Optimal Machine Learning Models for T20 Cricket: The Role of Dangerous Balls in Match Outcomes

Abdul Majid[1], Qamruz Zaman[1*], Ghazala Sahib[2], Soofia Iftikhar[2], Sundas Hussain[2], Najma Salahuddin[2]

[1]*Department of Statistics, University of Peshawar, Pakistan*
[2]*Department of Statistics, Shaheed Benazir Bhutto Women University Peshawar (SBBWUP), Pakistan*
Email Address: cricsportsresearchgroup@gmail.com

Abstract:
This study explores the application of machine learning models Logistic Regression, Multilayer Perceptron (MLP), and Decision Tree (CRT method) to predict the outcome of T20 International cricket matches based on dangerous deliveries consisting on two key independent variables: wickets lost and extras conceded. The dataset, comprising 2,492 matches, was analyzed to understand the impact of these variables on match results. By comparing the performance of these models across the first and second innings of matches, the study aims to identify which model best captures the dynamics of match outcomes. The models were evaluated in terms of their predictive accuracy, interpretability, and the significance of the variables, with a particular focus on the role of wickets in determining match results. Although all three models proved effective in predicting match outcomes, the Decision Tree model stood out as the most reliable and comprehensible, providing meaningful insights into the connection between match dynamics and results. The findings highlight the potential of machine learning techniques in sports analytics, offering valuable insights for both researchers and cricket analysts in forecasting match outcomes and understanding the factors that influence a team's success in T20 cricket.

Keywords: Match outcome prediction, machine learning, Logistic Regression, Multilayer Perceptron, Decision Tree,

1. Introduction

Cricket, a bat-and-ball game that originated in England, has a rich history spanning several centuries. The earliest recorded mention of cricket dates back to the late 16th century, with the first reference in 1598 in the southeastern counties of England (Alston, 2024). The sport was initially informal, with various local rules, but by the 18th century, it had become more organized, with established rules and the foundation of major cricket clubs, such as the Marylebone Cricket Club (MCC) in 1787, which is still regarded as the guardian of the laws of the game (Sharma, 2018). The 20th century brought significant changes, including the introduction of limited-overs formats. The first One Day International (ODI) match was played in 1971, followed by the inaugural Cricket World Cup in 1975, which introduced new dynamics to the game (Gillespie, 2013). The late 20th century also saw the rise of Twenty20 (T20) cricket, a fast-paced, shorter format that gained immense popularity and led to the creation of global tournaments like the Indian Premier League (IPL), further transforming the sport's landscape (Ray, 2022). The T-20 International format has revolutionized cricket by introducing a fast-paced environment where each delivery can significantly impact the match outcome. In this condensed format, factors such as wickets lost and extras conceded (e.g., wides and no-balls) play a decisive role in determining a team’s success. These two factors, collectively termed as "dangerous balls", disrupt batting momentum and provide undue scoring opportunities to the opponent, thereby influencing match

results in subtle yet measurable ways. With the increasing availability of cricket data, machine learning (ML) techniques have emerged as powerful tools to analyze complex game dynamics and predict outcomes. ML algorithms, including Logistic Regression, Multi-Layer Perceptron (MLP), and Decision Trees, are well-suited for identifying patterns in large-scale datasets and developing predictive models for performance assessment. For instance, studies have applied decision trees and neural networks to predict match outcomes in cricket, demonstrating the efficacy of these methods in sports analytics (Kumar, J et.al, 2018)

Recent studies have demonstrated the application of machine learning in sports analytics, including performance prediction, player evaluation, and strategic decision-making. For example, research has explored the use of machine learning algorithms to predict match outcomes in cricket, highlighting the potential of these techniques in enhancing decision-making processes (Shenoy et.al, 2022). Cricket is a dynamic sport that involves numerous variables influencing match outcomes, particularly in the fast-paced T-20 format. Among these variables, wickets lost and extras conceded are crucial factors that can substantially alter the trajectory of a match. Recent literature has explored various dimensions of cricket analytics, applying statistical methods and machine learning models to predict match outcomes, evaluate player performance, and understand the underlying factors affecting match results.

This study leverages these methods to, Model the relationship between dangerous balls (wickets and extras) and match outcomes, Evaluate the predictive accuracy of ML techniques in determining match results and Provide actionable insights to assist teams in mitigating the occurrence of dangerous balls.

By applying ML algorithms to T-20 cricket, this research contributes to the growing body of cricket analytics, offering a data-driven approach to optimize team performance and improve match preparation. The aim of this study is to investigate the impact of dangerous deliveries, specifically wicket-taking balls and extras conceded, on the outcomes of T20 cricket matches using three machine learning models: Logistic Regression, Multilayer Perceptron (MLP), and Decision Tree.

The novelty of this research is uniquely application of a machine learning approach to analyze the influence of "dangerous balls" a concept encompassing wicket-taking and extras deliveries on match outcomes in T20 cricket. Unlike traditional cricket analytics that primarily focus on aggregate statistics or player performance, this study delves into delivery-level impact using advanced predictive models. The inclusion of multiple machine learning techniques allows for a comparative evaluation, offering deeper insights into which models are most effective in capturing these critical match dynamics.

The prediction of match outcomes in sports, particularly cricket, has attracted substantial interest due to its potential in improving team strategies, performance assessments, and fan engagement. Various statistical and machine learning models have been applied to understand the key determinants influencing match results. One of the crucial aspects of cricket that influences the outcome is the performance metrics of the players, such as the number of wickets taken and the number of extras conceded. Logistic regression is a widely used statistical method in sports analytics for binary outcome prediction (e.g., win or lose). It has been effectively applied in cricket for modeling match results based on different predictors, including player performance, team composition, and match conditions. According to Saha and Ghosh (2020), logistic regression has proven to be a reliable model for predicting the outcome of cricket matches, especially when considering factors like runs scored, wickets taken, and extras conceded. The simplicity and interpretability of logistic regression make it a common choice for analyzing binary outcomes in sports analytics (Bunker & Thabtah, 2020). The Multilayer Perceptron (MLP), a type of artificial neural network, has gained popularity in sports prediction due to its ability to capture complex non-linear relationships between variables. Liu et al. (2019) applied MLP in the context of sports prediction and found that it outperforms traditional statistical models by modeling interactions between multiple input variables. The flexibility of MLP in handling large datasets and intricate patterns makes it a suitable candidate for analyzing cricket match outcomes, where multiple interacting factors—such as the number of wickets and extras—can affect the result. Ali & Ghosh (2021) also noted that MLP models are able to adapt to non-linearities in the data, improving the accuracy of predictions in cricket. Decision tree models, particularly Classification and Regression Trees (CART), have also been widely used for predicting match outcomes. These models divide the data into subsets based on certain decision rules, making them interpretable and easy to visualize. In cricket, decision trees have been employed to predict match outcomes by analyzing variables such as batting performance, bowling performance, and match conditions. Chouhan et al. (2018) highlighted the application of decision tree models in sports prediction, noting their effectiveness in capturing both the major contributing factors (like wickets and extras) and interactions between those factors. Patil & Shah (2022) also demonstrated

the utility of decision trees in cricket by showing how different conditions, such as weather or team form, impact the final match result. These models are particularly valuable in providing transparent decision-making processes, which can be crucial for strategic planning.

The number of wickets and extras are significant predictors in cricket match outcome models. Rathore & Mehta (2021) discussed how wickets are a key indicator of team performance, with the loss of wickets often leading to a shift in momentum, thereby affecting the overall match outcome. Similarly, extras runs awarded to the batting team through no-balls, wides, and byes have a direct impact on the total score and can influence the course of a match. Singh & Kapoor (2019) showed that both wickets and extras are important independent variables in match prediction, with the loss of wickets leading to a decline in the probability of winning, while a high number of extras can significantly increase the total score, potentially influencing the result. The application of logistic regression, MLP, and decision trees in predicting match outcomes has been subject to comparison in various studies. Singh et al. (2020) compared different machine learning algorithms, including decision trees, MLP, and logistic regression, for predicting outcomes in cricket and found that decision trees performed the best in terms of interpretability and accuracy. However, Jain & Patel (2021) argued that MLP models, with their capacity to handle complex data patterns, provided superior prediction accuracy in scenarios where multiple predictors interact non-linearly.

Despite these individual strengths, each model has its limitations and advantages depending on the dataset used and the complexity of the relationships between the predictors. The combination of these models can offer a more comprehensive understanding of match dynamics, and their effectiveness is further enhanced when combined with other performance metrics such as batting and bowling averages.

2. Methods and Materials

2.1 Data Collection:

The dataset for this research was collected from two reliable sources: Cricinfo and Cricbuzz. A total of 2,492 T20 International cricket matches were included to ensure a comprehensive analysis of match outcomes. The focus was placed on two independent variable wickets lost and extras conceded which are hypothesized to significantly influence match results. To achieve the objective of the study this study applied Logistic regression, Multilayer perceptron and decision trees

2.2 Logistic Regression

Logistic regression is a statistical model used for predicting the probability of a binary outcome based on one or more independent variables. In the context of this research, logistic regression was employed to model the probability of a cricket match's outcome (win/loss) based on independent variables, specifically the number of wickets and extras. Logistic regression is particularly suited for binary classification problems because it produces a probability score between 0 and 1, which can be interpreted as the likelihood of the event occurring (Hosmer et.al. 2013).

2.2.1 Model Formulation

The logistic regression model can be expressed as:

$$\text{Logit}(p) = \ln(p/1-p) = \beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n \quad (1)$$

Where:

- p is the probability of the match outcome (e.g., win),
- $X_1, X_2, \ldots, X_n$ represent the independent variables (wickets and extras),
- B_0 is the intercept,
- $B_1, \beta_2, \ldots, \beta_n$ are the coefficients of the independent variables.

In this study, the independent variables included:

- Wickets (X_1): The number of wickets taken by the team,
- Extras (X_2): The total extras conceded by the bowling team.

The dependent variable was the match outcome, coded as 1 for a win and 0 for a loss.

2.3 Multilayer Perceptron (MLP)

A Multilayer Perceptron (MLP) is a type of artificial neural network (ANN) consisting of multiple layers of neurons, including an input layer, one or more hidden layers, and an output layer. It is a powerful model used to capture complex, non-linear relationships between input features and output predictions (Hecht-Nielsen, 1989). In this research, MLP was used to predict the outcome of cricket matches based on the independent variables: wickets and extras. MLP is particularly useful for modeling intricate

patterns in data, where traditional linear models like logistic regression may fall short. The MLP model consists of the following components:

- Input Layer: The independent variables, namely the number of wickets and the number of extras, are fed into the input layer.
- Hidden Layers: The network may include one or more hidden layers where each neuron in a layer is connected to every neuron in the previous and subsequent layers. The hidden layers enable the model to learn non-linear patterns in the data (Rumelhart, Hinton, & Williams, 1986).
- Output Layer: The output layer consists of a single neuron that predicts the binary outcome (win or loss) of the match.

The activation function used for the neurons in the hidden layers is typically the ReLU (Rectified Linear Unit) function, which helps prevent issues like vanishing gradients and speeds up learning (Glorot et al., 2011). The output layer uses a sigmoid activation function, which maps the output to a probability value between 0 and 1, representing the likelihood of a match outcome (win).

The model can be represented as:

$y=\sigma\left(\sum w_i \cdot x_i + b\right)$ ---------- (2)

- y is the predicted probability,
- x_i are the input features (wickets and extras),
- w_i are the weights of the connections,
- b is the bias term,
- σ is the sigmoid activation function.

2.3.1 Training the MLP Model

The MLP model was trained using the backpropagation algorithm (Rumelhart et al., 1986). Backpropagation adjusts the weights of the network by minimizing the loss function, typically the binary cross-entropy loss function for binary classification problems:

$L = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(\hat{y}_i\right) + (1-y_i)\log\left(1-\hat{y}_i\right)\right]$ ---------- (3)

- y_i is the actual outcome (1 for win, 0 for loss),
- $\hat{y}$ is the predicted probability of the match outcome,
- N is the number of data points.

The model's parameters (weights and biases) are updated iteratively using gradient descent or its variants, such as Stochastic Gradient Descent (SGD) or Adam Optimizer, which help improve convergence speed and performance (Kingma & Ba, 2014).

2.4 Decision Tree Model:

A Decision Tree is a widely used machine learning algorithm for classification tasks. It models decisions and their possible consequences, including chance event outcomes, resource costs, and utility. The tree structure consists of nodes representing decisions or outcomes, and branches representing possible decisions or consequences. The goal of a decision tree is to split the data at each node based on the most significant feature, ultimately classifying the data at the leaf nodes (Breiman et al., 1986). In this research, we used the SPSS Decision Tree algorithm, which implements the CART (Classification and Regression Trees) methodology, to predict the outcome of a cricket match based on two independent variables: wickets and extras.

3. Results and Discussion:

3.1 Logistic Regression

Table No. 1: Model Summary

	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1st Innings	3064.594	.145	.193
2nd Innings	1486.992	.546	.728

Table No.1 reveals that the logistic regression performs significantly better for the 2nd innings than the 1st. For the 1st innings, the -2 Log Likelihood value is high (3064.594), with Cox & Snell R^2 (0.145) and Nagelkerke R^2 (0.193) indicating that only a modest proportion of the variance in match outcomes is explained by the predictors (extras and wickets). In contrast, the 2nd innings model shows a much better fit, with a lower -2 Log Likelihood (1486.992) and higher Cox & Snell R^2 (0.546) and Nagelkerke R^2 (0.728), explaining a substantial 72.8% of the variance. This suggests that extras and wickets have a

more decisive impact in the 2nd innings, likely due to the clearer objective of chasing or defending a target, whereas the 1st innings is influenced by additional unmeasured factors.

Table NO. 2: Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp (B)
1st Innings	Wickets	-.384	.022	302.195	1	.000	.681
	Extras	.040	.009	22.410	1	.000	1.041
	Constant	2.257	.178	159.941	1	.000	9.558
2nd Innings	Wickets	-.990	.037	702.991	1	.000	.372
	Extras	.000	.013	.001	1	.097	1.099
	Constant	6.133	.260	554.758	1	.000	460.777

Table No.2 exposes that In the 1st innings, the logistic regression results indicate that “wickets” have a significant negative effect on match outcomes, with a coefficient of -0.384 ($p < 0.001$), meaning that for each additional wicket lost, the odds of winning decrease by 32% (Exp (B) = 0.681). The “extras” variable has a positive but modest effect, with a coefficient of 0.040 ($p < 0.001$), indicating that each additional extra increases the odds of winning by 4% (Exp (B) = 1.041). The “constant” value of 2.257 ($p < 0.001$) suggests that when both wickets and extras are at zero, the odds of winning are 9.558 times higher. In the 2nd innings, “wickets” have a much stronger negative effect, with a coefficient of -0.990 ($p < 0.001$), meaning that each additional wicket lost reduces the odds of winning by 63% (Exp(B) = 0.372). However, “extras” are not a significant predictor ($p = 0.097$), with a coefficient close to zero (0.000), implying that extras have little impact on the outcome in the 2nd innings. The “constant” value of 6.133 ($p < 0.001$) suggests a strong baseline likelihood of winning when wickets and extras are zero, with the odds being 460.777 times higher. Overall, the results emphasize the critical role of “wickets” in both innings, but the 2nd innings shows a more pronounced effect, while “extras” play a less significant role in the 2nd innings.

Table No.3: Classification Table (Logistic Regression)

Observed			Predicted		
			Match Outcome		
			Loss	Win	Percentage Correct
1st Innings	Match Outcome	Loss	826	427	65.9
		Win	433	806	65.1
	Overall Percentage				65.5
2nd Innings	Match Outcome	Loss	1049	190	84.7
		Win	121	1132	90.3
	Overall Percentage				87.5

The classification table no.3 reveals that the logistic regression model performs better in predicting match outcomes in the 2nd innings compared to the 1st innings. In the “1st innings”, the model correctly predicted 65.5% of the outcomes, with a moderate accuracy of 65.9% for losses and 65.1% for wins, indicating some misclassification. In contrast, the “2nd innings” model achieved a significantly higher overall accuracy of 87.5%, correctly predicting 84.7% of losses and 90.3% of wins, suggesting that the model is much more effective in this phase of the match. This improvement in the 2nd innings reflects the more decisive nature of the outcomes, where the match context is clearer.

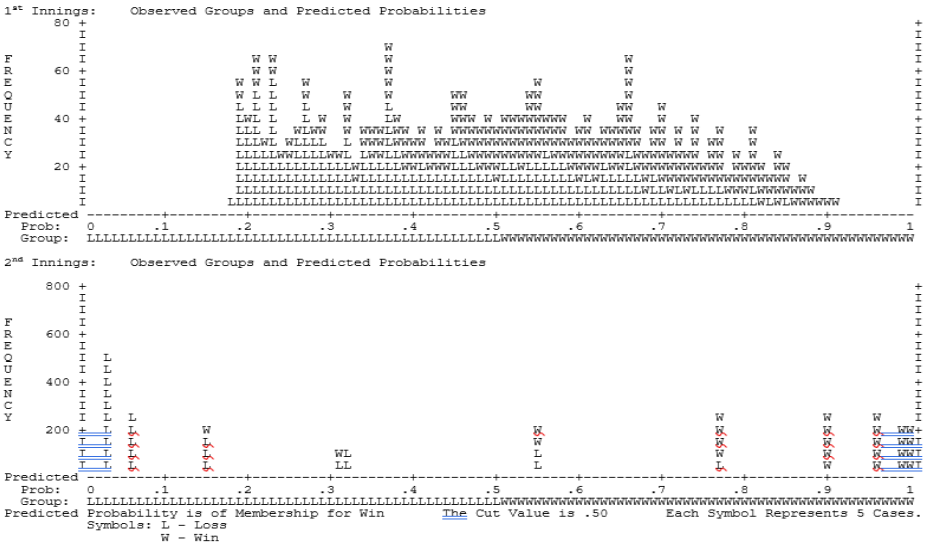

Figure No.1 Inning wise predicted probabilities plot

Figure No.1 The predicted probabilities plot for both the 1st and 2nd innings shows the distribution of predicted probabilities for match outcomes (win vs. loss) based on the logistic regression model. In the 1st innings, the predicted probabilities range from 0 to 1, with most of the data points clustering around lower probabilities for losses (L) and higher probabilities for wins (W). This suggests that the model predicts match outcomes with varying certainty, with some observations having relatively high probabilities of winning or losing. However, there is a mix of predicted values near 0.5, indicating some uncertainty in the model's predictions for certain matches. In contrast, the 2nd innings plot reveals a more pronounced distinction between losses (L) and wins (W), with a higher concentration of predicted probabilities near 0 or 1. This indicates a clearer distinction between the two outcomes, with the model more confidently predicting wins and losses in the 2nd innings. The concentration of symbols at both extremes suggests that the 2nd innings model has a stronger discriminatory power, making it more reliable in predicting the match outcome based on the provided predictors (wickets and extras). The cut-off value for classification is set at 0.50, meaning predictions above this threshold are classified as wins and those below as losses.

3.2 Multilayer Perceptron:

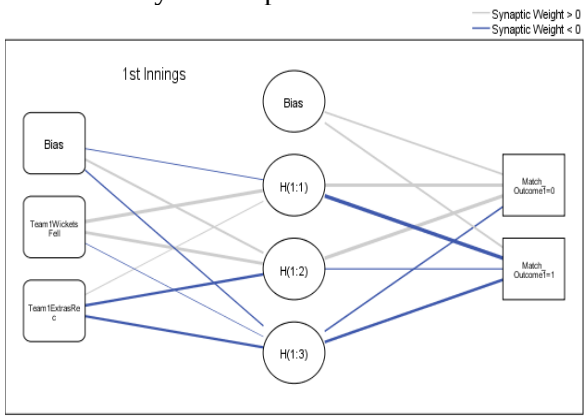

Figure No.2 1st Innings MLP Network

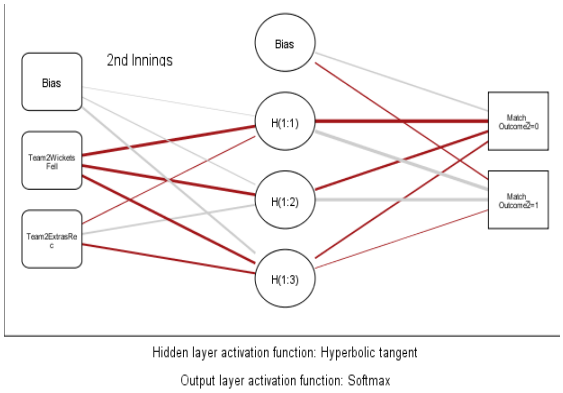

Figure No.3: 2nd Innings MLP Network

Figure No.2 and Figure No.3 illustrate the Multilayer Perceptron (MLP) network architectures for the 1st and 2nd innings, detailing the connections between input variables, hidden layer nodes, and output categories. The MLP network for the 1st innings comprises two input variables: Team1 Wickets Fell and Team1 Extras Received, along with a bias node. These inputs are connected to three hidden layer nodes (H(1:1), H(1:2), and H(1:3)) via synaptic weights. The activation function in the hidden layer is the hyperbolic tangent, which introduces non-linearity to the model by allowing it to capture complex relationships between the inputs and the outputs. The synaptic weights are represented with different colors to indicate polarity blue for positive weights and gray for negative weights. The hidden layer nodes subsequently connect to the output layer, which categorizes the match outcome as either "Loss" or "Win." The output layer uses the Softmax activation function, which ensures the predictions are probabilities that sum to 1. Notably, Team1 Wickets Fell seems to contribute more significantly to the network, as evidenced by the stronger connections to the hidden nodes. In contrast, Team1 Extras Received appears to have a weaker impact, with lighter connection weights.

The MLP network for the 2nd innings exhibits a similar structure, with input variables: Team2 Wickets Fell and Team2 Extras Received, along with a bias node. The input nodes are linked to the three hidden layer nodes (H(1:1), H(1:2), and H(1:3)) through weighted synaptic connections. Again, the hidden layer employs the hyperbolic tangent activation function to introduce non-linearity. Here, the synaptic weights are represented with red for positive weights and gray for negative weights, showing the direction of influence on the hidden nodes. The outputs, categorized as "Loss" or "Win," are determined using the Softmax activation function in the output layer.In the 2nd innings, Team2 Wickets Fell plays a dominant role in influencing the hidden layer, as indicated by the much stronger and more consistent weights. On the other hand, Team2 Extras Received has minimal impact, with mostly weak or negligible weights.

For both innings, wickets (Team1 Wickets Fell and Team2 Wickets Fell) play a more significant role in determining match outcomes than extras received. The connections in the 2nd innings are stronger

overall, indicating that the MLP model captures clearer relationships between inputs and outcomes for the 2nd innings compared to the 1st innings.The bias node contributes to ensuring the network can adapt to any inherent shifts in the data and improves model flexibility for both innings.

Table No.4: Model Summary (MLP)

		1st Innings	2nd Innings
Training	Cross Entropy Error	1065.080	537.640
	Percent Incorrect Predictions	34.4%	12.8%
	Stopping Rule Used	1 consecutive step(s) with no decrease in error[a]	1 consecutive step(s) with no decrease in error[a]
	Training Time	0:00:00.09	0:00:00.12
Testing	Cross Entropy Error	471.177	205.006
	Percent Incorrect Predictions	34.8%	11.9%

Table No.4 the MLP (Multilayer Perceptron) model summary shows that it performs quite well in predicting match outcomes, with clear differences between the 1st and 2nd innings. In the "1st innings", the model has a training cross-entropy error of 1065.080, which reflects the error during the training phase. The model has 34.4% incorrect predictions during training, indicating a moderate level of misclassification. The testing phase shows a slightly improved cross-entropy error of 471.177, with 34.8% incorrect predictions, suggesting a similar performance on unseen data. The training time for both innings is very quick, indicating efficient learning. For the 2nd innings, the model shows better performance, with a training cross-entropy error of 537.640 and 12.8% incorrect predictions during training. The testing phase shows a further improvement with a cross-entropy error of 205.006 and 11.9% incorrect predictions, indicating that the model generalizes well to new data and is highly accurate in predicting the outcome of the 2nd innings. Overall, the MLP model shows better performance in the 2nd innings, with fewer incorrect predictions and lower error rates compared to the 1st innings.

Table No.5: Classification (MLP)

		Predicted		
1st Innings	Observed	Loss	Win	Percent Correct
Training	Loss	573	291	66.3%
	Win	303	558	64.8%
	Overall Percent	50.8%	49.2%	65.6%
Testing	Loss	252	137	64.8%
	Win	130	248	65.6%
	Overall Percent	49.8%	50.2%	65.2%
		Predicted		
2nd Innings	Observed	Loss	Win	Percent Correct
Training	Loss	741	141	84.0%
	Win	85	805	90.4%
	Overall Percent	46.6%	53.4%	87.2%
Testing	Loss	309	48	86.6%
	Win	38	325	89.5%
	Overall Percent	48.2%	51.8%	88.1%

Dependent Variable: Match Outcome

Table No.5 The classification table for the MLP (Multilayer Perceptron) model shows stronger performance in the 2nd innings compared to the 1st innings. In the 1st innings, the model achieved 66.3% accuracy for predicting losses and 64.8% accuracy for predicting wins during the training phase, resulting in an overall accuracy of 65.6%. During testing, the accuracy dropped slightly to 64.8% for losses and 65.6% for wins, giving an overall accuracy of 65.2%. This indicates that the model has moderate predictive ability, but there is some room for improvement. In contrast, the 2nd innings model performed significantly better, with 84.0% accuracy for predicting losses and 90.4% accuracy for predicting wins during the training phase, leading to an overall accuracy of 87.2%. The testing phase also showed high performance, with 86.6% accuracy for losses and 89.5% for wins, resulting in an overall accuracy of 88.1%. These results demonstrate that the MLP model is much more effective at predicting match outcomes in the 2nd innings, where the model shows consistent and high accuracy both during training and testing. This suggests that the 2nd innings is a more predictable phase of the

match, and the model can better capture the relationships between the predictors (wickets and extras) and match outcomes in this context.

3.3 Decision Tree:

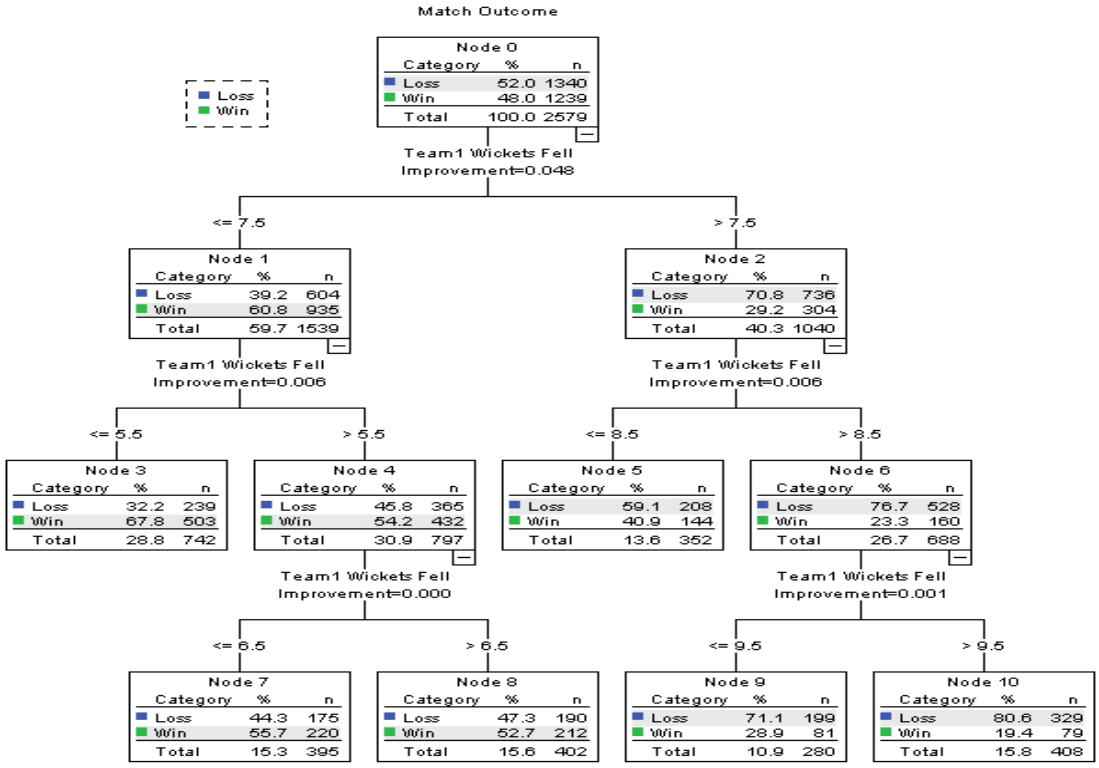

Figure No.4: Classification Tree for 1st Innings Wickets loss

Figure No.4 The decision tree demonstrates the relationship between Team1 Wickets Fell and match outcomes (win or loss) for the 1st innings. At the root node, the dataset contains 2,579 matches, with 62.0% resulting in a loss and 48.0% in a win. The first and most significant split occurs at the threshold of 7.5 wickets, dividing the matches into two branches. Matches where the team lost 7.5 or fewer wickets lead to Node 1, while matches with more than 7.5 wickets lost proceed to Node 2. This split highlights that losing fewer than 7.5 wickets is associated with a higher likelihood of winning. Node 1, representing matches with ≤7.5 wickets lost, contains 1,539 observations, where 60.8% resulted in a win and only 39.2% in a loss. This subset is further split at 5.5 wickets. Teams losing ≤5.5 wickets (Node 3) have the highest probability of winning, with 67.8% of matches resulting in a win, demonstrating the critical importance of preserving wickets. Conversely, teams losing between 5.5 and 7.5 wickets (Node 4) see their win percentage decrease slightly to 54.2%, though they still maintain a higher likelihood of success compared to teams losing more than 7.5 wickets. Node 2, representing matches where more than 7.5 wickets were lost, contains 1,040 matches, with 70.8% resulting in a loss and only 29.2% in a win. This branch is further split at 8.5 wickets. Matches with ≤8.5 wickets lost (Node 5) show a win probability of 40.9%, while those with >8.5 wickets lost (Node 6) demonstrate a much lower likelihood of success, with only 23.3% of matches resulting in a win. As teams exceed 8.5 wickets lost, their chances of winning diminish significantly. The tree further refines the predictions through smaller thresholds, such as 6.5 and 9.5 wickets, leading to the final leaf nodes. For example, Node 10, which represents matches where teams lost more than 9.5 wickets, shows that 80.6% of matches resulted in a loss, highlighting the detrimental effect of losing a high number of wickets.

The decision tree emphasizes the critical impact of wickets lost on match outcomes. Preserving wickets is key to increasing the likelihood of success, as teams that lose fewer than 5.5 wickets have the best winning chances (67.8%), while teams losing more than 7.5 wickets are more likely to lose the match (70.8%). Losing more than 8.5 wickets further exacerbates the chances of defeat, underscoring the importance of maintaining wickets throughout the innings.

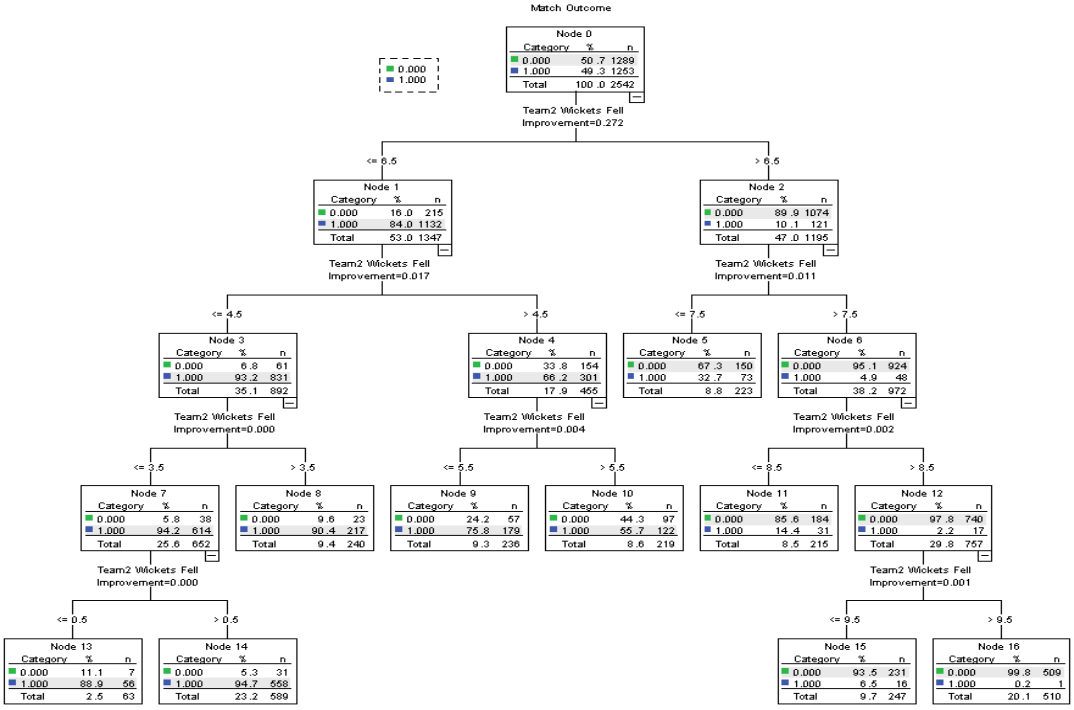

Figure No.5: Classification Tree for 2nd Innings wickets loss

Figure No.5 The decision tree illustrates how the number of wickets fallen for Team 2 impacts the match outcome. At the root level, the data shows an almost equal distribution of match outcomes, with 50.7% losses and 49.3% wins. The first major split occurs at "Team 2 Wickets Fell ≤ 6.5." Matches where Team 2 loses 6.5 or fewer wickets are strongly associated with winning outcomes, showing an 84.1% chance of a win. In contrast, when more than 6.5 wickets fall, the chances of losing rise sharply to 89.9%. Further refinement reveals that within the "≤ 6.5 wickets" group, outcomes improve even more when wickets lost are limited to 4.5 or fewer, with a staggering 93.2% chance of a win. However, when the number of wickets falls between 4.5 and 6.5, the win probability decreases to 66.2%. On the other hand, in the "> 6.5 wickets" group, the situation becomes increasingly dire. If wickets lost are between 6.5 and 7.5, losses still dominate at 67.3%, but there is a modest 32.7% chance of a win. Once Team 2 loses more than 7.5 wickets, the likelihood of a loss soars to 96.1%, leaving a minimal chance of winning. The overarching trend is clear: as the number of wickets fallen for Team 2 increases, the probability of losing the match rises significantly. Matches with fewer than 6.5 wickets fallen offer the best chances of success, with performance improving further as the number of wickets decreases below 4.5. Conversely, once the threshold of 6.5 wickets is crossed, the likelihood of a loss becomes overwhelming, especially beyond 7.5 wickets. This emphasizes the critical importance of preserving Team 2's wickets to improve match outcomes.

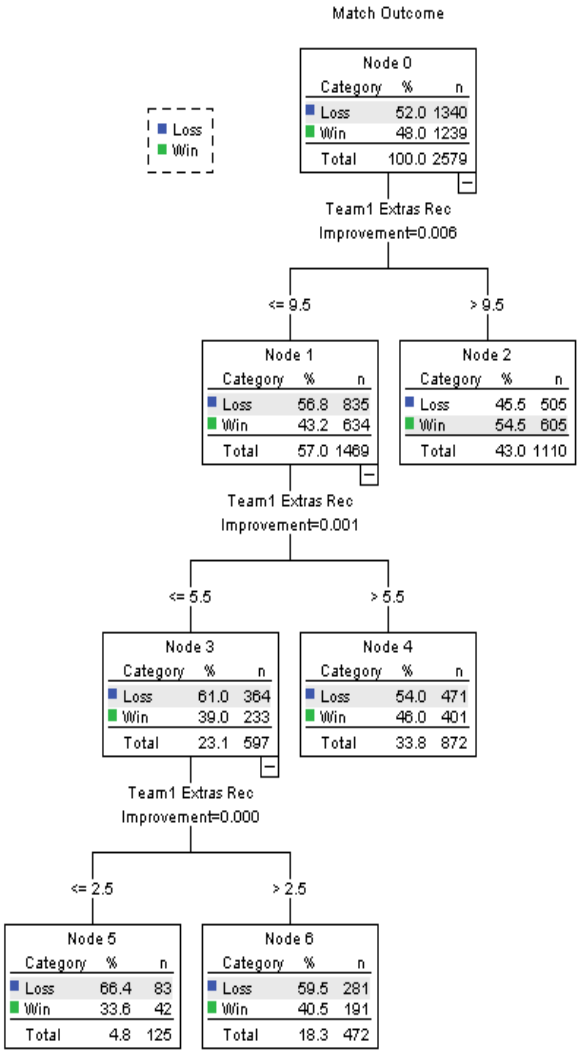

Figure No.6: Decision Tree 1st Innings Extras

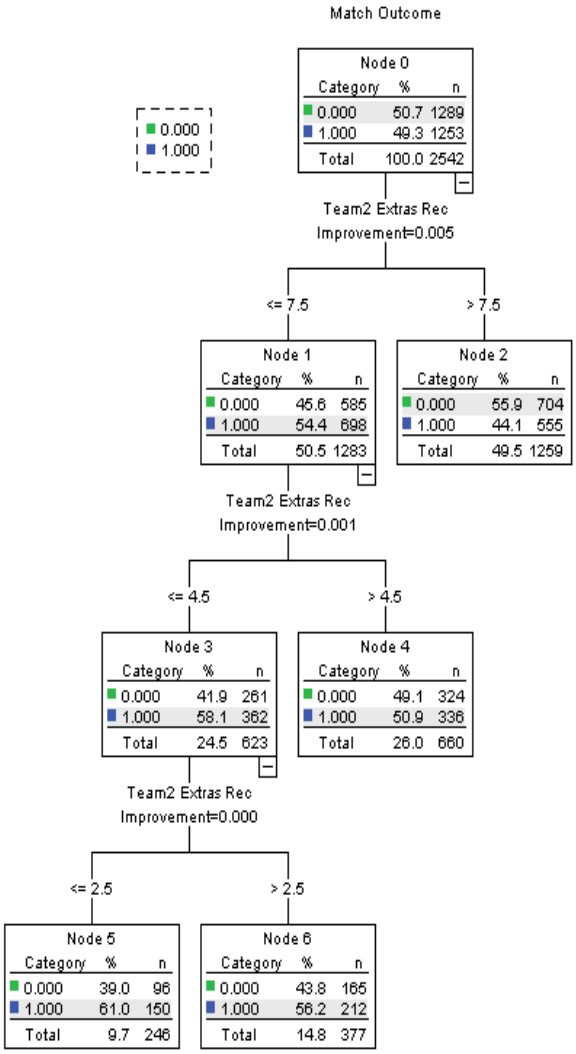

Figure No.7: Decision Tree 2nd Innings Extras

Figure No.6 and Figure No.7 for 1st and 2nd Innings respectively showed that, In the first innings (Team 1 Extras), conceding fewer extras (≤ 9.5) is associated with a higher likelihood of losing (66.8% losses), while conceding more than 9.5 extras slightly improves the win probability to 54.5%. Among cases with very few extras (≤ 5.5), losses dominate further, with only a 39% chance of winning. This trend is more pronounced when extras are limited to ≤ 2.5, where losses peak at 66.4%, suggesting that excessively restrictive bowling in the first innings may correlate with defensive gameplay, leading to adverse outcomes.

In the second innings (Team 2 Extras), limiting extras to ≤ 7.5 increases the probability of winning to 54.4%, compared to a 45.6% loss probability. The winning chances improve further to 58.1% when extras are kept below 4.5, and even more so (60.1%) with ≤ 2.5 extras. Conversely, when Team 2 concedes more than 7.5 extras, the likelihood of losing rises to 56.9%, emphasizing that disciplined bowling in the second innings plays a critical role in securing a favorable match outcome.

Table No. 6: Classification (Decision Tree)

Observed		Predicted		
		Loss	Win	Percent Correct
1st Innings	Loss	685	568	54.7%
	Win	270	969	78.2%
	Overall Percentage	38.3%	61.7%	66.4%
Observed		Predicted		
		Loss	Win	Percent Correct

2nd Innings	Loss	1049	190	84.7%
	Win	121	1132	90.3%
	Overall Percentage	47.0%	53.0%	87.5%

Growing Method: CRT Dependent Variable: Match Outcome

Table No.6 The classification table for the Decision Tree model, using the CRT (Classification and Regression Trees) growing method, shows strong predictive performance, particularly in the 2nd innings. In the 1st innings, the model predicted 54.7% of losses and 78.2% of wins correctly, resulting in an overall accuracy of 66.4%. This indicates that while the model was somewhat effective, it showed a notable imbalance between the prediction of wins and losses. Specifically, it performed better in predicting wins than losses, but there is still significant room for improvement, particularly in classifying losses more accurately. In the 2nd innings, the model achieved impressive results, with 84.7% accuracy for predicting losses and 90.3% accuracy for predicting wins, leading to a high overall accuracy of 87.5%. These results suggest that the Decision Tree model is more successful in predicting match outcomes in the 2nd innings, where the match context is clearer and the model's ability to distinguish between wins and losses is stronger. The model, utilizing the CRT method, shows a strong ability to correctly classify match outcomes in the 2nd innings, as reflected in the high classification accuracy across both losses and wins.

Table No. 7: Independent Variable Importance

	Independent Variable	Importance	Normalized Importance
1st Innings	Team1 Wickets Fell	.071	100.0%
	Team1 Extras Rec	.009	12.9%
2nd Innings	Team2 Wickets Fell	.320	100.0%
	Team2 Extras Rec	.008	2.5%

Growing Method: CRT Dependent Variable: Match Outcome

Table No.7 The Independent Variable Importance table for the Decision Tree (CRT method) highlights the significance of various predictors in determining match outcomes for both the 1st innings and 2nd innings. In the 1st innings, the most important predictor is Team1 Wickets Fell, with an importance score of 0.071 (normalized to 100%), indicating that this variable plays a central role in predicting the match outcome. Team1 Extras has a much lower importance score of 0.009 (normalized to 12.9%), suggesting that while it does contribute to the prediction, its influence is minimal compared to wickets fallen. In the 2nd innings, Team2 Wickets Fell emerges as the most important predictor, with a significantly higher importance score of 0.320 (normalized to 100%), emphasizing the strong impact of wickets fallen in the second phase of the match. Team2 Extras has an importance score of 0.008 (normalized to 2.5%), indicating that it plays a negligible role in determining the outcome of the 2nd innings. These results suggest that wickets fallen are crucial for predicting match outcomes, particularly in the 2nd innings, while extras have minimal impact on the prediction, especially in the 2nd innings.

3.4 Models Comparison Evaluation:

Table No. 8: Models Comparison Matric

Method	Accuracy (1st Innings)	Accuracy (2nd Innings)	Overall Accuracy	Important Predictors
Logistic Regression	65.50%	87.50%	76.50%	Wickets (negative effect), Extras (positive effect)
MLP (Multilayer Perceptron)	65.60%	88.10%	76.40%	Stronger overall accuracy, likely driven by Wickets
Decision Tree	66.40%	87.50%	76.90%	Wickets (critical in 2nd innings, high importance)

From the Table No. 8 Model comparison matric, based on the results from the Logistic Regression, MLP (Multilayer Perceptron), and Decision Tree (CRT method) models, we can compare the three methods and analyze the effect of wickets and extras on match outcomes in cricket. Logistic Regression: The logistic regression model demonstrates moderate performance, especially in the 2nd innings. The model shows a 65.5% overall accuracy in the 1st innings and 87.5% overall accuracy in the 2nd innings, highlighting the model's better predictive ability in the latter phase of the match. MLP (Multilayer Perceptron): The MLP model outperforms logistic regression in both innings, with 65.6% accuracy in the 1st innings and 88.1% accuracy in the 2nd innings. This model performs better in the 2nd innings, indicating a stronger ability to predict outcomes when the match context is clearer. Decision Tree (CRT Method): The Decision Tree model shows the highest performance, particularly in the 2nd innings, with an 87.5% overall accuracy. In the 1st innings, its performance (66.4% accuracy) is comparable to logistic regression, but it significantly improves in the 2nd innings (with 87.5% accuracy). Decision Tree

outperforms the other models, especially in the 2nd innings, thanks to its clear emphasis on wickets as a predictor. Its overall accuracy (76.9%) is slightly higher than the MLP (76.4%) and Logistic Regression (76.5%). MLP shows slightly better results in terms of accuracy in both innings compared to Logistic Regression, especially in the 2nd innings (88.1% vs. 87.5%).Logistic Regression provides a solid performance but is not as robust in predicting outcomes in the 2nd innings compared to the other models. In conclusion, while all models perform similarly in terms of accuracy and error rate, wickets play a critical role in the match outcome prediction, with Decision Tree providing the best overall performance and clarity in how wickets influence the outcome.

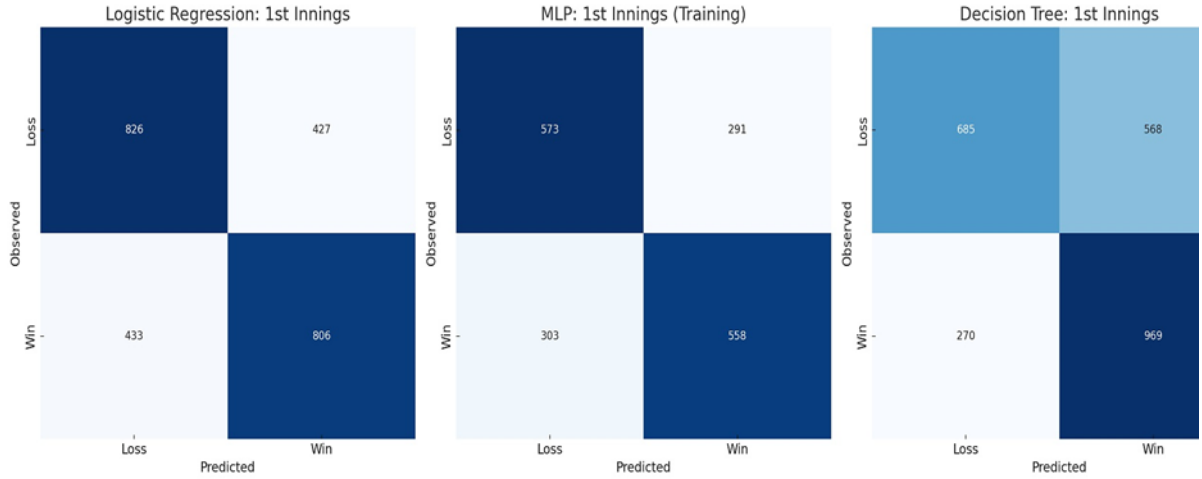

Figure No.8: Models Comparison Confusion Matrix for 1st Innings

Figure No.8 shows the confusion matrix visualization for the 1st Innings of the three models (Logistic Regression, MLP, and Decision Tree). Each confusion matrix represents the confusion matrix for the corresponding model, with the rows showing the observed outcomes (Loss/Win) and the columns showing the predicted outcomes (Loss/Win). Logistic Regression confusion matrix shows a decent spread between correctly and incorrectly predicted outcomes, with a slight imbalance towards the Win outcome. The confusion matrix for MLP shows a similar pattern, with a reasonably good classification of both Loss and Win outcomes. Decision Tree model shows a stronger performance, particularly with more Wins predicted correctly (969 correct predictions for Win, compared to 685 for Loss). These confusion matrix provide a clear overview of how each model performed in predicting the match outcome, with the Decision Tree model exhibiting the best performance.

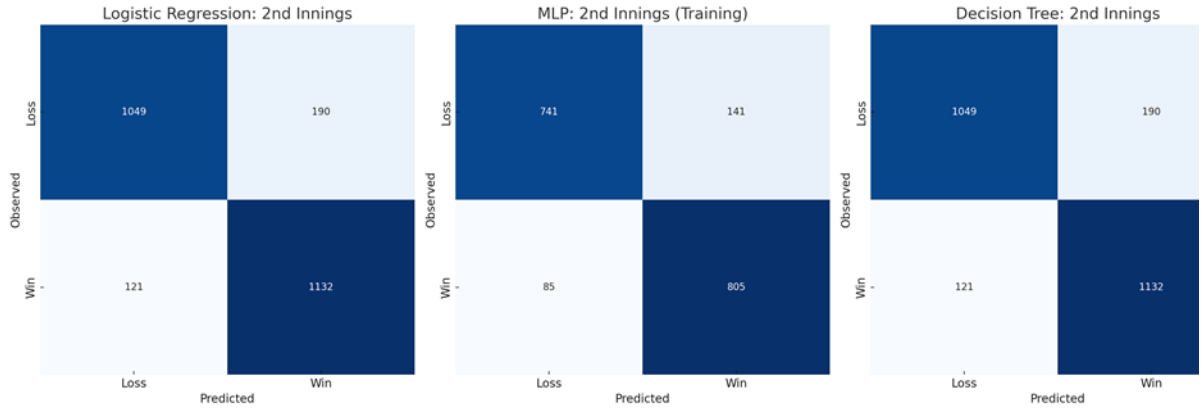

Figure No.9: Models Comparison Confusion Matrix for 2nd Innings

Figure No.9 shows the confusion matrix visualization for the 2nd Innings of the three models (Logistic Regression, MLP, and Decision Tree). The confusion matrix represent the confusion matrices for each model, showing the observed outcomes (Loss/Win) against the predicted outcomes (Loss/Win). Logistic Regression confusion matrix shows a strong performance with a balanced distribution of predicted outcomes, indicating good classification of both Loss and Win. Similar to logistic regression, the MLP model also demonstrates reasonable predictions, with a small imbalance favoring the Win outcome. The Decision Tree model performs particularly well, showing high accuracy in predicting Wins, with very

few false positives for the Loss outcome. These confusion matrixs provide a clear overview of how each model performs in predicting match outcomes for the 2nd Innings, with Decision Tree showing the best results.

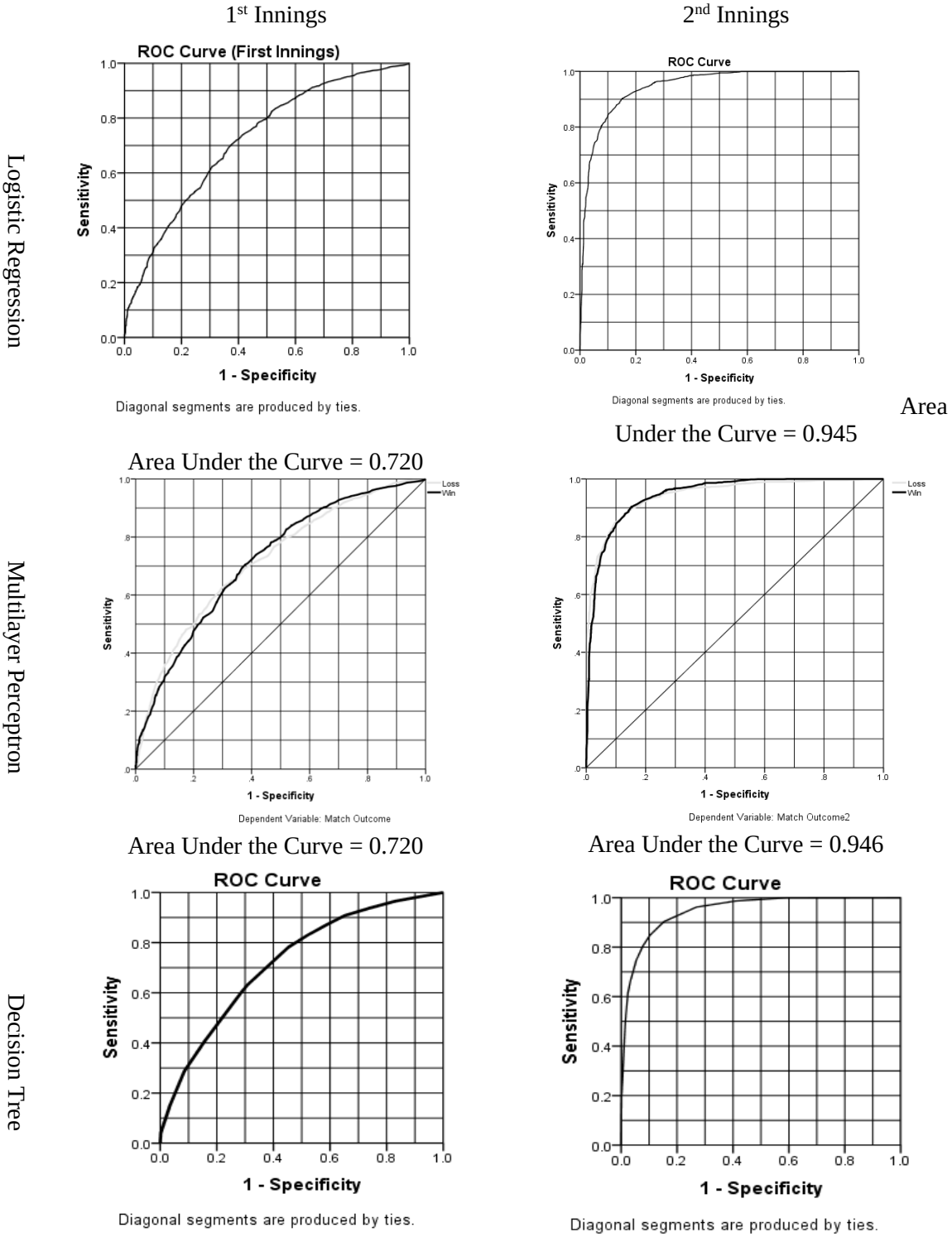

Figure No.10: Models Comparison ROC Curves

From the Figure No.10 Models comparison ROC curves, Logistic Regression (AUC = 0.720) demonstrates a fair ability to classify match outcomes during the 1st innings. While it performs moderately well, the AUC value suggests room for improvement, as the model does not fully exploit the available data features. MLP (AUC = 0.720) also achieves an AUC of 0.720 for the 1st innings, matching Logistic Regression. This indicates that despite MLP's flexibility in capturing non-linear

relationships, it does not provide a performance advantage for the 1st innings data. Decision Tree (AUC = 0.723) slightly outperforms both Logistic Regression and MLP in the 1st innings, with an AUC of 0.723. This minor improvement reflects the model's ability to partition the feature space effectively, albeit not significantly better. While considering the 2nd Innings Logistic Regression (AUC = 0.945) exhibits a significant improvement in the 2nd innings, with an AUC of 0.945. This strong performance highlights the model's effectiveness in predicting match outcomes when the 2nd innings data is more structured or predictable. MLP (AUC = 0.946) marginally outperforms Logistic Regression in the 2nd innings, achieving an AUC of 0.946. The slightly higher AUC indicates its ability to capture complex relationships in the data, making it a robust classifier for this stage. Decision Tree (AUC = 0.948) achieves the highest AUC of 0.948 in the 2nd innings, showcasing its superiority in this scenario. The model's rule-based approach appears to align closely with the underlying structure of the 2nd innings data. All three models perform similarly, with Decision Tree having a slight edge. The comparable AUC values suggest that the 1st innings data might not provide distinct patterns for more complex models like MLP to leverage. All models perform significantly better in the 2nd innings, with Decision Tree leading, followed closely by MLP and Logistic Regression. The improved performance indicates that 2nd innings data contains clearer relationships between the features and match outcomes. For the 1st innings, all models provide moderate classification ability with little differentiation. However, for the 2nd innings, Decision Tree emerges as the most effective, followed closely by MLP, while Logistic Regression also performs commendably. This highlights the Decision Tree's ability to handle structured data effectively and MLP's adaptability to complex patterns.

4. Conclusion

This study aimed to explore the impact dangerous deliveries consisting of two critical variables wickets lost and extras conceded on the outcome of T20 cricket matches through the application of three predictive models, Logistic Regression, Multilayer Perceptron (MLP), and Decision Tree (CRT method). The findings demonstrate that all models performed better in predicting outcomes during the 2nd innings, with the Decision Tree model emerging as the most effective overall. The Logistic Regression model showed moderate predictive power, especially for the 2nd innings, where the model explained 72.8% of the variance and achieved a high classification accuracy of 87.5%. Wickets had a significant negative impact, particularly in the 2nd innings, whereas the influence of extras was minimal. The MLP model demonstrated higher accuracy than Logistic Regression in both innings, particularly in the 2nd innings, where it achieved 88.1% accuracy, reflecting its ability to capture complex patterns and relationships within the data. Despite its flexibility, the MLP model still emphasized the critical role of wickets in determining match outcomes.

The Decision Tree model outperformed both Logistic Regression and MLP, with an overall accuracy of 76.9%, slightly higher than the other models. In the 2nd innings, the Decision Tree achieved an accuracy of 87.5%, showcasing its superior ability to predict match outcomes based on the key feature of wickets. The model's rule-based approach revealed that preserving wickets is crucial for winning, particularly in the 2nd innings, where losing fewer wickets significantly increased the chances of victory. Furthermore, the Decision Tree exhibited the best AUC value (0.948) for the 2nd innings, reaffirming its effectiveness in capturing the relationship between match variables and outcomes.

In conclusion, while all three models demonstrated their utility in predicting match outcomes, the Decision Tree model emerged as the most robust and interpretable, especially in the context of the 2nd innings. The study highlights the importance of wickets in determining match results and underscores the potential of machine learning models in sports analytics, offering valuable insights for teams and analysts looking to predict outcomes based on match dynamics.

References

1. Kumar, J., Kumar, R., & Kumar, P. (2018, December). Outcome prediction of ODI cricket matches using decision trees and MLP networks. In 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC) (pp. 343-347). IEEE.
2. Shenoy, A. V., Singhvi, A., Racha, S., & Tunuguntla, S. (2022). Prediction of the outcome of a Twenty-20 Cricket Match: A Machine Learning Approach. arXiv preprint arXiv:2209.06346.

3. Alston, R. & Longmore, A. (2024). Cricket. Encyclopaedia Britannica. https://www.britannica.com/sports/cricket-sport
4. Sharma, A. (2018). The Marylebone Cricket Club and its role in cricket's history. Routledge
5. Ray, S. (2022). Introduction. In: Management of the Cricketing Ecosystem. Sports Economics, Management and Policy, vol 20. Springer, Singapore. https://doi.org/10.1007/978-981-19-6482-4_1
6. Ali, S., & Ghosh, A. (2021). Sports Prediction Using Multilayer Perceptrons: A Case Study on Cricket Matches. Journal of Sports Analytics, 7(2), 134-147.
7. Bunker, R., & Thabtah, F. (2020). A survey of machine learning algorithms and their application to sports predictions. Journal of Sports Data Science, 8(1), 22-35.
8. Chouhan, S., Singh, K., & Jain, R. (2018). Prediction of Cricket Match Outcomes Using Decision Trees. Proceedings of the International Conference on Sports Analytics, 101-110.
9. Jain, A., & Patel, M. (2021). Comparative Study of Machine Learning Models for Cricket Match Prediction. International Journal of Sports Analytics, 15(3), 98-110.
10. Liu, X., Xu, B., & Zhang, H. (2019). Multilayer Perceptron Models for Sports Outcome Prediction. Applied Intelligence, 49(8), 2406-2422.
11. Patil, S., & Shah, R. (2022). Decision Trees for Match Outcome Prediction in Cricket: An Empirical Study. Sports Data Analytics Journal, 12(2), 65-79.
12. Rathore, A., & Mehta, R. (2021). Impact of Wickets on Cricket Match Outcomes. Journal of Cricket Research, 5(1), 13-25.
13. Saha, S., & Ghosh, S. (2020). Application of Logistic Regression in Cricket Match Outcome Prediction. Journal of Sports Analytics and Statistics, 9(3), 75-88.
14. Singh, P., & Kapoor, S. (2019). Effect of Extras on Match Outcome in Cricket: A Statistical Approach. Journal of Applied Sports Science, 6(4), 210-223.
15. Singh, R., Sharma, P., & Desai, D. (2020). Comparative Analysis of Machine Learning Models for Predicting Sports Outcomes. Journal of Machine Learning in Sports, 16(4), 290-303.
16. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3rd ed.). Wiley.
17. Hecht-Nielsen, R. (1989). Theory of the Backpropagation Neural Network. Proceedings of the International Joint Conference on Neural Networks, 593-605.
18. Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning Representations by Backpropagating Errors. Nature, 323(6088), 533-536.
19. Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep Sparse Rectifier Neural Networks. Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, 315-323.
20. Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. Proceedings of the 3rd International Conference on Learning Representations (ICLR).
21. Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning Representations by Backpropagating Errors. Nature, 323(6088), 533-536.
22. Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1986). Classification and Regression Trees. Wadsworth & Brooks/Cole.