FACULTAD DE INGENIERÍA

Deep Learning
Juan José Alzate Molina.
C.C.1007232151

Entrega Final: Proyecto Tennis Predictions con Deep Learning

1. Contexto

La predicción de resultados en el deporte profesional ha sido históricamente un área de gran interés tanto para analistas deportivos como para el sector de las apuestas. Con la creciente disponibilidad de datos detallados —como estadísticas de jugadores, históricos de enfrentamientos, condiciones de juego y métricas en tiempo real—, los métodos tradicionales de análisis han dado paso a enfoques más avanzados basados en inteligencia artificial (IA).

La introducción del deep learning ha revolucionado este campo, permitiendo construir modelos que aprenden representaciones complejas de los datos y capturan patrones no evidentes para el análisis tradicional. Modelos como redes neuronales recurrentes (RNN), redes neuronales convolucionales (CNN) adaptadas a series temporales, y arquitecturas basadas en transformers han mostrado un gran potencial en la predicción de resultados deportivos.

En este contexto, la combinación de datos históricos, información en tiempo real y algoritmos de IA avanzados abre nuevas posibilidades tanto para mejorar la precisión de las predicciones como para entender mejor los factores que influyen en el desempeño deportivo.

2. Objetivo del Proyecto

El objetivo de este proyecto es desarrollar un modelo de deep learning capaz de predecir el porcentaje de probabilidad de victoria de un jugador en partidos de tenis profesional masculino. Para ello, se utilizarán los datos históricos disponibles en el repositorio TML-Database, entrenando y evaluando el desempeño del modelo en

función de su capacidad de generalizar sobre nuevos encuentros. En busca del mejor resultado posible se exploró una segunda base de datos. Ambas bases son descritas en el siguiente numeral.

3. Bases de Datos

3.1.1 TML Database

Los datos utilizados provienen del repositorio público 'TML-Database' (https://github.com/Tennismylife/TML-Database), el cual contiene archivos CSV con resultados de partidos del circuito ATP desde el año 2000. Para este proyecto se consideraron los archivos correspondientes al período 2015-2025, con el fin de trabajar con datos actuales y relevantes.

Cada archivo anual incluye columnas que describen los torneos, jugadores y estadísticas de desempeño durante el partido. A continuación, se detallan algunas de las columnas más importantes:

- winner_name / loser_name: nombres del jugador ganador y perdedor del partido.
- winner_rank / loser_rank: posición en el ranking ATP de ambos jugadores al momento del partido.
- surface: tipo de superficie donde se jugó el partido (Hard, Clay, Grass).
- tourney_date: fecha del torneo.
- w_ace / l_ace: cantidad de aces por cada jugador.
- w_df / l_df: doble faltas cometidas.
- w_svpt / l_svpt: puntos al servicio jugados.
- w_1stIn / l_1stIn: primeros servicios dentro.
- w_1stWon / l_1stWon: puntos ganados con primer servicio.
- w_2ndWon / l_2ndWon: puntos ganados con segundo servicio.

Estas variables permiten caracterizar el rendimiento de los jugadores y su estilo de juego, factores relevantes para predecir el resultado de futuros encuentros.

3.1.1 Procesamiento TML Database

A la base de datos se le realizó una limpieza de valores nulos y fuertemente atípicos, para evitar que estos afectarán la capacidad predictiva. Para evitar la fuga de información, se descartó el uso de estadísticas obtenidas durante el mismo partido que se desea predecir. En su lugar, se calcularon estadísticas históricas previas para cada jugador basadas en sus últimos partidos: promedio de aces, porcentaje de victorias, superficie favorita y ranking actual. Además, para cada partido se crearon dos instancias: una donde el jugador 1 es el ganador (target=1) y otra donde es el perdedor (target=0), generando un dataset simétrico y balanceado.

3.2.1 Tennis UK Database

En busca de la posibilidad de mejores resultado se tomó una segunda base de datos, en este caso desde http://www.tennis-data.co.uk/alldata.php.

La base de datos original contiene información detallada de partidos de tenis ATP, incluyendo metadatos del torneo, jugadores participantes, rankings, puntos ATP, superficie, ronda del torneo y probabilidades de casas de apuestas.

Campos iniciales más relevantes:

- Date: Fecha del partido
- Tournament, Location, Series, Court, Surface: Información del torneo
- Winner / Loser: Jugadores participantes
- WRank / LRank: Ranking ATP antes del partido
- WPts / LPts: Puntos ATP antes del partido
- AvgW / AvgL: Probabilidades promedio de apuestas

Se realizó una limpieza de columnas post-partido que no pueden usarse para predicción, como resultado en sets, juegos ganados, score y estadísticas derivadas del marcador. También se eliminaron columnas redundantes o con información irrelevante.

3.2.2 Tennis UK Database Enriquecido

Se agregaron múltiples columnas derivadas del historial entre jugadores, utilizando únicamente información disponible antes del partido:

- Wh2h / Lh2h: Diferencia de partidos ganados contra ese oponente antes de la fecha del partido.
- Wh2h_surf / Lh2h_surf: Lo mismo pero restringido a la misma superficie.
- Whist / Lhist: Diferencia de victorias-derrotas totales del jugador hasta ese momento.
- Whist_surf / Lhist_surf: Historial total pero específico a la superficie del partido.
- W10 / L10: Rendimiento en los últimos 10 partidos.
- W10_surf / L10_surf: Rendimiento en los últimos 10 partidos en la misma superficie.

4. Post-procesamiento para modelado

Para evitar el sesgo de posición (ya que el jugador ganador siempre está en la columna 'Winner'), se duplicó cada fila del dataset invirtiendo las posiciones del ganador y perdedor, así como sus respectivas métricas. Se añadió una columna 'target' con valor 1 para la fila original (ganó el jugador en columna izquierda) y 0 para la fila invertida.

El dataset final fue codificado utilizando Label Encoding en variables categóricas y normalizado con StandardScaler. Se dividió el conjunto en entrenamiento y prueba utilizando train_test_split con estratificación en el target.

En el caso del dataset TML se hicieron iteraciones considerando diferente cantidad de partidos previos para el cálculo de las estadísticas promedio (5, 10 y 50 partidos previos)

5. Arquitecturas utilizadas

5.1 Perceptrón Multicapa (MLP)

El modelo MLP es una red neuronal feedforward compuesta por múltiples capas densas. Es ideal para trabajar con datos tabulares ya que puede modelar relaciones no lineales entre variables. En este proyecto se utilizó una arquitectura profunda con capas de 256, 128, 64, 32 y 16 neuronas, todas con función de activación ReLU, y una salida con activación sigmoide.

Ventajas:

- Buena capacidad de generalización con regularización (Dropout).
- Sencilla de entrenar y ajustar.

Desventajas:

- Requiere la normalización previa de los datos.
- Poco interpretable comparado con modelos como árboles de decisión.

5.2 Red Convolucional 1D (CNN 1D)

Se exploró una arquitectura CNN 1D aplicada sobre vectores estructurados que contienen estadísticas concatenadas de ambos jugadores. Esta arquitectura intenta captar patrones locales entre bloques de características del jugador 1 frente al jugador 2. Sin embargo, debido a la baja dimensionalidad del vector de entrada (pocos features), esta red tiende a perder información con el uso de capas de pooling.

Ventajas:

- Puede detectar patrones entre combinaciones específicas de estadísticas.

Desventajas:

- Menor desempeño que el MLP en datos tabulares.

- Problemas con reducción de dimensión en datasets pequeños.

5.3 XGBoost

XGBoost es un modelo de boosting basado en árboles de decisión que ha demostrado gran eficacia en tareas de clasificación tabular. No requiere normalización y es capaz de modelar interacciones complejas entre variables. Fue utilizado como comparación frente a las redes neuronales.

Ventajas:

- Alta precisión en datos estructurados.
- Manejo nativo de datos faltantes.
- Interpretabilidad de features mediante importancia de variables.

Desventajas:

- No es adecuado para tareas con entradas secuenciales o espaciales como imágenes o texto.

6. Resultados

6.1 TML Database

Durante el entrenamiento y validación de los modelos, se observaron desempeños consistentes en torno al 65% de precisión en el conjunto de prueba, tanto para el modelo de perceptrón multicapa (MLP) como para el modelo XGBoost. Las métricas complementarias también respaldan este rendimiento.

Un hallazgo relevante fue que al incrementar el número de partidos históricos considerados para calcular las métricas previas (como winrate o desempeño en superficie), se evidenció una ligera pero consistente mejora en la capacidad predictiva del modelo. Este resultado sugiere que tener una muestra más representativa y amplia del rendimiento reciente de cada jugador permite estimaciones más estables y precisas de su probabilidad de victoria. Sin embargo el cambio no fue lo suficientemente notorio

6.2 UK Database

Los modelos entrenados con la base de datos original (sin métricas adicionales, pero con una buena selección de variables y normalización adecuada) alcanzaron una precisión cercana al 69%, lo cual representa un resultado altamente competitivo en comparación con los modelos reportados en la literatura especializada para predicción de partidos de tenis ATP. Este resultado valida la utilidad del dataset base y confirma que, aun con variables limitadas, es posible capturar patrones relevantes de rendimiento entre jugadores.

Por otro lado, al utilizar el dataset enriquecido con métricas históricas adicionales (head-to-head, historial reciente, rendimiento por superficie), no se observaron mejoras significativas en la precisión. De hecho, algunos modelos regresaron a una precisión cercana al azar (50%) cuando el procesamiento no fue cuidadoso o se introdujeron linealidades no manejadas correctamente. Por ende es necesario profundizar en la simetrización de estas estadísticas adicionales y los modelos efectivos, pues XGBoost presentó un rendimiento mejor que los demás (lo cual fue un caso atípico)

7. Conclusiones

Este estudio demuestra que es posible construir un modelo predictivo efectivo para partidos de tenis profesional utilizando únicamente información disponible antes del encuentro. La estrategia de construir un dataset simétrico, balanceado y fundamentado en métricas históricas probó ser clave para evitar fugas de información y lograr una generalización adecuada.

Entre los modelos evaluados:

El MLP profundo mostró un rendimiento competitivo y robusto, siendo sensible a una buena representación de las estadísticas históricas.

La CNN 1D, si bien interesante desde un enfoque arquitectónico, se vio limitada por la baja dimensionalidad del input y la estructura no secuencial de los datos.

El modelo XGBoost destacó por su facilidad de entrenamiento, interpretabilidad y rendimiento comparable al del MLP. Sin embargo exceptuando el database enriquecido, no presentó un rendimiento mejor que otros modelos

Una de las conclusiones más relevantes es que la calidad y profundidad del historial estadístico de los jugadores es un factor determinante en la eficacia del modelo. Cuantos más partidos se usen para generar features previos, mejores serán las predicciones. A futuro, se recomienda incluir otras variables contextuales como la edad, la experiencia en torneos específicos y el historial directo entre jugadores, para enriquecer aún más las predicciones.

8. Referencias

1. **Beal, R., Middleton, S. E., Norman, T. J., & Ramchurn, S. D. (2021). Combining Machine Learning and Human Experts to Predict Match Outcomes in Football: A Baseline Model. Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21), 15447–15450. https://github.com/RyanBeal7/GuardianPreviewData**

2. **Lisi, F., & Zanella, G. (2013). Tennis betting: Can statistics beat bookmakers? Electronic Journal of Applied Statistical Analysis, 10(3), 790–807. https://doi.org/10.1285/i20705948v10n3p790**

3. **Dumović, M. (s. f.). *Tennis match predictions using neural networks*. Stanford University. Repositorio**