuadrático medio (MSE):", mse_prueba2, "\n")
```

```
cat("Coeficiente de determinación (R²):", r2_entrenamiento2, "\n")

cat("Coeficiente de determinación (R²):", r2_prueba2, "\n")

# Gráfico de dispersión para datos de entrenamiento

plot(datos_entrenamiento$puntuacion_PL_N, prediccion_entrenamiento2, main = "Valores reales vs. Valores predichos (Entrenamiento)", xlab = "Valores reales", ylab = "Valores predichos")

# Gráfico de dispersión para datos de prueba

plot(datos_prueba$puntuacion_PL_N, predicciones_prueba2, main = "Valores reales vs. Valores predichos (Prueba)", xlab = "Valores reales", ylab = "Valores predichos")

model2<- lm(puntuacion_PL_N~ mejor_marca_personal_N + altura_N + comp_encasa_N + Performance_PL_N + Transitions_PL_N, data=datos_entrenamiento)

summary(model2)


#MODELO MEDIANTE STEPWISE

# Ajustar un modelo inicial con todas las variables predictoras

model3 <- lm(puntuacion_PL_N ~ ., data = datos_entrenamiento)

# Aplicar el método stepwise hacia adelante y hacia atrás

step_model <- step(model3, direction = "both")

# Imprimir el resumen del modelo final seleccionado

summary(step_model)

predicciones_prueba3<- predict(step_model, newdata = datos_prueba)

#COMPROBACIÓN DEL MODELO2

# Calcula el error cuadrático medio (MSE)

mse3 <- mean((predicciones_prueba3- datos_prueba$puntuacion_PL_N)^2)

# Calcula el coeficiente de determinación (R²)

r_squared3 <- 1 - sum((datos_prueba$puntuacion_PL_N - predicciones_prueba3)^2) / sum((datos_prueba$puntuacion_PL_N - mean(datos_entrenamiento$puntuacion_PL_N))^2)

cat("Error cuadrático medio (MSE):", mse3, "\n")

cat("Coeficiente de determinación (R²):", r_squared3, "\n")
```

Modelo definitivo: máquinas de vectores de soporte programa corto

```
# Importar los datos desde el archivo Excel

datosPC <- readxl::read_excel("VarPC.xlsx")

library(dplyr) # Cargar la librería de manipulación de dataframes "dply

NuevosPC<- select(datosPC,puntuación_PC, edad_MMP, mejor_marca_personal, altura, Skating_skills_PC, Performance_PC, Composition_PC, comp_encasa, ordenPC, Edad, Misma_Nac_PC, Transitions_PC, Interpretation_of_music_PC, tiempo)


# Renombrar las columnas del dataset

colnames(NuevosPC) <- c("puntuacion_PC", "edad_MMP", "mejor_marca_personal", "altura", "Skating_skills_PC", "Performance_PC", "Composition_PC", "comp_encasa", "ordenPC", "Edad", "Misma_Nac_PC", "Transitions_PC", "Interpretation_of_music_PC", "tiempo")


#Normalizar los datos

NormalesPC<- scale(NuevosPC)

#scale devuelve una matriz-> convertirlo en un dataframe necesario para SVR

NormalesPC <- as.data.frame(NormalesPC)


#Renombrar columnas de la base de datos normalizada

colnames(NormalesPC) <- c("puntuacion_PC_N", "edad_MMP_N", "mejor_marca_personal_N", "altura_N", "Skating_skills_PC_N", "Performance_PC_N", "Composition_PC_N", "comp_encasa_N", "ordenPC_N", "Edad_N", "Misma_Nac_PC_N", "Transitions_PC_N", "Interpretation_of_music_PC_N", "tiempo_N")


#dividir datos en entrenamiento y prueba

datos_conjuntos<- sample(c(TRUE, FALSE), nrow(NormalesPC), replace=T,prob=c(0.7,0.3))

datos_entrenamiento <- NormalesPC[datos_conjuntos,]

datos_prueba <- NormalesPC[!datos_conjuntos,]


#Creación del modelo SVR
```

```
library(e1071)
# Creación del modelo SVR con los parámetros óptimos
# Definir los valores de epsilon y cost para la búsqueda de cuadrícula
#epsilon_values <- seq(0, 1, 0.1)
#cost_values <- 2^(2:9)

# Realizar la búsqueda de cuadrícula
#tuneResult <- tune(svm, puntuacion_PC_N ~ ., data = datos_entrenamiento,
#ranges = list(epsilon = epsilon_values, cost = cost_values))

# Imprimir los resultados de la búsqueda de cuadrícula
#print(tuneResult)
# Definir los valores de gamma, cost y epsilon
gamma_value <- 0.1
cost_value <- 256
epsilon_value <- 0
modelo_svr <- svm(puntuacion_PC_N ~ ., data = datos_entrenamiento, kernel = "radial",
          gamma = gamma_value, cost = cost_value, epsilon = epsilon_value, scale = TRUE)

# Predecir valores en los datos de entrenamiento
prediccion_entrenamiento <- predict(modelo_svr, datos_entrenamiento)
# Predecir valores en los datos de prueba
prediccion_prueba <- predict(modelo_svr, datos_prueba)


#COMPROBACIÓN DEL MODELO
summary(modelo_svr)
# Calcular MSE en los datos de entrenamiento
mse_entrenamiento <- mean((datos_entrenamiento$puntuacion_PC_N - prediccion_entrenamiento)^2)
```

```
# Calcular MSE en los datos de prueba

mse_prueba <- mean((datos_prueba$puntuacion_PC_N - prediccion_prueba)^2)

# Imprimir los valores de MSE

print(paste("MSE en datos de entrenamiento:", mse_entrenamiento))

print(paste("MSE en datos de prueba:", mse_prueba))

#Calcular R cuadrado en los datos de entrenamiento

r2_entrenamiento <- 1 - sum((datos_entrenamiento$puntuacion_PC_N - prediccion_entrenamiento)^2) / sum((datos_entrenamiento$puntuacion_PC_N - mean(datos_entrenamiento$puntuacion_PC_N))^2)

# Calcular R cuadrado en los datos de prueba

r2_prueba <- 1 - sum((datos_prueba$puntuacion_PC_N - prediccion_prueba)^2) / sum((datos_prueba$puntuacion_PC_N - mean(datos_entrenamiento$puntuacion_PC_N))^2)

# Imprimir los valores de R cuadrado

print(paste("R cuadrado en datos de entrenamiento:", r2_entrenamiento))

print(paste("R cuadrado en datos de prueba:", r2_prueba))


# Gráfico de dispersión para datos de entrenamiento

plot(datos_entrenamiento$puntuacion_PC_N, prediccion_entrenamiento, main = "Valores reales vs. Valores predichos (Entrenamiento)", xlab = "Valores reales", ylab = "Valores predichos")

# Gráfico de dispersión para datos de prueba

plot(datos_prueba$puntuacion_PC_N, prediccion_prueba, main = "Valores reales vs. Valores predichos (Prueba)", xlab = "Valores reales", ylab = "Valores predichos")


# Calcular la desviación estándar de cada variable en relación a la variable de respuesta

importancia_variables <- apply(NormalesPC[, -1], 2, function(x) sd(x) * sd(NormalesPC$puntuacion_PC_N))

# Crear un data frame con los nombres de las variables y su importancia

variables_influyentes <- data.frame(Variable = colnames(NormalesPC)[-1], Importancia = importancia_variables)

# Ordenar las variables por su importancia en orden descendente

variables_influyentes <- variables_influyentes[order(abs(variables_influyentes$Importancia), decreasing = TRUE), ]
```

```
# Mostrar las variables más influyentes

print("Variables más influyentes:")

print(variables_influyentes)

#EJEMPLO DE APLICACIÓN

nuevos_datos <- readxl::read_excel("pruebaPC.xlsx")

library(dplyr) # Cargar la librería de manipulación de dataframes "dply

ejemploPC<- select(nuevos_datos,puntuación_PC, edad_MMP, mejor_marca_personal, altura, Skating_skills_PC, Performance_PC, Composition_PC, comp_encasa, ordenPC, Edad, Misma_Nac_PC, Transitions_PC, Interpretation_of_music_PC, tiempo)


colnames(ejemploPC) <- c("puntuacion_PC", "edad_MMP", "mejor_marca_personal", "altura", "Skating_skills_PC", "Performance_PC", "Composition_PC", "comp_encasa", "ordenPC", "Edad", "Misma_Nac_PC", "Transitions_PC", "Interpretation_of_music_PC", "tiempo")


# Convertir columnas no numéricas a numéricas y manejar valores faltantes

ejemploPC <- ejemploPC %>%

  mutate_if(~ !is.numeric(.), ~ as.numeric(as.character(.)))


# Verificar si hay NAs en el dataframe

if (anyNA(ejemploPC)) {

  print("Se encontraron valores faltantes (NA) en el dataframe.")

  # Tratar los valores faltantes según tus necesidades

  # ...

} else {

  print("No se encontraron valores faltantes (NA) en el dataframe.")

}

# Imputar valores faltantes con la media

ejemploPC <- apply(ejemploPC, 2, function(x) ifelse(is.na(x), mean(x, na.rm = TRUE), x))


# Normalizar los datos

ejemploPC_normalizados <- scale(ejemploPC)
```

```
ejemploPC_normalizados <- as.data.frame(ejemploPC_normalizados)


colnames(ejemploPC_normalizados) <- c("puntuacion_PC_N", "edad_MMP_N", "mejor_marca_personal_N", "altura_N", "Skating_skills_PC_N", "Performance_PC_N", "Composition_PC_N", "comp_encasa_N", "ordenPC_N", "Edad_N", "Misma_Nac_PC_N", "Transitions_PC_N", "Interpretation_of_music_PC_N", "tiempo_N")


# Asignar valor predeterminado de 0.0 a comp_encasa_N si todos los valores son 0 o NA

if (all(is.na(ejemploPC_normalizados$comp_encasa_N)) || all(ejemploPC_normalizados$comp_encasa_N == 0)) {

  ejemploPC_normalizados$comp_encasa_N <- ifelse(is.na(ejemploPC_normalizados$comp_encasa_N), 0.0, ejemploPC_normalizados$comp_encasa_N)

}

prediccion_nuevos <- predict(modelo_svr, ejemploPC_normalizados)

print(prediccion_nuevos)


# Obtener los parámetros de escala del conjunto de entrenamiento original

mean_entrenamiento <- attr(NormalesPC$puntuacion_PC_N, "scaled:center")

sd_entrenamiento <- attr(NormalesPC$puntuacion_PC_N, "scaled:scale")


# Transformación inversa para obtener los valores en la escala original

prediccion_nuevos_originales <- (prediccion_nuevos * sd(NuevosPC$puntuacion_PC)) + mean(NuevosPC$puntuacion_PC)


# Imprimir los valores en la escala original

print(prediccion_nuevos_originales)
```

Modelo definitivo: máquinas de vectores de soporte programa largo

```
# Importar los datos desde el archivo Excel

datosPL <- readxl::read_excel("VarPL.xlsx")

library(dplyr) # Cargar la librería de manipulación de dataframes "dply
```

NuevosPL<- select(datosPL,puntuación_PL, edad_MMP, mejor_marca_personal, altura, Skating_Skills_PL, Performance_PL, Composition_PL, comp_encasa, ordenPL, Edad, Misma_Nac_PL, Transitions_PL, Interpretation_of_the_Music_PL, tiempo)

 # Renombrar las columnas del dataset

colnames(NuevosPL) <- c("puntuacion_PL", "edad_MMP", "mejor_marca_personal", "altura", "Skating_Skills_PL", "Performance_PL", "Composition_PL", "comp_encasa", "ordenPL", "Edad", "Misma_Nac_PL", "Transitions_PL", "Interpretation_of_music_PL", "tiempo")

#Normalizar los datos

NormalesPL<- scale(NuevosPL)

#scale devuelve una matriz-> convertirlo en un dataframe necesario para SVR

NormalesPL <- as.data.frame(NormalesPL)

#Renombrar columnas de la base de datos normalizada

colnames(NormalesPL) <- c("puntuacion_PL_N", "edad_MMP_N", "mejor_marca_personal_N", "altura_N", "Skating_skills_PL_N", "Performance_PL_N", "Composition_PL_N", "comp_encasa_N", "ordenPL_N", "Edad_N", "Misma_Nac_PL_N", "Transitions_PL_N", "Interpretation_of_music_PL_N", "tiempo_N")

#dividir datos en entrenamiento y prueba

datos_conjuntos<- sample(c(TRUE, FALSE), nrow(NormalesPL), replace=T,prob=c(0.7,0.3))

datos_entrenamiento <- NormalesPL[datos_conjuntos,]

datos_prueba <- NormalesPL[!datos_conjuntos,]

#Creación del modelo SVR

library(e1071)

 # Creación del modelo SVR con los parámetros óptimos

 # Definir los valores de epsilon y cost para la búsqueda de cuadrícula

#epsilon_values <- seq(0, 1, 0.1)

#cost_values <- 2^(2:9)

```
# Realizar la búsqueda de cuadrícula
#tuneResult <- tune(svm, puntuacion_PL_N ~ ., data = datos_entrenamiento,
#ranges = list(epsilon = epsilon_values, cost = cost_values))
# Imprimir los resultados de la búsqueda de cuadrícula
#print(tuneResult)
# Definir los valores de gamma, cost y epsilon
gamma_value <- 0.1
cost_value <- 512
epsilon_value <- 0
modelo_svr <- svm(puntuacion_PL_N ~ ., data = datos_entrenamiento, kernel = "radial",
            gamma = gamma_value, cost = cost_value, epsilon = epsilon_value, scale = TRUE)
summary(modelo_svr)

# Predecir valores en los datos de entrenamiento
prediccion_entrenamiento <- predict(modelo_svr, datos_entrenamiento)
# Predecir valores en los datos de prueba
prediccion_prueba <- predict(modelo_svr, datos_prueba)

#COMPROBACIÓN DEL MODELO
# Calcular MSE en los datos de entrenamiento
mse_entrenamiento <- mean((datos_entrenamiento$puntuacion_PL_N - prediccion_entrenamiento)^2)
# Calcular MSE en los datos de prueba
mse_prueba <- mean((datos_prueba$puntuacion_PL_N - prediccion_prueba)^2)
# Imprimir los valores de MSE
print(paste("MSE en datos de entrenamiento:", mse_entrenamiento))
print(paste("MSE en datos de prueba:", mse_prueba))
#Calcular R cuadrado en los datos de entrenamiento
r2_entrenamiento <- 1 - sum((datos_entrenamiento$puntuacion_PL_N - prediccion_entrenamiento)^2) / sum((datos_entrenamiento$puntuacion_PL_N - mean(datos_entrenamiento$puntuacion_PL_N))^2)
```

```
# Calcular R cuadrado en los datos de prueba

r2_prueba <- 1 - sum((datos_prueba$puntuacion_PL_N - prediccion_prueba)^2) / sum((datos_prueba$puntuacion_PL_N - mean(datos_entrenamiento$puntuacion_PL_N))^2)

# Imprimir los valores de R cuadrado

print(paste("R cuadrado en datos de entrenamiento:", r2_entrenamiento))

print(paste("R cuadrado en datos de prueba:", r2_prueba))


# Gráfico de dispersión para datos de entrenamiento

plot(datos_entrenamiento$puntuacion_PL_N, prediccion_entrenamiento, main = "Valores reales vs. Valores predichos (Entrenamiento)", xlab = "Valores reales", ylab = "Valores predichos")

# Gráfico de dispersión para datos de prueba

plot(datos_prueba$puntuacion_PL_N, prediccion_prueba, main = "Valores reales vs. Valores predichos (Prueba)", xlab = "Valores reales", ylab = "Valores predichos")


# Calcular la desviación estándar de cada variable en relación a la variable de respuesta

importancia_variables <- apply(NormalesPL[, -1], 2, function(x) sd(x) * sd(NormalesPL$puntuacion_PL_N))


# Crear un data frame con los nombres de las variables y su importancia

variables_influyentes <- data.frame(Variable = colnames(NormalesPL)[-1], Importancia = importancia_variables)


# Ordenar las variables por su importancia en orden descendente

variables_influyentes <- variables_influyentes[order(abs(variables_influyentes$Importancia), decreasing = TRUE), ]


# Mostrar las variables más influyentes

print("Variables más influyentes:")

print(variables_influyentes)

#EJEMPLO DE APLICACIÓN

nuevos_datos <- readxl::read_excel("pruebaPL.xlsx")
```

```
library(dplyr) # Cargar la librería de manipulación de dataframes "dply

ejemploPL<- select(nuevos_datos,puntuación_PL, edad_MMP, mejor_marca_personal, altura, Skating_skills_PL, Performance_PL, Composition_PL, comp_encasa, ordenPL, Edad, Misma_Nac_PL, Transitions_PL, Interpretation_of_music_PL, tiempo)


colnames(ejemploPL) <- c("puntuacion_PL", "edad_MMP", "mejor_marca_personal", "altura", "Skating_skills_PL", "Performance_PL", "Composition_PL", "comp_encasa", "ordenPL", "Edad", "Misma_Nac_PL", "Transitions_PL", "Interpretation_of_music_PL", "tiempo")


# Convertir columnas no numéricas a numéricas y manejar valores faltantes

ejemploPL <- ejemploPL %>%

  mutate_if(~ !is.numeric(.), ~ as.numeric(as.character(.)))


# Verificar si hay NAs en el dataframe

if (anyNA(ejemploPL)) {

  print("Se encontraron valores faltantes (NA) en el dataframe.")

  # Tratar los valores faltantes según tus necesidades

  # ...

} else {

  print("No se encontraron valores faltantes (NA) en el dataframe.")

}

# Imputar valores faltantes con la media

ejemploPL <- apply(ejemploPL, 2, function(x) ifelse(is.na(x), mean(x, na.rm = TRUE), x))


# Normalizar los datos

ejemploPL_normalizados <- scale(ejemploPL)

ejemploPL_normalizados <- as.data.frame(ejemploPL_normalizados)


colnames(ejemploPL_normalizados) <- c("puntuacion_PL_N", "edad_MMP_N", "mejor_marca_personal_N", "altura_N", "Skating_skills_PL_N", "Performance_PL_N", "Composition_PL_N", "comp_encasa_N", "ordenPL_N", "Edad_N", "Misma_Nac_PL_N", "Transitions_PL_N", "Interpretation_of_music_PL_N", "tiempo_N")
```

```
# Asignar valor predeterminado de 0.0 a comp_encasa_N si todos los valores son 0 o NA
if (all(is.na(ejemploPL_normalizados$comp_encasa_N)) || all(ejemploPL_normalizados$comp_encasa_N == 0)) {
  ejemploPL_normalizados$comp_encasa_N <- ifelse(is.na(ejemploPL_normalizados$comp_encasa_N), 0.0, ejemploPL_normalizados$comp_encasa_N)
}
prediccion_nuevos <- predict(modelo_svr, ejemploPL_normalizados)
print(prediccion_nuevos)

# Obtener los parámetros de escala del conjunto de entrenamiento original
mean_entrenamiento <- attr(NormalesPL$puntuacion_PL_N, "scaled:center")
sd_entrenamiento <- attr(NormalesPL$puntuacion_PL_N, "scaled:scale")

# Transformación inversa para obtener los valores en la escala original
prediccion_nuevos_originales <- (prediccion_nuevos * sd(NuevosPL$puntuacion_PL)) + mean(NuevosPL$puntuacion_PL)

# Imprimir los valores en la escala original
print(prediccion_nuevos_originales)
```

Obtención de las gráficas de dispersión

```
#PROGRAMA CORTO
# Importar los datos desde el archivo Excel
datosPC <- readxl::read_excel("VarPC.xlsx")
library(dplyr) # Cargar la librería de manipulación de dataframes "dply
NuevosPC<- select(datosPC,puntuación_PC, edad_MMP, mejor_marca_personal, altura, Skating_skills_PC, Performance_PC, Composition_PC, comp_encasa, ordenPC, Edad, Misma_Nac_PC, Transitions_PC, Interpretation_of_music_PC, tiempo)

# Renombrar las columnas del dataset
```

colnames(NuevosPC) <- c("puntuacion_PC", "edad_MMP", "mejor_marca_personal", "altura", "Skating_skills_PC", "Performance_PC", "Composition_PC", "comp_encasa", "ordenPC", "Edad", "Misma_Nac_PC", "Transitions_PC", "Interpretation_of_music_PC", "tiempo")

#Normalizar los datos

NormalesPC<- scale(NuevosPC)

#scale devuelve una matriz-> convertirlo en un dataframe necesario para SVR

NormalesPC <- as.data.frame(NormalesPC)

#Renombrar columnas de la base de datos normalizada

colnames(NormalesPC) <- c("puntuacion_PC_N", "edad_MMP_N", "mejor_marca_personal_N", "altura_N", "Skating_skills_PC_N", "Performance_PC_N", "Composition_PC_N", "comp_encasa_N", "ordenPC_N", "Edad_N", "Misma_Nac_PC_N", "Transitions_PC_N", "Interpretation_of_music_PC_N", "tiempo_N")

#dividir datos en entrenamiento y prueba

datos_conjuntosPC<- sample(c(TRUE, FALSE), nrow(NormalesPC), replace=T,prob=c(0.7,0.3))

datos_entrenamientoPC <- NormalesPC[datos_conjuntosPC,]

datos_pruebaPC <- NormalesPC[!datos_conjuntosPC,]

#PROGRAMA LARGO

 # Importar los datos desde el archivo Excel

datosPL <- readxl::read_excel("VarPL.xlsx")

library(dplyr) # Cargar la librería de manipulación de dataframes "dply

NuevosPL<- select(datosPL,puntuación_PL, edad_MMP, mejor_marca_personal, altura, Skating_Skills_PL, Performance_PL, Composition_PL, comp_encasa, ordenPL, Edad, Misma_Nac_PL, Transitions_PL, Interpretation_of_the_Music_PL, tiempo)

 # Renombrar las columnas del dataset

colnames(NuevosPL) <- c("puntuacion_PL", "edad_MMP", "mejor_marca_personal", "altura", "Skating_Skills_PL", "Performance_PL", "Composition_PL", "comp_encasa", "ordenPL", "Edad", "Misma_Nac_PL", "Transitions_PL", "Interpretation_of_music_PL", "tiempo")

#Normalizar los datos

NormalesPL<- scale(NuevosPL)

#scale devuelve una matriz-> convertirlo en un dataframe necesario para SVR

```
NormalesPL <- as.data.frame(NormalesPL)

#Renombrar columnas de la base de datos normalizada

colnames(NormalesPL) <- c("puntuacion_PL_N", "edad_MMP_N", "mejor_marca_personal_N", "altura_N", "Skating_skills_PL_N", "Performance_PL_N", "Composition_PL_N", "comp_encasa_N", "ordenPL_N", "Edad_N", "Misma_Nac_PL_N", "Transitions_PL_N", "Interpretation_of_music_PL_N", "tiempo_N")

#dividir datos en entrenamiento y prueba

datos_conjuntosPL<- sample(c(TRUE, FALSE), nrow(NormalesPL), replace=T,prob=c(0.7,0.3))

datos_entrenamientoPL <- NormalesPL[datos_conjuntosPL,]

datos_pruebaPL <- NormalesPL[!datos_conjuntosPL,]


#DIAGRAMAS DE DISPERSION

library(ggplot2)

#puntuación-edad

ggplot(NuevosPC, aes(x =Edad , y =puntuacion_PC)) +

  geom_point(color = "blue") +

  labs(x = "Edad", y = "Puntuación PC", title = "Diagrama de Dispersión - Puntuación programa corto vs. Edad")

ggplot(NuevosPL, aes(x =Edad , y =puntuacion_PL)) +

  geom_point(color = "red") +

  labs(x = "Edad", y = "Puntuación PL", title = "Diagrama de Dispersión - Puntuación programa largo vs. Edad")

library(dplyr)

# Contar el número de patinadoras por edad PC

tabla_edadesPC <- NuevosPC %>% count(Edad)

# Mostrar la tabla de frecuencias

print(tabla_edadesPC)

# Contar el número de patinadoras por edad PL

tabla_edadesPL <- NuevosPL %>% count(Edad)

# Mostrar la tabla de frecuencias

print(tabla_edadesPL)

#Puntuación-altura
```

```
ggplot(NuevosPC, aes(x =altura , y =puntuacion_PC)) +
  geom_point(color = "blue") +
  labs(x = "Altura", y = "Puntuación PC", title = "Diagrama de Dispersión - Puntuación programa corto vs. Altura")
ggplot(NuevosPL, aes(x =altura , y =puntuacion_PL)) +
  geom_point(color = "red") +
  labs(x = "Altura", y = "Puntuación PL", title = "Diagrama de Dispersión - Puntuación programa largo vs. Altura")
library(dplyr)
# Contar el número de patinadoras por edad PC
tabla_alturaPC <- NuevosPC %>% count(altura)
# Mostrar la tabla de frecuencias
print(tabla_alturaPC)
# Contar el número de patinadoras por edad PL
tabla_alturaPL <- NuevosPL %>% count(altura)
# Mostrar la tabla de frecuencias
print(tabla_alturaPL)
#Puntuación-Mejor marca personal
ggplot(NuevosPC, aes(x =mejor_marca_personal , y =puntuacion_PC)) +
  geom_point(color = "blue") +
  labs(x = "Mejor marca personal", y = "Puntuación PC", title = "Diagrama de Dispersión - Puntuación programa corto vs. Mejor marca ")
ggplot(NuevosPL, aes(x =mejor_marca_personal , y =puntuacion_PL)) +
  geom_point(color = "red") +
  labs(x = "Mejor marca personal", y = "Puntuación PL", title = "Diagrama de Dispersión - Puntuación programa largo vs. Mejor marca")
#Puntuación-orden
ggplot(NuevosPC, aes(x =ordenPC, y =puntuacion_PC)) +
  geom_point(color = "blue") +
  labs(x = "Orden", y = "Puntuación PC", title = "Diagrama de Dispersión - Puntuación programa corto vs. Orden")
```

```
ggplot(NuevosPL, aes(x =ordenPL , y =puntuacion_PL)) +
  geom_point(color = "red") +
  labs(x = "Orden", y = "Puntuación PL", title = "Diagrama de Dispersión - Puntuación programa largo vs. Orden")
#skating skills
ggplot(NuevosPC, aes(x =Skating_skills_PC, y =puntuacion_PC)) +
  geom_point(color = "blue") +
  labs(x = "Habilidades de la patinadora en PC", y = "Puntuación PC", title = "Diagrama de Dispersión - Puntuación programa corto vs. Orden")
ggplot(NuevosPL, aes(x =Skating_Skills_PL , y =puntuacion_PL)) +
  geom_point(color = "red") +
  labs(x = "Habilidades de la patinadora en PL", y = "Puntuación PL", title = "Diagrama de Dispersión - Puntuación programa largo vs. Orden")
#performance
ggplot(NuevosPC, aes(x =Performance_PC, y =puntuacion_PC)) +
  geom_point(color = "blue") +
  labs(x = "Actuación de la patinadora en PC", y = "Puntuación PC")
ggplot(NuevosPL, aes(x =Performance_PL , y =puntuacion_PL)) +
  geom_point(color = "red") +
  labs(x = "Actuación de la patinadora en PL", y = "Puntuación PL")
#transitions
ggplot(NuevosPC, aes(x =Transitions_PC, y =puntuacion_PC)) +
  geom_point(color = "blue") +
  labs(x = "Transiciones de la patinadora en PC", y = "Puntuación PC")
ggplot(NuevosPL, aes(x =Transitions_PL , y =puntuacion_PL)) +
  geom_point(color = "red") +
  labs(x = "Transiciones de la patinadora en PL", y = "Puntuación PL")
#Compisition
ggplot(NuevosPC, aes(x =Composition_PC, y =puntuacion_PC)) +
  geom_point(color = "blue") +
  labs(x = "Composición del  PC", y = "Puntuación PC")
ggplot(NuevosPL, aes(x =Composition_PL , y =puntuacion_PL)) +
```

```
  geom_point(color = "red") +
  labs(x = "Composición del PL", y = "Puntuación PL")
#Int
ggplot(NuevosPC, aes(x =Interpretation_of_music_PC, y =puntuacion_PC)) +
  geom_point(color = "blue") +
  labs(x = "Interpretación de la música en el PC", y = "Puntuación PC")
ggplot(NuevosPL, aes(x =Interpretation_of_music_PL , y =puntuacion_PL)) +
  geom_point(color = "red") +
  labs(x = "Interpretación de la música en el PL", y = "Puntuación PL")
```