FACULTAD DE INGENIERÍA

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

Entrega 1: Proyecto Tennis Predictions con Deep Learning

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.

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.

Datos

El repositorio a usar se basó originalmente en el de Jeff Sackmann. Mantuvieron los mismos patrones en los nombres de las columnas y su orden. Sin embargo, al utilizar los archivos CSV de Sackmann, encontraron datos faltantes.Otra diferencia significativa con respecto a la base de datos de Sackmann es el uso de los identificadores de jugadores de la ATP, lo que facilita el cálculo de registros y la búsqueda de datos en el sitio web de la ATP.

Las columnas son:

Estructura Principal del Dataset - TML-Database

- tourney_id
- tourney_name
- surface
- draw_size
- tourney_level
- match_num
- winner_id
- winner_name
- loser_id
- loser_name
- score
- best_of
- round
- minutes
- w_ace
- w_df
- w_svpt
- l_ace
- l_df
- l_svpt

Todos los campos son tipo string y cuenta con más de 180.000 registros de partidos pasados, que puedan permitir la identificación de patrones. Adicional se intentará generar métricas adicionales provenientes del mismo dataset, como el head to head o racha de los últimos N partidos.

Evaluación

Evaluación de Machine Learning:

El modelo será evaluado principalmente como un modelo de predicción probabilística, dado que predice la probabilidad de victoria de un jugador. Se utilizarán métricas específicas como:

- **Log Loss (Logaritmo de pérdida):** mide la precisión de las probabilidades predichas; penaliza las predicciones muy seguras pero incorrectas.
- **Brier Score:** mide la media cuadrática de la diferencia entre las probabilidades predichas y los resultados reales.
- **AUC-ROC**: mide la capacidad de discriminación del modelo entre ganadores y perdedores.
- **Accuracy (exactitud):** en versiones donde se tome la predicción más probable como clasificación directa (opcional).

Evaluación de Negocio:

En el contexto de apuestas deportivas y predicciones de resultados:

- **Return on Investment (ROI) simulado:** evaluar si apostar sistemáticamente en función de las probabilidades predichas generaría ganancias a lo largo del tiempo.
- **Hit Rate:** porcentaje de predicciones en las que el jugador predicho como favorito gana el partido.

Referencias y Resultados Previos

Estudios previos han demostrado que es posible alcanzar precisiones superiores al 70% en la predicción de partidos de tenis utilizando modelos de machine learning clásicos como Random Forests y Gradient Boosting. Trabajos recientes basados en deep learning, como redes neuronales profundas y modelos de secuencias (RNN, LSTM), han mostrado mejoras adicionales en la calidad de las predicciones. Además, investigaciones como las de Klaassen y Magnus (2001) y diversos análisis de Kaggle han evidenciado la relevancia de factores como el tipo de superficie, el ranking y el historial entre jugadores en la predicción de resultados.