Checkpoint 2: Problem Definition & Topic Justification

Nicholas Condos

November 10, 2025

Project Title & Updated Topic Summary

Title: *Modeling ATP Match Outcomes Using Serve, Return, and Performance Efficiency Metrics*

Building on Checkpoint 1, this project expands its original serve-focused approach to include a broader set of performance metrics that influence ATP match outcomes.

Using match data from 2008–2024, the goal is to determine which different metrics (serve accuracy, return performance, and efficiency ratios such as winners-to-unforced-errors) most strongly predict match success across different playing surfaces.

Problem Definition & Justification

The guiding question for this study is:

> *Which serve, return, and performance-efficiency factors most significantly influence ATP match outcomes across hard, clay, and grass surfaces?*

While serve statistics are known to impact match success, many existing datasets isolate serve metrics without looking at returns, unforced errors, and other important statistics. This research takes a more holistic view of player performance, aligning with the increasing use of analytics in professional tennis for decision-making and preparation.

Stakeholder Relevance:

- **Coaches & Performance Teams:** Use statistical insights to tailor match strategy and training programs.
- **Players & Support Staff:** Identify measurable strengths and weaknesses for targeted development.
- **Sports Analysts & Consultants:** Build predictive dashboards and reports using model results.
- **Media & Tournament Organizations:** Translate findings into contextual insights and broadcast analysis.

This work aims to bridge the gap between raw match statistics and actionable competitive insights in professional and high level tennis.

Checkpoint 2: Problem Definition & Topic Justification

Key Metrics (KPIs):

KPI	Description	Expected Impact on Win Probability
First Serve Percentage (%1stIn)	Accuracy of first serves	Higher % correlates with more service points won
Aces and Double Faults	Serve power vs. control	Balance between aggression and precision
Break Points Saved/Won	Pressure performance	Key for momentum conversion
Return Points Won (%RPW)	Effectiveness in return games	Higher values signal well-rounded players
Winners to Unforced Errors Ratio	Efficiency metric	Reflects shot selection and composure

Unit of Analysis & Scope

- **Unit of Analysis:** Individual ATP match
- **Observation Period:** 2008 – 2024
- **Filters:** Main-draw matches only; exclude walkovers, retirements, and incomplete matches
- **Segmentation:** Surface type (hard, clay, grass) for comparative modeling

This timeframe ensures consistency with Checkpoint 1 and provides coverage with reliable match statistics.

Data Sources & Access Plan

Primary dataset: **“tennis_atp” by Jeff Sackmann** (GitHub).

- Annual CSV files (atp_matches_YYYY.csv) containing match-level data for each season.
- Supplementary files (atp_players.csv, atp_rankings.csv) for player attributes and ranking context.
- Total coverage: 60,000 + matches spanning three decades.

Checkpoint 2: Problem Definition & Topic Justification

- ○ To be condensed to 2008-2024 data

The dataset is freely available, open-source, and well-documented (matches_data_dictionary.txt), making it fully suitable for academic and reproducible analysis.

Literature Scan

1. **Wei et al. (2016)** – *Journal of Sports Sciences*: First-serve success and break-point conversion identified as primary predictors of ATP match outcomes.

2. **Crespo & Reid (2019)** – *ITF Coaching Review*: Argued that serve dominance and return consistency together drive overall performance efficiency.

3. **Kovalchik (2021)** – *Machine Learning in Sports*: Applied classification models to ATP data, finding that combining serve and return variables improves predictive accuracy by ~15%.

These studies validate the approach of this project and reinforce its relevance to coaching and sports analytics.

Exploratory & Preprocessing Plan

Exploratory Data Analysis (EDA):

- Compute summary statistics and correlations for serve/return metrics.
- Visualize distributions/performance patterns by surface.
- Detect missing values and outliers (e.g., extreme serve counts).

Preprocessing Steps:

- Handle missing values using mean/median imputation.
- Normalize numeric variables with z-score scaling.
- Encode categorical fields (surface, handedness).
- Create new features such as *serve-return balance index* and *efficiency ratio* for modeling.

Checkpoint 2: Problem Definition & Topic Justification

Modeling:

The analysis will begin with:

- **Logistic Regression:** A baseline model to predict match outcomes (win/loss) using serve, return, and efficiency metrics.
- A secondary model to capture non-linear patterns and identify the most important features.

Model performance will be evaluated using **accuracy, AUC, precision, and recall** scores.

Risks & Ethical Considerations

- **Missing Data:** Will be handled through imputation and clear documentation.
- **Surface Imbalance:** Addressed through stratified sampling across hard, clay, and grass courts.
- **Ethical Use:** The dataset contains only public match records, so there are no privacy concerns.

Next Steps

1. Load and review the 2008–2024 ATP match data for completeness and accuracy.
2. Conduct EDA and create visuals to compare performance by surface.
3. Develop new combined metrics (e.g., serve-return balance, winners/errors ratio).
4. Compile findings and a data quality report for Checkpoint 3.