Modelling the Probability of Match Outcomes in Tier 1 Rugby Union International Test Matches

Alexander White
WHTALE015@myuct.ac.za
Computer Science, University of Cape Town

Nicolas Wise
WSXNIC001@myuct.ac.za
Computer Science, University of Cape Town

1 Problem Formulation

This project addresses the lack of objective, data-driven tools for predicting rugby outcomes in South Africa, a gap that limits the analytical depth, accuracy, and credibility of sports punditry and media coverage.

Rugby[1] is one of South Africa's most passionately followed sports, with international test matches commanding nationwide attention and emotional investment [2, 3]. Despite this popularity, much of the public and media discourse surrounding rugby remains driven by narrative, intuition, or historical bias rather than by data-informed reasoning [4]. Analysts and commentators frequently rely on subjective assessments of team strength or form, leading to predictions that are often inconsistent, anecdotal, or unsupported by empirical evidence [4].

Figure 1: The Springboks lift the William Web-Ellis Cup after winning the 2023 Men's Rugby World Cup.

This presents a clear analytical problem: while modern rugby analytics have advanced at the team and coaching level, there remains no generally accessible system for systematically predicting match outcomes at the international level using statistical or machine learning methods [4]. As a result, pre-match coverage in mainstream media tends to prioritise storytelling over statistical inference, reducing opportunities for data-driven insight and objective discussion [4].

To address this, the project develops a **neural network–based AI system** that predicts the outcome of Tier 1 international rugby test matches as one of three classes: *Home Win*, *Home Loss*, or *Draw*. The task is formulated as a **multi-class supervised classification problem**, with inputs derived from historical match data such as teams, competition, venue, and hosting country. The model outputs a probabilistic prediction, offering a quantitative complement to traditional qualitative analysis.

[1]For brevity, this report refers to the sport of "rugby union" simply as *rugby*. Given the limited presence of rugby league in South Africa, this is unlikely to cause confusion.

1.1 Usefulness

The system's usefulness lies in its ability to introduce a more rigorous, data-informed perspective into rugby punditry and reporting. For broadcasters, journalists, and analysts, the model provides an empirical foundation for discussion — grounding predictions in quantifiable evidence rather than opinion. Match outcomes can be predicted on some empirical basis, and inferences can be drawn from historical trends (e.g. the impact of whether a particular fixture takes place during a World Cup or not). This has the potential to elevate the quality of pre- and post-match analysis, enhance audience understanding, and increase trust in media predictions.

Beyond journalism, such a system could support academic or commercial research in sports analytics by enabling comparative evaluation of predictive models across different competitions or eras. While not designed for betting purposes, it is acknowledged that predictive models of this nature may attract attention from the betting community. As such, careful communication and ethical framing are essential to ensure responsible interpretation of results.

1.2 Ethics

From an ethical standpoint, it is critical that the model be communicated and applied responsibly. Predictions should always be expressed probabilistically rather than as certainties, and the system should never be marketed as a betting tool. Developers and users must remain transparent about model limitations, potential biases (e.g., favouring historically dominant teams), and the probabilistic nature of AI-generated forecasts. The model's role is to inform and enhance analytical discourse, not to promote speculation or overconfidence.

The dataset used in this project consists exclusively of publicly available historical Tier 1 international rugby test match results and associated team statistics sourced from Kaggle [1]. Since this data exists entirely in the public domain and pertains to team-level outcomes rather than personal or biometric information, its use presents minimal ethical risk. All information reflects factual, publicly recorded match data, aligning with standards of transparency, reproducibility, and responsible data use.

Importantly, the model is designed to **augment** rather than replace human expertise. Rugby analysts, commentators, and journalists contribute contextual knowledge, intuition, and storytelling that AI systems cannot replicate. The model instead functions as an assistive analytical tool, providing a layer of computational insight

that enhances human judgment. In this sense, it supports existing roles within sports media by automating tasks that are computationally complex for humans, enabling professionals to focus on interpretation, creativity, and narrative depth.

While the possibility remains that individuals may use such systems for betting, the project takes the ethical stance that the primary value of this work lies in advancing data literacy and analytical rigour within rugby media and research communities.

2 Dataset

The dataset used for this study is the *International Rugby Union Results from 1871–2024* dataset, sourced from Kaggle [1]. It contains over 2400 international rugby union test matches, capturing detailed results across major competitions and eras. For this project, we focused specifically on Tier 1 test matches (representing fixtures between top international rugby nations) ensuring consistent data quality and competitive comparability across teams and tournaments.

3 Baseline Model: Naïve Bayes Classifier

3.1 Rationale and Implementation

A Naïve Bayes (NB) classifier was selected as the baseline for the task of predicting outcomes of Tier 1 rugby test matches as (*HomeWin*, *HomeLoss*, or *Draw*). Following the assignment guideline to implement the "simplest possible approach" to the problem [**?**], Naïve Bayes provides a probabilistic foundation that is interpretable, fast to train, and computationally lightweight. It assumes conditional independence between features given the class label, enabling straightforward application to categorical data.

The baseline model was implemented using *scikit-learn*'s *CategoricalNB* class. Six categorical and Boolean predictors were selected: `home_team`, `away_team`, `competition`, `country`, `neutral`, and `world_cup`. These features capture contextual information about the match without incorporating any numerical indicators such as team form, ranking, or score differentials. Boolean variables were cast explicitly as binary indicators, and missing competitions were imputed as "Unknown" to preserve feature completeness. The dataset was split into 80% training and 20% test data, with stratification to maintain the empirical class distribution. A small hyperparameter grid search over Laplace smoothing parameters $\alpha \in \{0.5, 1.0, 2.0\}$ was conducted using five-fold cross-validation.

3.2 Results and Evaluation

The best model achieved a test accuracy of 0.6643. Table 1 summarises the key metrics.

Table 1: Performance of Baseline Naïve Bayes Model (Test Set).

Class	Precision	Recall	F1-score
Draw	0.000	0.000	0.000
HomeLoss	0.643	0.468	0.542
HomeWin	0.672	0.849	0.750
Macro avg	0.439	0.439	0.431
Weighted avg	0.632	0.664	0.637
Overall Accuracy		0.6643	

The confusion matrix (Table 2) reveals that the model is most accurate when predicting *HomeWin* outcomes but struggles to identify *Draws* and *HomeLosses*. Draws are particularly underrepresented and were never predicted, reflecting a common issue in sports modelling where draws constitute a small minority of especially unlikely match outcomes.

Table 2: Confusion Matrix for Baseline Predictions. Rows = True Labels, Columns = Predicted Labels.

	Draw	HomeLoss	HomeWin
Draw	0	8	16
HomeLoss	0	101	115
HomeWin	0	48	269

The Naïve Bayes baseline establishes a performance benchmark of approximately 66% accuracy. While respectable, this figure primarily reflects the model's bias toward predicting the majority class (*HomeWin*), rather than genuine pattern recognition. The zero recall and precision for the *Draw* class confirm the model's difficulty handling class imbalance and the oversimplification inherent in its independence assumption.

Key limitations include:

- **Feature independence:** The assumption that features such as *home_team* and *country* are independent given the outcome is unrealistic, as home advantage strongly interacts with venue and competition.
- **Limited contextual awareness:** The model lacks numerical indicators (e.g., Elo ratings or recent form), constraining predictive depth.
- **Class imbalance:** The rarity of draws biases both training and evaluation, making macro-averaged metrics more reflective of real performance disparities.

Despite these shortcomings, the baseline performs its intended role: providing a transparent and reproducible reference for evaluating the added value of more complex neural models. Future iterations (e.g., multilayer perceptrons or recurrent networks) will test whether richer representations and learned interactions can

exceed this baseline, particularly in minority-class recall and generalisation performance.

4 Model Design

Our neural network-based AI system was implemented as a Feedforward Neural Network (FNN) using the PyTorch framework. A feedforward architecture was chosen because the dataset represents independent rugby test matches rather than sequential data; therefore, temporal dependencies (as in RNNs or LSTMs) are not relevant. The model aims to predict one of three possible match outcomes: *Home Win*, *Home Loss*, or *Draw*.

4.1 Architecture Overview

The final model architecture comprises:

(1) **Input Layer:** A concatenated feature vector combining contextual and categorical features for each match.
(2) **Hidden Layer 1:** 128 neurons using the Rectified Linear Unit (ReLU) activation function.
(3) **Hidden Layer 2:** 64 neurons using ReLU.
(4) **Output Layer:** 3 neurons corresponding to the output classes, followed by a softmax activation to produce class probabilities.

The model uses `CrossEntropyLoss`, which internally applies a softmax to logits and computes the difference between predicted and true labels. ReLU was chosen as it efficiently introduces nonlinearity and mitigates vanishing gradients, allowing the network to model complex interactions between features such as team strength, form, and match context.

4.2 Feature Engineering

Contextual Features:

- `elo_home_pre` and `elo_away_pre`: team strength indicators derived from Elo ratings prior to each match.
- Rolling averages of team form: points, goal differences, and win/draw rates computed over 3-, 5-, and 10-match windows.
- `neutral` flag: indicates whether the match was played at a neutral venue (no home advantage).

Contextual features describe the environment in which the match occurs, rather than the main participants themselves. For instance, home advantage has been statistically shown to influence outcomes through factors such as crowd support, travel fatigue, and environmental familiarity. Including a neutral-site flag helps the network avoid over-attributing success to teams that frequently play at home.

Categorical Features:

- `home_id`: the home team identifier.
- `away_id`: the away team identifier.
- `comp_id`: the competition or tournament identifier.

These categorical variables are embedded as dense vectors through trainable embedding layers, allowing the model to learn latent representations of teams and competitions. This enables the network to capture structural patterns such as:

- Competitive balance in regional tournaments (e.g., Six Nations).
- Strong dominance by top-tier nations in the Rugby Championship.
- Increased variability in friendly tests and mismatches in World Cups.

For example, embedding `comp_id` allows the model to understand that a result in the Six Nations differs statistically from a World Cup fixture, reflecting competition-specific dynamics. These embeddings are concatenated with continuous contextual features to produce the final match-level feature vector.

4.3 Regularisation

To prevent overfitting, the model employs:

- **Dropout:** Randomly deactivates neurons during training (rate = 0.2–0.3) to encourage redundancy and robustness.
- **Weight decay:** Penalises large weights to prevent memorisation of training data.

These mechanisms ensure the model generalises to unseen matches rather than simply reproducing historical outcomes.

5 Model Validation

Hyperparameters were selected by monitoring both validation and test accuracy across multiple training configurations. Each configuration was trained for 100 epochs using the Adam optimiser and evaluated on separate validation and test splits. The experiments were structured into phases, progressively adjusting parameters to improve generalisation and class balance.

5.1 Phase 1: Baseline Neural Model

Configuration: Normal CrossEntropyLoss; learning rate = 0.001; weight decay = 0.0004; dropout = 0.2; hidden layers (128, 64, 3); team embeddings $\{8, 16, 32\}$; competition embeddings $\{4, 8, 16\}$. This setup achieved a **validation accuracy of 75.4%** and a **test accuracy of 72%**. However, the model failed to identify the *Draw* class (precision and recall = 0.00), demonstrating the effects of strong class imbalance.

5.2 Phase 2: Weighted-Class Cross Entropy

To mitigate class imbalance, the loss function was modified to assign higher weights to rare classes (particularly *Draws*). While this encouraged the network to predict more draws, overall performance declined: **validation accuracy dropped to 75%** and **test accuracy to 69%**. The model still produced no true positives for draws, suggesting that penalising misclassification alone was insufficient due to the extreme rarity of draw outcomes.

5.3 Phase 3: Adjusted Regularisation Parameters

Here, dropout was increased to 0.3 and weight decay reduced to 0.001 to enhance generalisation. This yielded a **validation accuracy of 75%** and a **test accuracy of 72%**, marginally improving model stability but not significantly enhancing recall for minority classes.

5.4 Phase 4: Reduced Learning Rate

Lowering the learning rate to 0.0005 (with dropout 0.3 and weight decay 0.001) aimed to smooth convergence and prevent overfitting. The model achieved a **validation accuracy of 76%** and a **test accuracy of 73.5%**, improving training stability and slightly narrowing the generalisation gap.

5.5 Phase 5: Increased Model Capacity

Expanding model capacity (256 and 128 neurons in the two hidden layers) increased representation power. With dropout = 0.3, learning rate = 0.0005, and weight decay = 0.001, the model achieved a **validation accuracy of 78%** and a **test accuracy of 74%**. A reduced-capacity variant (64 and 32 neurons) achieved 76% validation and 72% test accuracy, confirming that a larger network slightly enhances performance without overfitting.

6 Evaluation and Results

The final model achieved a **validation accuracy of 75%** and a **test accuracy of 72%**, outperforming the Naïve Bayes baseline (66% accuracy). The neural model effectively learned to distinguish between wins and losses but continued to struggle with predicting draws due to their low frequency (4.35% of all matches, or 121 out of 2783). The confusion matrix below summarises model behaviour across classes:

Table 3: Confusion Matrix of Neural Network Predictions (rows = true labels, columns = predicted labels).

	Draw	HomeLoss	HomeWin
HomeLoss	0	98	64
Draw	0	3	8
HomeWin	0	41	204

Precision and recall per class are shown in Table 4.

Table 4: Classification Metrics for Neural Network Model.

Class	Precision	Recall	F1-score
HomeLoss	0.690	0.605	0.645
Draw	0.000	0.000	0.000
HomeWin	0.739	0.833	0.783
Macro avg	0.476	0.479	0.476
Weighted avg	0.701	0.723	0.709
Overall Accuracy		0.7225	

The model converged steadily over 100 epochs, with validation accuracy improving from 0.61 in epoch 1 to a peak of 0.77 by epoch 11, before stabilising between 0.72–0.75. This indicates successful learning without overfitting.

7 Analysis of Performance

The neural model improves upon the Naïve Bayes baseline by approximately 6% in overall accuracy and achieves notably higher recall for wins and losses. Its ability to generalise from contextual and embedding features demonstrates that combining continuous and categorical information enhances predictive power compared to count-based baselines.

However, the persistent inability to predict draws reveals two key insights:

(1) **Data imbalance:** The rarity of draws (less than 5% of matches) limits the model's exposure to positive examples. Standard cross-entropy loss functions assume roughly balanced classes, causing the model to under-optimise minority outcomes.
(2) **Predictive saturation:** Rugby results are strongly deterministic; elite teams rarely draw, and the probabilistic space between win and loss is narrow. Thus, even an optimal classifier may find draws statistically unpredictable from pre-match features alone.

Despite this, the model demonstrates effective representation learning: validation and test accuracy remain stable across epochs, and the generalisation gap never exceeds 3–4%. Compared to the baseline, the neural model captures nuanced contextual relationships (e.g., team embeddings and form indicators), supporting the conclusion that neural architectures meaningfully enhance predictive accuracy while maintaining interpretability through feature inspection.

References

[1] Lyle Begbie. 2022. International Rugby Union Results from 1871–2022. https://www.kaggle.com/datasets/lylebegbie/international-rugby-union-results-from-18712022. Accessed October 2025, Kaggle Dataset.
[2] Simon Borchardt. 2025. *Springbok Women break broadcast records.* https://www.sarugby.co.za/news-features/articles/2025/10/09/springbok-women-break-broadcast-records/ SA Rugby News Features – record viewership for Springbok Women.
[3] Thinus Ferreira. 2024. *South African rugby continues to drive TV ratings.* https://teeveetee.blogspot.com/2024/11/south-african-rugby-continues-to-drive.html Blog post on TVwithThinus – News & analysis on South African television.
[4] Ben Jermy. 2022. How Our Kick Predictor Is Enhancing Rugby Storytelling. https://www.statsperform.com/resource/how-our-new-kick-predictor-is-enhancing-rugby-storytelling/. Stats Perform – Case study on rugby broadcast analytics.