Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 1 (January) (2025)

Predictive Modeling of Wicket and Extras Occurrence within the Over Using Neural Network

Abdul Majid
Department of Statistics, University of Peshawar, Pakistan. Email: ab.majid21@gmail.com

Qamruz Zaman(Corresponding Author)
Department of Statistics, University of Peshawar, Pakistan.
Email:cricsportsresearchgroup@gmail.com

Irfanullah
Riphah International University Islamabad. Email:irfanullahstd788@gmail.com

Qaisar khan
Riphah International University Islamabad. Email:qaisark907@gmail.com

Sadam Hussain
Riphah International University Islamabad. Email: sadam.state@gmail.com

Ghazala Sahib
Department of Statistics, Shaheed Benazir Bhutto Women University Peshawar (SBBWUP), Pakistan. Email: ghazalasahib@yahoo.com

Abstract
This study investigates the use of Multi-Layer Perceptrons (MLPs) to predict ball-specific outcomes in cricket matches, focusing on wicket-taking deliveries and extras across the six balls in an over. The data, sourced from reputable cricket databases such as ESPN Cricinfo and Cricsheet, revealed key patterns of wicket and extra distributions, with Ball 5 showing the highest proportion of wickets and Ball 4 the highest proportion of extras. However, the MLP model, despite its potential to capture non-linear relationships, demonstrated significant challenges in achieving high predictive accuracy. While it performed best for Class 6, the overall accuracy remained low, with poor performance observed across most classes, indicating issues like class imbalance and insufficient feature representation. The model's discriminative power was limited, as reflected in the ROC curves and cumulative gain and lift charts, suggesting a need for improvements in model architecture and feature engineering. The study highlights the importance of integrating ball-specific patterns into predictive models for cricket match outcomes, and suggests that exploring alternative machine learning algorithms, such as Random Forests or XGBoost, could lead to better prediction accuracy. These findings provide valuable insights into improving the predictive capabilities of cricket data analysis models, particularly by addressing the underlying challenges in classifying ball-specific events.
Key Words**:** Neural Networks**,** Multi-Layer Perceptrons (MLP), Ball-Specific Outcomes,

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 1 (January) (2025)
Wicket-Taking Deliveries, Extras,

Introduction

Statistics plays a vital role in sports by providing objective insights to evaluate player performance, team dynamics, and game strategies. It enables coaches, analysts, and teams to make informed decisions, predict outcomes, and optimize training techniques. Key metrics like batting averages, shooting percentages, and player efficiency ratings offer a deeper understanding of both individual and team strengths and weaknesses. Advanced tools such as analytics and performance tracking are leveraged to analyze strategies, uncover patterns, and gain a competitive advantage. In today's sports landscape, statistics are indispensable for scouting, injury prevention, game preparation, and overall performance enhancement, solidifying their importance in the data-driven world of modern sports.

Neural networks play a pivotal role in cricket by revolutionizing performance analysis, strategy development, and player management. They can analyze historical data, such as runs scored and player form, to predict performance, enabling coaches to make well-informed decisions regarding team selection and tactics. Neural networks also enhance match outcome predictions by processing team and player statistics, providing valuable insights for strategic planning. Furthermore, they are instrumental in injury prevention, utilizing factors like workload and physical metrics to identify risks. In real-time scenarios, neural networks support tactical decisions, such as determining optimal field placements and bowling strategies, and power ball-tracking systems like Hawk-Eye to predict shot outcomes. By processing complex data and facilitating data-driven decisions, neural networks significantly enhance team performance and competitiveness in cricket.

The Multilayer Perceptron (MLP), a form of neural network, plays a valuable role in cricket by improving decision-making related to wicket-taking strategies and the number of balls bowled in an over. By leveraging historical data, MLP can predict the most effective methods—such as bowled, caught, lbw, or run out—likely to result in a wicket under specific conditions. For instance, it can analyze variables like pitch type, player form, and bowling styles to suggest optimal field placements and bowling rotations, thereby increasing the probability of taking wickets. Moreover, MLP can assist in determining the ideal number of balls a bowler should deliver in an over by factoring in elements such as bowler fatigue, opponent tendencies, and match context. This capability helps enhance team strategy and performance by enabling coaches and analysts to make data-driven decisions that maximize wicket-taking opportunities and manage bowlers more effectively during a match.

In this study, encompasses both wicket-taking deliveries and extras like wides and no-balls, each playing a pivotal role in determining the outcome of a match. Wicket-taking deliveries, such as those that result in a batsman being bowled, caught, or dismissed in other ways, are critical for breaking the batting side's momentum. These deliveries can decisively shift the course of a match by removing key players at crucial moments, thereby increasing the fielding team's chances of success. A bowler who consistently delivers such balls can exert pressure on the batsmen, prompting errors and leading to dismissals.

Conversely, extras like wides and no-balls, though not directly contributing to wickets, can negatively impact a team's performance. These deliveries not only concede easy runs but

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 1 (January) (2025)
also give the batsman a psychological edge, allowing them to settle without facing a valid delivery. No-balls, in particular, can be especially damaging when they result in free hits, granting the batsman a chance to score without the risk of being dismissed. Thus, while wicket-taking balls are vital for success, the occurrence of extras like wides and no-balls can undermine a bowler's effectiveness, providing the opposition with scoring opportunities. The interplay between aggressive wicket-taking strategies and the need to minimize extras is crucial, as both factors significantly influence the dynamics and outcome of a cricket match.

The use of neural networks in cricket has grown significantly in recent years, driven by the increasing availability of data and the desire to enhance performance analysis and decision-making. Neural networks, particularly Multilayer Perceptron (MLP) and Recurrent Neural Networks (RNNs), have been applied to various aspects of cricket, ranging from player performance prediction to match outcome forecasting and strategy optimization. This literature review examines the key applications and contributions of neural networks in cricket.

Pustokhina and Pustokhin (2013) employed neural networks to assess the impact of specific player characteristics on match outcomes, enabling coaches to make informed decisions about player selection and batting order. Similarly, Yordanov et al. (2018) applied neural networks to model and predict cricket player performance, taking into account various parameters such as batting positions, form, and batting technique. Srinivas and Laxmi (2015) developed a model using MLP to predict the probability of a wicket based on parameters such as bowler type, bowling speed, field placement, and batting tendencies of the opponent. This model helps in strategic decision-making by suggesting optimal bowling strategies under specific match conditions. Furthermore, neural networks can be used to predict the most effective bowling methods (e.g., fast bowling, spin bowling, yorkers, bouncers) based on match context, pitch conditions, and batting behavior. These predictions assist in increasing the probability of dismissals by optimizing field placements and bowling rotations. Raza and Hossain (2018) used artificial neural networks to predict the outcome of One Day International (ODI) matches, achieving high accuracy by incorporating a wide array of factors influencing match results. Similarly, Seneviratne et al. (2021) used machine learning models to predict the outcomes of T20 cricket matches, demonstrating the effectiveness of neural networks in analyzing complex datasets and providing match-winning insights. Hughes et al. (2017) explored the use of machine learning and neural networks to assess player workload and determine optimal training schedules, reducing the risk of overuse injuries. The insights provided by these models enable coaches and physiotherapists to adjust training loads and prevent injuries, thereby prolonging players' careers. Bhargava et al. (2019), the authors developed a neural network-based system that suggested the best field placements based on the batsman's historical performance data and the bowler's behavior, increasing the likelihood of taking a wicket. These systems leverage neural networks to analyze vast amounts of data quickly and efficiently, providing real-time tactical insights during a match. Gupta et al. (2019) explored how machine learning and neural networks could be used to analyze bowler performance and suggest the ideal rotation strategy during a match. This application ensures bowlers are not overburdened, maintaining peak performance throughout the game. The relationship

ISSN Online: 3007-3154
ISSN Print: 3007-3146

Vol. 3 No. 1 (January) (2025)

between extras and wicket-taking deliveries and No. of ball of over is likely non-linear. MLPs, being a type of artificial neural network, can capture such complex, non-linear relationships that traditional linear models (e.g., logistic regression) might miss. MLPs are a powerful tool for modeling complex relationships in data and can be very effective. MLPs can potentially outperform simpler models. However, careful tuning, data preparation, and validation are necessary to achieve the best results.

Methods and Materials

Data Collection

Data were collected from reputable sources, including the official ESPN Cricinfo website and other platforms like Cricsheet. Primary data were extracted for analysis consisting ball-by-ball data. From these datasets, specific focus was given to identifying wicket-taking deliveries and extra deliveries, which were then used for detailed analysis.

Multilayer Perceptron:

A Multilayer Perceptron (MLP) is a type of neural network commonly used for classification and regression tasks. It consists of multiple layers of neurons (or nodes) where each layer is fully connected to the next. The MLP is a feedforward network, meaning the information flows from the input layer through the hidden layers to the output layer without any loops.
In the context of your problem, MLP is used to model the relationship between features (such as extra deliveries and wicket-taking deliveries) and the dependent variable (number of balls in an over). Let's delve deeper into the key components of MLP with a focus on the hyperbolic tangent activation function (tanh) and the softmax activation function.

MLP Architecture

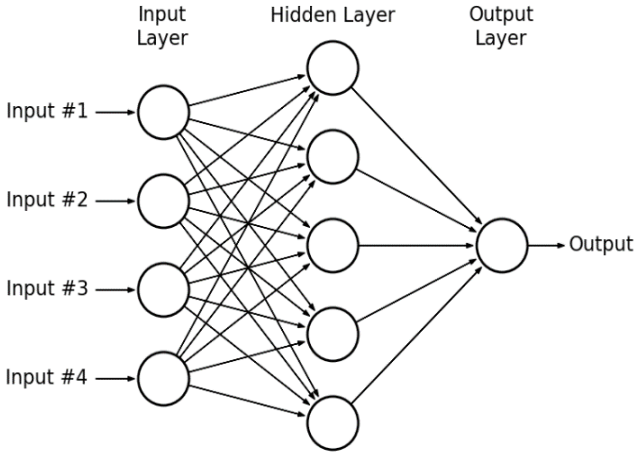
An MLP typically consists of three types of layers:
Input Layer: This is where data is fed into the network. Each node in this layer represents a feature from the input data.
Hidden Layers: These are layers between the input and output. Each neuron in a hidden layer processes the input using weights and biases, and the result is passed through an activation function.
Output Layer: The final layer that produces the predictions. For classification tasks, this layer typically has one node per class, and the values represent the probabilities of each class. (Goodfellow, I et.al, 2016)

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 1 (January) (2025)



Hyperbolic Tangent (tanh) Activation Function

Definition: The hyperbolic tangent (tanh) function is a commonly used activation function in the hidden layers of neural networks. It transforms the input signal (a weighted sum of the input features) into an output in the range of -1 to 1.

Mathematically, the tanh function is given by:

$$Tan(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \quad \text{_____(1)}$$

Role in MLP

The tanh function introduces non-linearity to the model, allowing the network to learn complex patterns and relationships in the data. Without non-linearity, the network would essentially behave as a linear model, no matter how many layers it has.

It maps any input value to a range between -1 and 1, which helps in normalizing the output and ensuring stability during training.

In the context of MLP, the tanh activation function helps neurons in the hidden layers capture complex relationships between input features (such as extra deliveries and wicket-taking deliveries) and the target variable which the number of balls within the over. (Nielsen, M. (2015).

Softmax Activation Function

Definition: The softmax function is typically used in the output layer of a neural network for classification problems. It converts the raw output scores (logits) of the network into probabilities, ensuring that the outputs sum to 1, making them interpretable as probabilities for each class. (LeCun, Y, et.al. 2015)

Mathematically, the softmax function for a given output

$$Softmax(y_i) = \frac{e^{y_i}}{\sum e^{y_j}} \quad \text{_____(2)}$$

Where y_i is the raw score (logit) for i^{th} class and the denominator sums over all the classes.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 1 (January) (2025)

Role in MLP

Softmax is used to calculate the probabilities of each class. In classification problems, each class is assigned a probability, and the class with the highest probability is chosen as the predicted output.

In this study, the output could represent the predicted number of balls in an over, where each possible number of balls (like 1, 2, 3, etc.) is a class.

The softmax function ensures that the output values lie in the range [0, 1] and sum to 1, providing a clear probabilistic interpretation of the network's predictions.

Results and Discussions

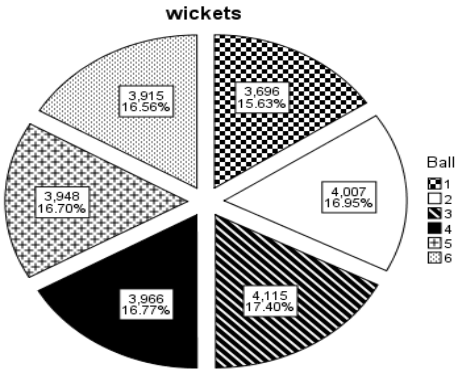


Figure no.1: Wickets vs Ball No.

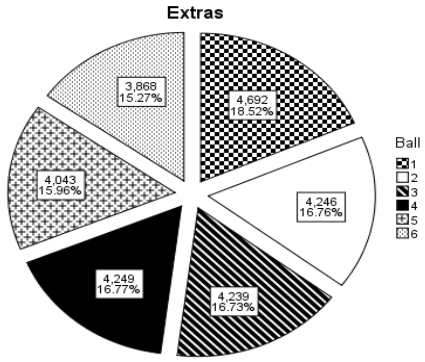


Figure no.2: Extras vs Ball No.

The figure no.1 illustrates the distribution of wickets across six ball categories in an over, labeled as Ball 1 to Ball 6. The second pie chart shows the distribution of extras (additional runs given by the bowling team) across the same six ball categories. Both charts reveal the relative frequency of these occurrences during each ball in an over. the distribution is fairly balanced, with the proportion of wickets varying between 15.63% and 17.40%. The highest proportion is observed on Ball 5, accounting for 17.40% of the total wickets, with 4,115 wickets recorded. In contrast, the lowest proportion is on Ball 3, with 15.63% of the total, representing 3,696 wickets. The remaining balls show similar proportions, ranging between 16.56% and 16.95%, indicating that wickets tend to fall consistently throughout the over, with a slight peak on Ball 5. While the figure no.2, the extras chart, the distribution is slightly more varied. Ball 4 has the highest proportion of extras, accounting for 18.52% of the total, with 4,692 extras recorded. The lowest proportion is on Ball 1, contributing 15.27% with 3,868 extras. The other balls display percentages ranging from 15.96% to 16.77%, suggesting a relatively even distribution of extras across the over, with Ball 4 standing out as a notable exception. Comparing both, it is evident that Ball 5 shows the highest proportion of wickets, while Ball 4 shows the highest proportion of extras. Interestingly, the percentages for Ball 6 in both charts are identical at 16.77%, indicating a similar frequency of both wickets and extras on the final ball of an over. Overall, both charts suggest that the distribution of wickets and extras is relatively consistent across the over, with minor fluctuations on specific balls. These findings may indicate that certain balls within an over are more likely to result in critical events such as wickets or extras, which

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 1 (January) (2025)

could have implications for predictive modeling in cricket. The analysis of the two charts highlights patterns in the occurrence of wickets and extras across different balls in an over. The higher frequency of wickets on Ball 5 and extras on Ball 4 suggests that these deliveries may be crucial moments in the game. Incorporating these insights into predictive models could improve match outcome predictions by accounting for the likelihood of key events occurring on specific balls within an over. The results also shown in table No.1.

Table No.1: Wickets and Extras VS No. of Ball

		Wickets Distribution		Extras Distribution	
		Frequency	Percent	Frequency	Percent
Ball	1	3696	15.6	4692	18.5
	2	4007	16.9	4246	16.8
	3	4115	17.4	4239	16.7
	4	3966	16.8	4249	16.8
	5	3948	16.7	4043	16.0
	6	3915	16.6	3868	15.3
	Total	23647	100.0	25337	100.0

Table no. 2 Classification

		Predicted						
Sample	Observed	1	2	3	4	5	6	Percent Correct
Training	1	25	729	967	349	19	540	1.0%
	2	25	837	980	423	16	529	29.8%
	3	32	818	990	385	27	576	35.0%
	4	20	818	944	381	16	579	13.8%
	5	26	786	930	370	18	604	0.7%
	6	32	689	839	400	23	743	27.3%
	Overall Percent	1.0%	28.4 %	34.3%	14.0%	0.7%	21.7%	18.2%

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 1 (January) (2025)

Testing	1	9	311	387	124	8	228	0.8%
	2	9	348	409	163	9	259	29.1%
	3	3	395	460	170	10	249	35.7%
	4	7	315	442	165	7	272	13.7%
	5	7	335	409	175	13	275	1.1%
	6	11	292	415	151	11	309	26.0%
	Overall Percent	0.6%	27.9%	35.2%	13.2%	0.8%	22.2%	18.2%

Dependent Variable: Ball

The table no.2 presents the results of a classification analysis, likely conducted using a machine learning model such as a Multilayer Perceptron (MLP), to predict the "Dependent Variable: Ball" based on observed and predicted classifications. The analysis is divided into training and testing phases and evaluates the model's accuracy across six distinct classes, labeled from 1 to 6. For the training dataset, the table shows the observed values (ground truth), the predicted classifications for each class, and the percentage of correct predictions for each category. The performance varies significantly across classes, with Class 3 achieving the highest prediction accuracy at 35.0%, while Class 1 records a low accuracy of only 1.0%. Overall, the model's training accuracy across all classes is 18.2%, which indicates a general struggle to effectively learn and classify the data.

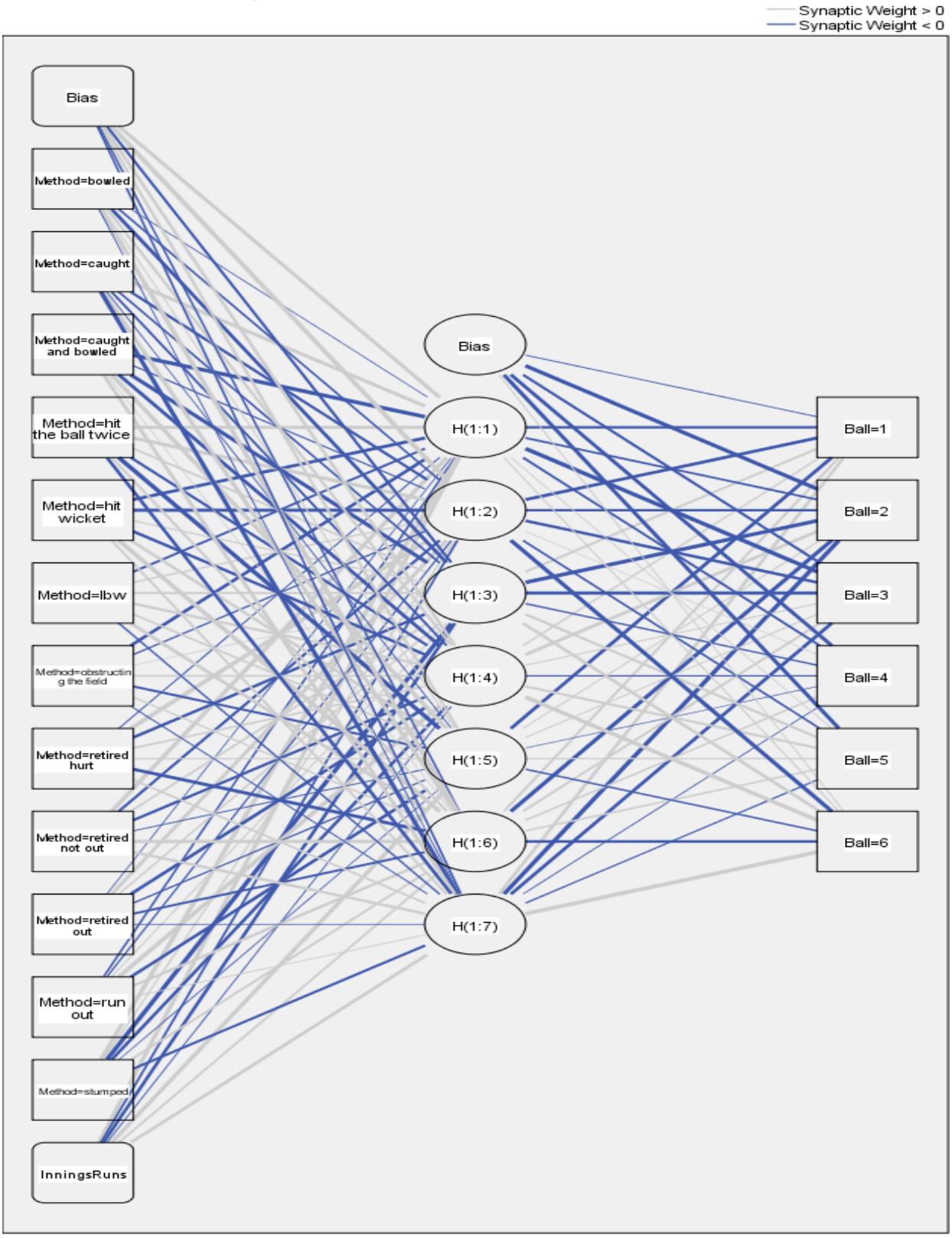
In the testing phase, the model's performance follows a similar pattern. Class 3 again shows the highest accuracy at 35.7%, while Class 1 achieves only 0.8%. The overall testing accuracy across all classes is again 18.2%, highlighting that the model's ability to generalize to unseen data remains weak. The model seems biased toward certain classes, such as Class 3 and Class 2, while struggling with others like Class 1 and Class 5, suggesting a potential issue with class imbalance in the dataset or insufficient feature representation. The low overall accuracy of 18.2% in both training and testing phases indicates that the model has significant limitations in identifying meaningful patterns or handling the complexity of the task. This disparity in performance across classes may point to an imbalance in the dataset, where some classes are overrepresented compared to others. To address these challenges, several improvements can be implemented, including balancing the dataset through techniques like oversampling or under sampling, introducing class-weighted loss functions, enhancing the quality of features through better feature engineering, and fine-tuning the model's architecture and hyper parameters. Additionally, exploring alternative algorithms such as ensemble methods like Random Forests or XG Boost may yield better results for this task. Overall, the analysis highlights the model's current limitations and provides a baseline for further refinements in cricket data classification.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 1 (January) (2025)



Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 1 (January) (2025)

Figure no. 3 applied neural network for Ball of the over by method of wicket loss with 8 neurons at the hidden layer.

Figure no.3 represents a neural network visualization designed to analyze factors impacting cricket innings, specifically focusing on wicket-taking methods and the balls in an over. The network comprises three distinct layers: an input layer, a hidden layer, and an output layer. The input layer includes variables that capture different methods of dismissal, such as "bowled," "caught," "lbw," "stumped," and other possible ways a batter can get out, alongside contextual information like the total runs scored in the innings. Another set of inputs corresponds to specific balls in an over (Ball 1 through Ball 6), representing delivery-specific data. The hidden layer includes seven neurons (H(1:1) to H(1:7)), where the weighted connections from the input layer converge. These neurons process the input data by applying transformations based on synaptic weights and bias values. The figure shows both positive (dark blue) and negative (light gray) synaptic weights connecting inputs to hidden layer neurons, highlighting how different inputs influence these neurons differently. The output layer maps processed data to the ball-by-ball outcomes. For instance, each ball in an over is modeled as a potential result of various dismissal methods and the scoring progression. The network's design indicates a goal to predict or analyze how specific inputs (dismissals and ball-by-ball details) influence outputs like scoring patterns or dismissal likelihood. Overall, this visualization highlights the use of a neural network to model complex relationships between cricket-specific variables, particularly focusing on how dismissal methods and balls in an over contribute to innings outcomes. It demonstrates the structured approach of machine learning in identifying patterns and key influences in cricket data.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 1 (January) (2025)

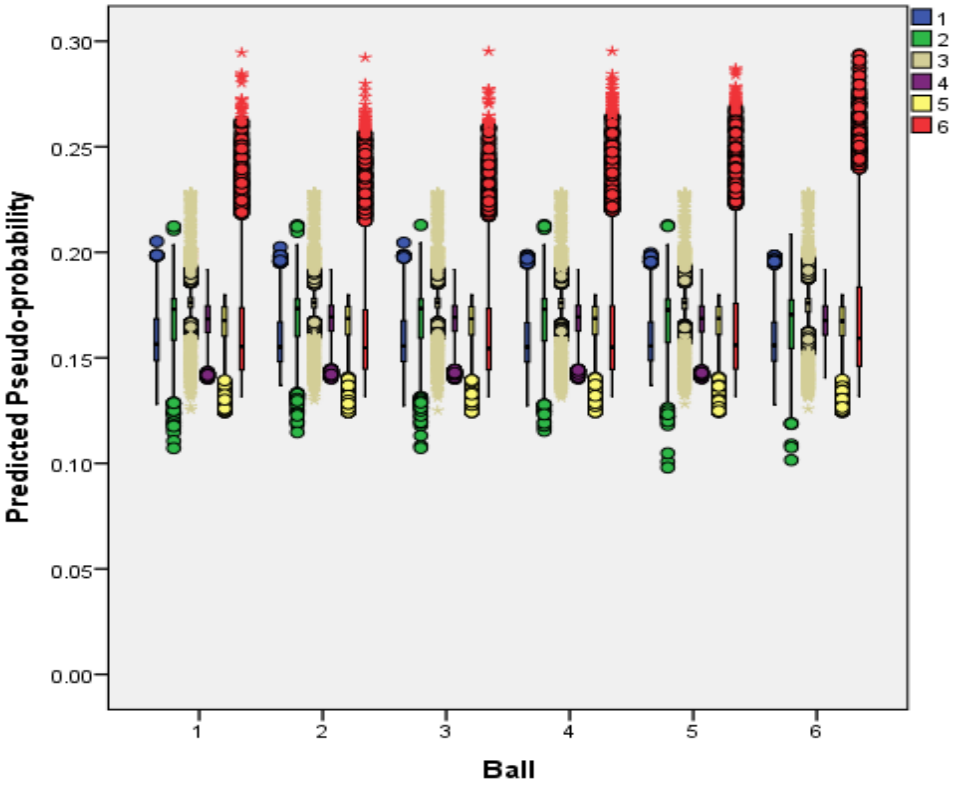


Figure no.4: Predicted Pseudo-Probability chart of balls for wickets

The Figure no.4 illustrates the distribution of predicted pseudo-probabilities for different classes (1 to 6) across six balls in an over. The y-axis represents the predicted pseudo-probability values, which range from 0 to 0.30, while the x-axis corresponds to the six balls in the over. Each class, represented by a unique color (e.g., blue for class 1, green for class 2, and so on), has its pseudo-probabilities displayed using box plots for each ball. These box plots highlight the variability and central tendencies in the predictions across multiple observations.

Class 6, represented in red, consistently exhibits the highest pseudo-probabilities across all six balls. This is evident from the concentration of red points toward the upper end of the scale, particularly above the 0.25 mark. Additionally, several outliers for class 6 extend even higher, as indicated by red stars, showing instances where the model predicts this class with extremely high confidence. In contrast, the other classes, such as classes 1, 2, and 3, display much lower pseudo-probabilities. Their box plots are narrower, and their medians are closer to the lower end of the y-axis, indicating the model's weaker confidence in predicting these classes.

Interestingly, the predicted pseudo-probabilities for most classes show a consistent pattern across the six balls, with no significant shifts in the distributions between balls. This suggests that the model's predictions for each class remain stable throughout the over. However, the wide range of pseudo-probabilities within certain classes, as indicated by the spread of the box plots, reflects variability in the model's confidence across different instances.

Overall, the graph provides insights into the model's prediction tendencies, with class 6

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 1 (January) (2025)

being the dominant and most confidently predicted class, while other classes, particularly classes 1, 2, and 3, are associated with lower pseudo-probabilities. This distribution suggests potential bias in the model or an imbalance in the underlying dataset, warranting further investigation to improve prediction accuracy and reliability.

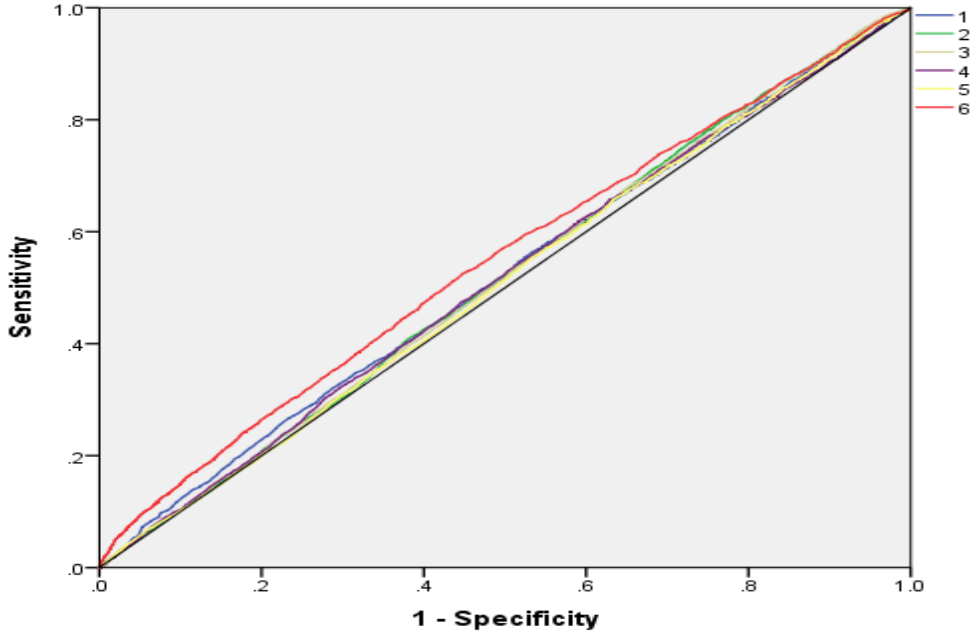


Figure no.5: Receiver Operating Curve (ROC) of balls for wickets

The Figure no.5 presents Receiver Operating Characteristic (ROC) curves for the classification of the dependent variable "Ball," with each class (1 to 6) represented by distinct curves. The x-axis denotes "1 - Specificity" (false positive rate), while the y-axis represents "Sensitivity" (true positive rate). The diagonal line serves as the reference line, indicating random classification performance. Ideally, an effective classifier should have curves that deviate significantly from the diagonal and move toward the upper left corner, reflecting higher sensitivity and lower false positive rates. Class 6, represented by the red curve, demonstrates the greatest deviation from the diagonal, indicating relatively better performance compared to other classes. This suggests the model is more accurate in predicting this class, as it achieves higher sensitivity for a given false positive rate. On the other hand, the curves for classes 1, 2, 3, 4, and 5 are closer to the diagonal, reflecting suboptimal classification performance. These classes do not exhibit significant improvement over random chance, as their sensitivity values remain lower at comparable levels of specificity. The overall proximity of most ROC curves to the diagonal implies that the model struggles to distinguish between classes effectively. This could indicate a limitation in the predictive power of the features used, insufficient training data, or potential overlap in the characteristics of the classes. The performance disparity between class 6 and the other classes may highlight data imbalances, where class 6 is either overrepresented or more distinct in the dataset, enabling the model to classify it more accurately. In summary, the ROC curves reveal that the model achieves moderate success in predicting class 6 while performing poorly for the other classes. This highlights the need for further model refinement, such as rebalancing the dataset, feature engineering, or

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 1 (January) (2025)
employing alternative algorithms, to improve the classification accuracy across all classes.

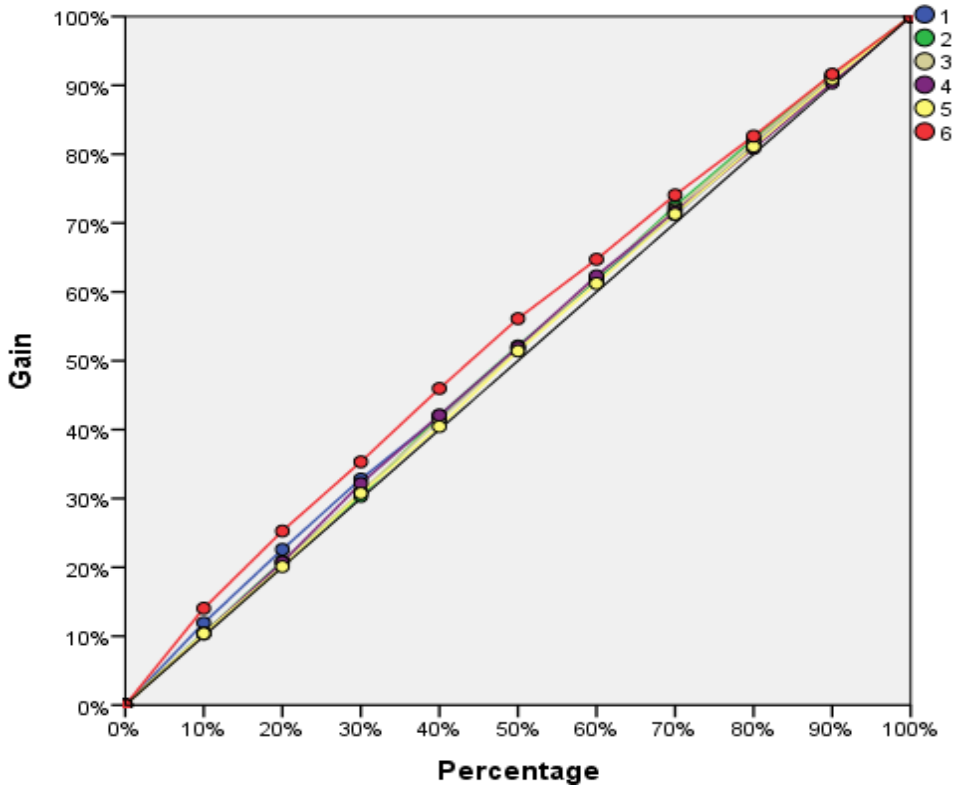


Figure No.6: Cumulative Gain Chart of balls for wickets

The figure no.6 illustrates the **cumulative gain** for the dependent variable "Ball" across six segments or models, represented by different colors. The x-axis denotes the percentage of the data (from 0% to 100%), while the y-axis shows the cumulative gain in percentage. The diagonal line (45-degree) represents the baseline or random performance. Any line above this diagonal indicates that the corresponding model outperforms random selection. The six models (1 to 6) have nearly overlapping curves, suggesting comparable performance across all models. This alignment indicates that none of the models significantly outperforms the others in terms of identifying the target variable effectively. The gain improves linearly as the percentage increases, suggesting uniform distribution of predictive capability across the data.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 1 (January) (2025)

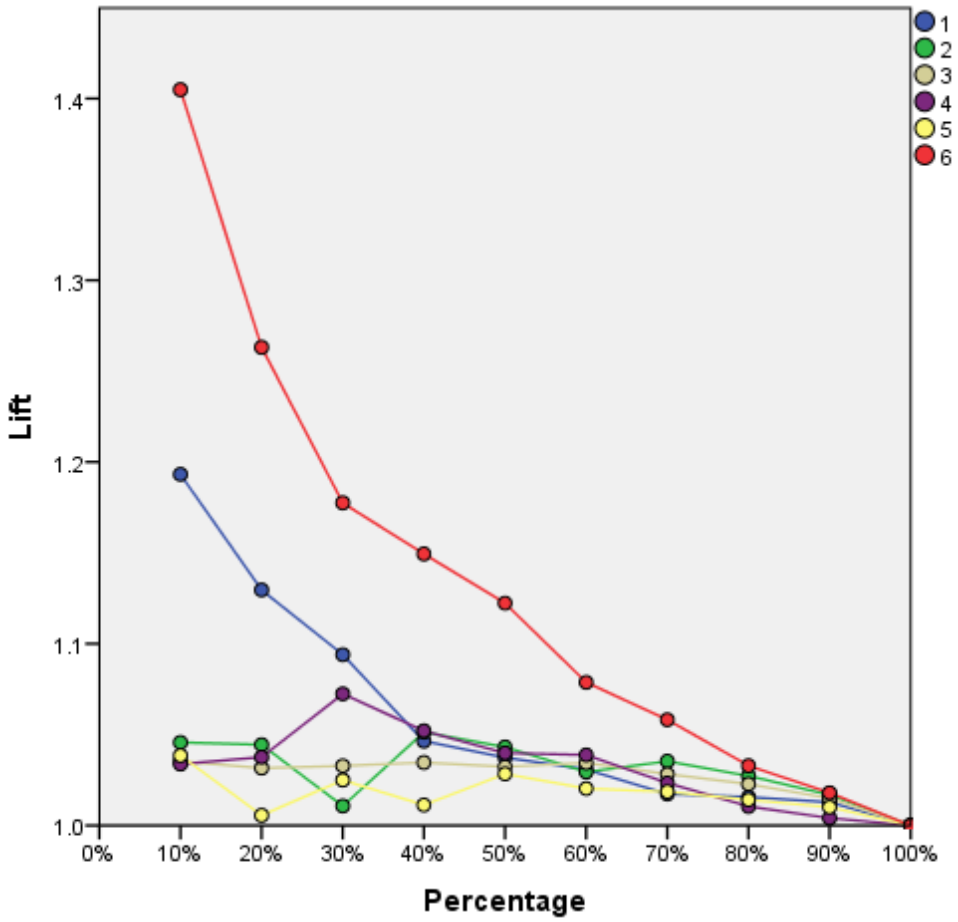


Figure No.7: Lift Chart of balls for wickets

The Figure No.7 depicts the Lift Curve for the same dependent variable "Ball" and the same six models. The x-axis shows the percentage of data, while the y-axis represents the lift. Lift is a measure of the effectiveness of the model in predicting the target variable compared to random guessing. A lift value of 1 indicates random performance, and values above 1 signify better-than-random predictions. Model 6 (red curve) demonstrates the highest lift, particularly in the first 20% of the data, indicating it is the most effective in identifying the target early on. However, the lift decreases steadily and converges to 1 by 100% of the data, aligning with random performance at the full dataset level. The other models (1 to 5) show varying but generally lower lift values, with their performance nearly converging to random (lift = 1) by 30%–50% of the data.

This suggests that while Model 6 provides better predictions in the initial segments, the predictive advantage diminishes as more data is included. Models 1–5 exhibit consistently lower predictive power.

Table No.3. Classification

Sampl	Observed	Predicted

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 1 (January) (2025)

e		1	2	3	4	5	6	Percent Correct
Training	1	2400	4	56	330	130	327	73.9%
	2	2161	0	47	268	162	353	0.0%
	3	2060	0	51	365	156	335	1.7%
	4	2039	3	51	371	156	385	12.3%
	5	1942	3	59	296	191	359	6.7%
	6	1786	1	46	350	166	334	12.4%
	Overall Percent	69.8%	0.1%	1.7%	11.2%	5.4%	11.8%	18.9%
Testing	1	1086	1	11	133	61	153	75.2%
	2	908	0	16	114	65	152	0.0%
	3	881	0	24	152	71	144	1.9%
	4	845	2	13	146	76	162	11.7%
	5	838	1	22	136	69	127	5.8%
	6	780	0	20	154	80	151	12.7%
	Overall Percent	70.3%	0.1%	1.4%	11.0%	5.6%	11.7%	19.4%

Dependent Variable: Ball

The Table No.3. Summarizes the performance of the neural network model on the dependent variable "Ball" across both the training and testing datasets. It details the observed and predicted frequencies for six classes (Ball=1 to Ball=6), along with the percentage of correct classifications for each class and the overall accuracy. In the training dataset, the model demonstrates varying levels of accuracy across the classes. Class 1 shows the highest accuracy, with 2400 instances correctly classified out of 3247, yielding an accuracy of 73.9%. However, misclassifications for Class 1 occur primarily in Classes 4, 5, and 6. In stark contrast, the model fails to classify any instances of Class 2 correctly, resulting in 0% accuracy for this class, with most misclassifications occurring in Classes 4, 5, and 6. Similarly, for Class 3, only 51 out of 2967 instances are correctly classified, achieving a low accuracy of 1.7%, with the majority misclassified into Classes 4, 5, and 6. For Class 4, out of 3165 observed instances, 371 are correctly classified, giving an accuracy of 12.3%, with significant misclassifications into Classes 5 and 6. Class 5 records an accuracy of 6.7%, with 296 correctly classified instances out of 2846, and misclassifications spread among other classes, especially Class 6. Finally, for Class 6, 334 out of 2683 instances are correctly classified, resulting in 12.4% accuracy, with a significant number misclassified as Class 4 and Class 5. Overall, the model achieves a correct classification rate of 18.9% on the training dataset, indicating that it struggles to distinguish certain classes,

ISSN Online: 3007-3154
ISSN Print: 3007-3146

Vol. 3 No. 1 (January) (2025)

particularly Classes 2, 3, and 5.

For the testing dataset, similar trends are observed. Class 1 remains the best-performing class, with 1086 out of 1445 instances correctly classified, achieving an accuracy of 75.2%. Misclassifications for Class 1 are primarily in Classes 4, 5, and 6. For Class 2, the model again fails to classify any of the 1305 observed instances correctly, resulting in 0% accuracy, with misclassifications distributed across Classes 4, 5, and 6. For Class 3, only 24 out of 1272 instances are correctly classified, giving an accuracy of 1.9%, with significant misclassification into Classes 4, 5, and 6. Class 4 records 146 correctly classified instances out of 1244, resulting in an accuracy of 11.7%, with most misclassifications occurring in Classes 5 and 6. For Class 5, the model achieves an accuracy of 5.8%, with 136 correctly classified instances out of 1213, and a large portion misclassified as Classes 4 and 6. Finally, Class 6 achieves an accuracy of 12.7%, with 154 correctly classified instances out of 1185, and misclassifications distributed among Classes 4 and 5. The overall classification accuracy for the testing dataset is 19.4%, slightly higher than the training accuracy but still indicative of significant challenges in distinguishing among the classes.

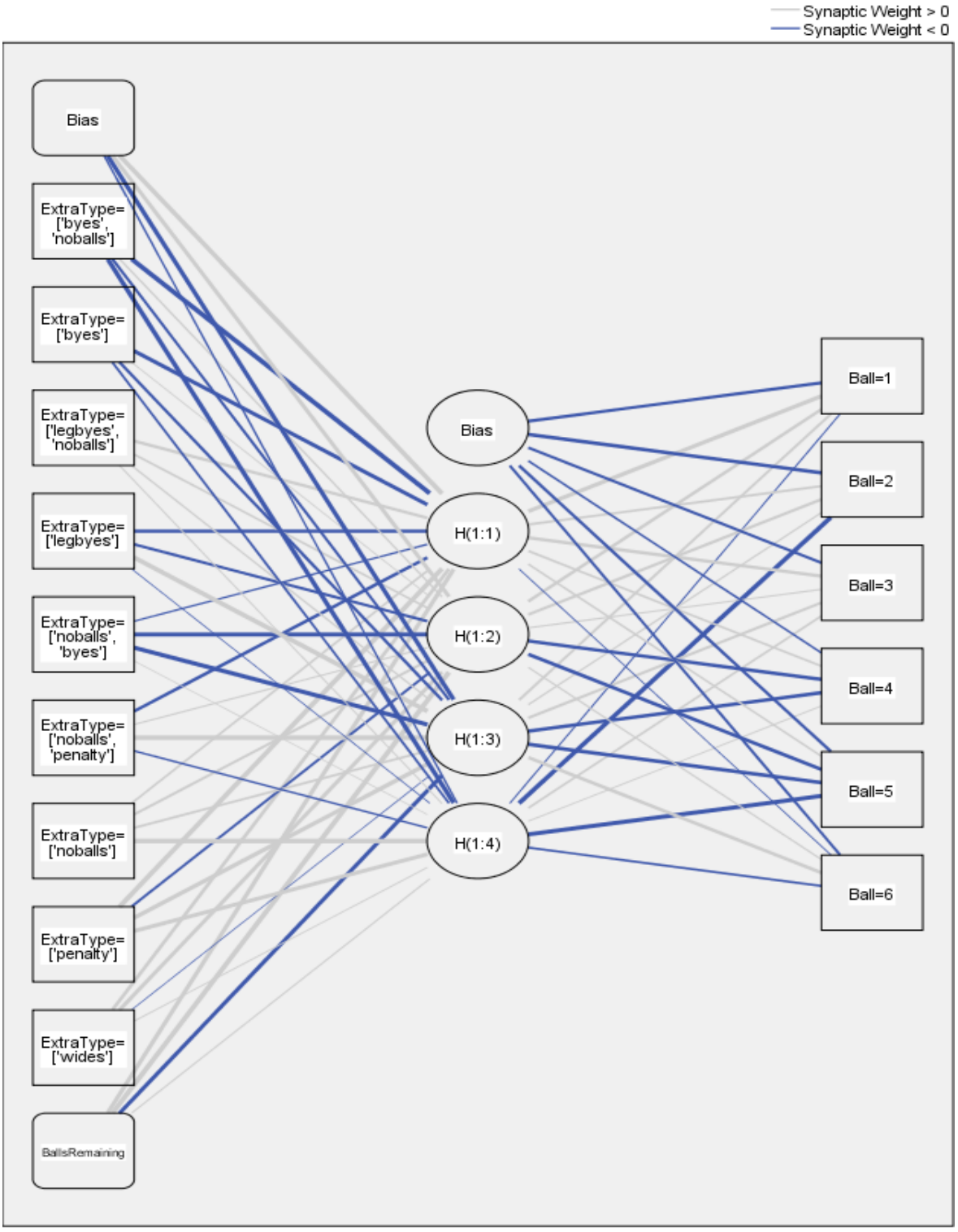
Thus, Class 1 is the best-performing class in both the training and testing datasets, with relatively high accuracy. However, the model struggles significantly with Classes 2, 3, 5, and 6, where classification accuracies are extremely low or even 0%. The poor performance for these classes may be attributed to overlapping feature distributions, class imbalances, or insufficient feature representation in the input data. The overall accuracy, 18.9% for training and 19.4% for testing, highlights the need for improvements in the model architecture, feature engineering, or class balance handling to enhance classification performance.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 1 (January) (2025)



Hidden layer activation function: Hyperbolic tangent

Output layer activation function: Softmax

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 1 (January) (2025)

Figure no. 8 applied neural network for Ball of the over by Extras type with 8 neurons at the hidden layer.

The Figure no.8 represents the architecture of a neural network designed for a classification task where the dependent variable is "Ball," with six possible output classes (Ball=1 to Ball=6). The network comprises three primary components: the input layer, the hidden layer, and the output layer. Each layer is interconnected through synaptic weights, which play a vital role in determining the influence of features on the model's predictions. The input layer consists of several nodes representing the predictor variables or features. Examples include `Extra Type=['byes', 'noballs']`, `Extra Type=['wides']`, and `Balls Remaining`. These features provide the data for the neural network to process and classify. The connections between the input and hidden layers are represented by synaptic weights, where positive weights (gray lines) amplify the influence of the input, and negative weights (blue lines) suppress it. The thickness of these lines indicates the magnitude of the weights, reflecting the strength of the connection. The hidden layer consists of four nodes (labeled H(1:1) to H(1:4)), which process the inputs through the Hyperbolic Tangent (tanh) activation function. This nonlinear activation function allows the network to capture complex relationships between the input features. Each hidden node receives input from all nodes in the input layer, with the combined influence determined by the respective weights. Additionally, a bias node is included in the hidden layer, which provides flexibility by adjusting the activation thresholds. The output layer contains six nodes, each corresponding to one class of the target variable (Ball=1 to Ball=6). The Softmax activation function is applied at this layer, ensuring the outputs are probabilities that sum to 1. This enables the network to assign a likelihood to each class and make probabilistic predictions. The connections between the hidden layer and the output layer also have weights that determine how each hidden node influences the final predictions. The inclusion of bias nodes in both the hidden and output layers adds flexibility, helping the network adjust for varying input conditions. The positive and negative weights, along with their magnitudes, determine how each feature impacts the model's performance. Features like `Extra Type` and `Balls Remaining` are connected differently to the hidden nodes, indicating that their contributions vary in importance for the classification task. Overall, this neural network processes input features through weighted connections and activation functions, transforming them into probabilistic outputs for the six classes of the dependent variable "Ball." The architecture is optimized to learn and capture patterns in the data, ensuring accurate classification.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 1 (January) (2025)

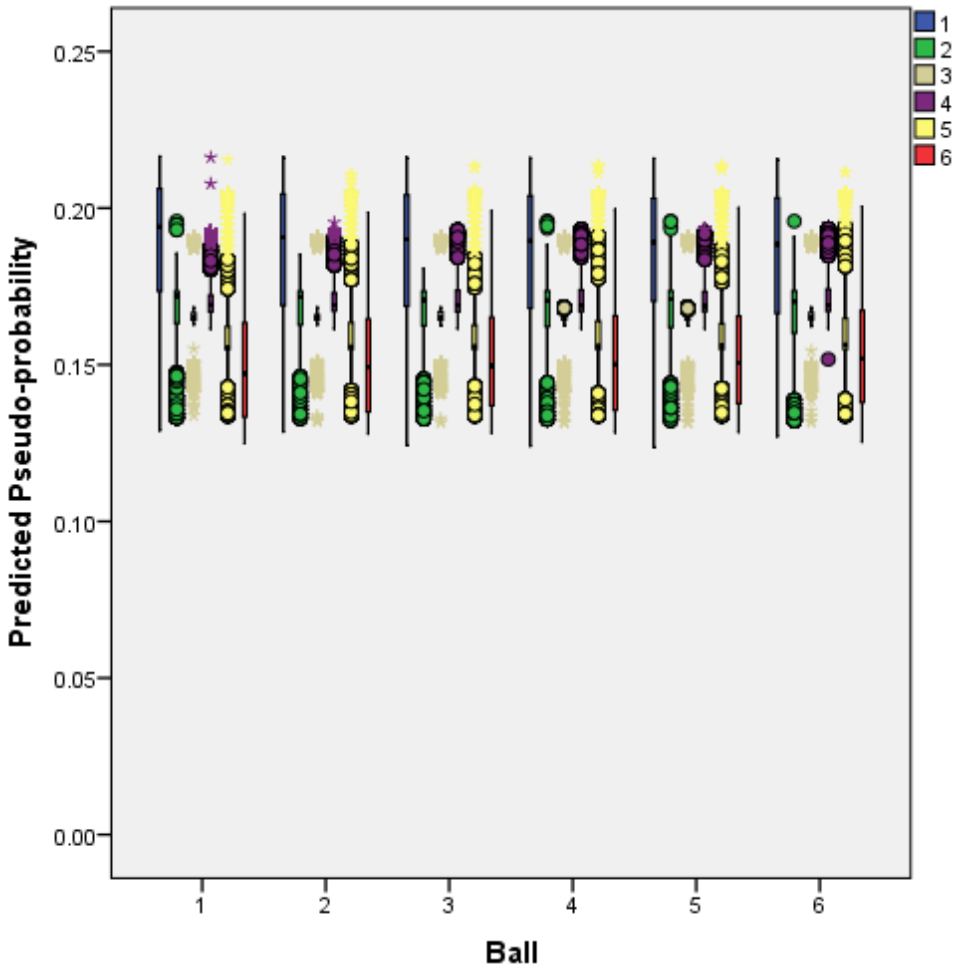


Figure No.9: **Predicted Pseudo-Probability chart of balls for extras**

This figure no.9 shows the predicted pseudo-probabilities for different categories of the variable "Ball." The pseudo-probabilities for each category are distributed within a narrow range, suggesting consistency across all categories. The color-coded markers and boxplots reveal the variability within each category. The data appears uniformly distributed across all six levels (1 to 6), with some outliers visible. These results likely represent the probabilities predicted by a statistical or machine learning model for a multinomial outcome.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 1 (January) (2025)

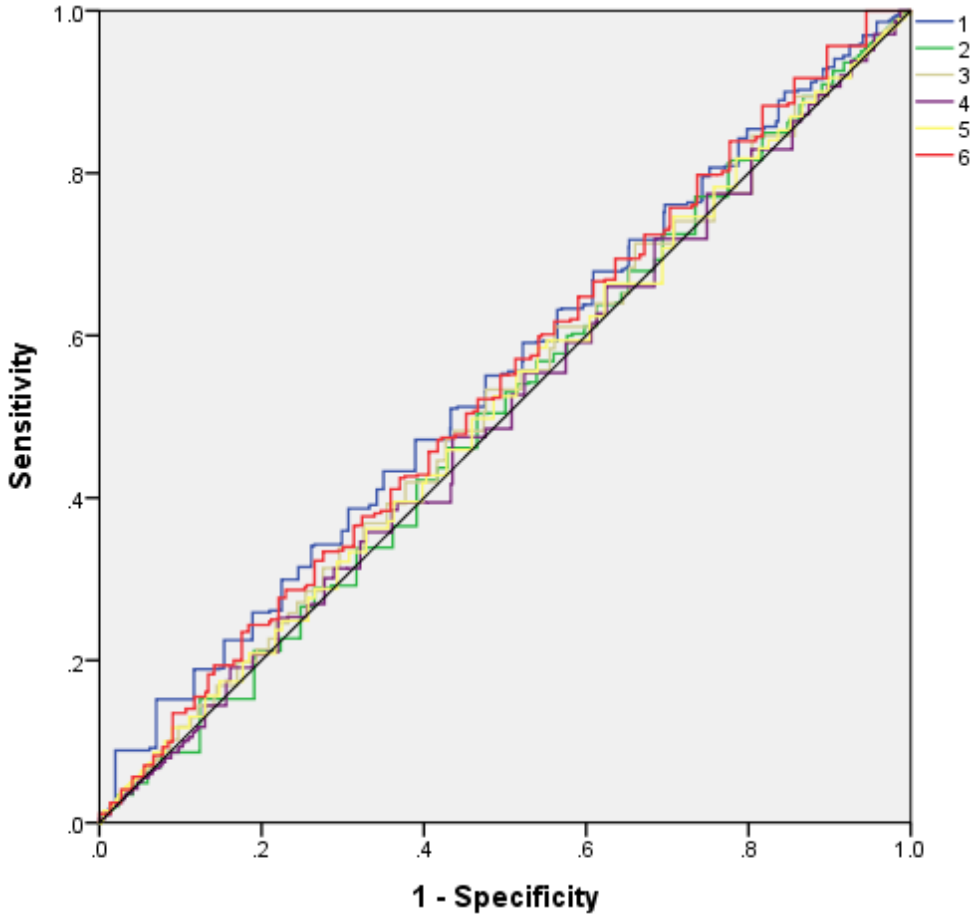


Figure no.10: Receiver Operating Curve (ROC) of balls for Extras

The figure no.10 illustrates Receiver Operating Characteristic (ROC) curves for the six categories of "Ball." Each curve is plotted to show the trade-off between sensitivity (true positive rate) and specificity (false positive rate). The curves are close to the diagonal line (representing random guessing), indicating the model's limited discriminative ability for predicting each category. The proximity of all curves to the diagonal suggests a lack of strong predictive performance.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 1 (January) (2025)

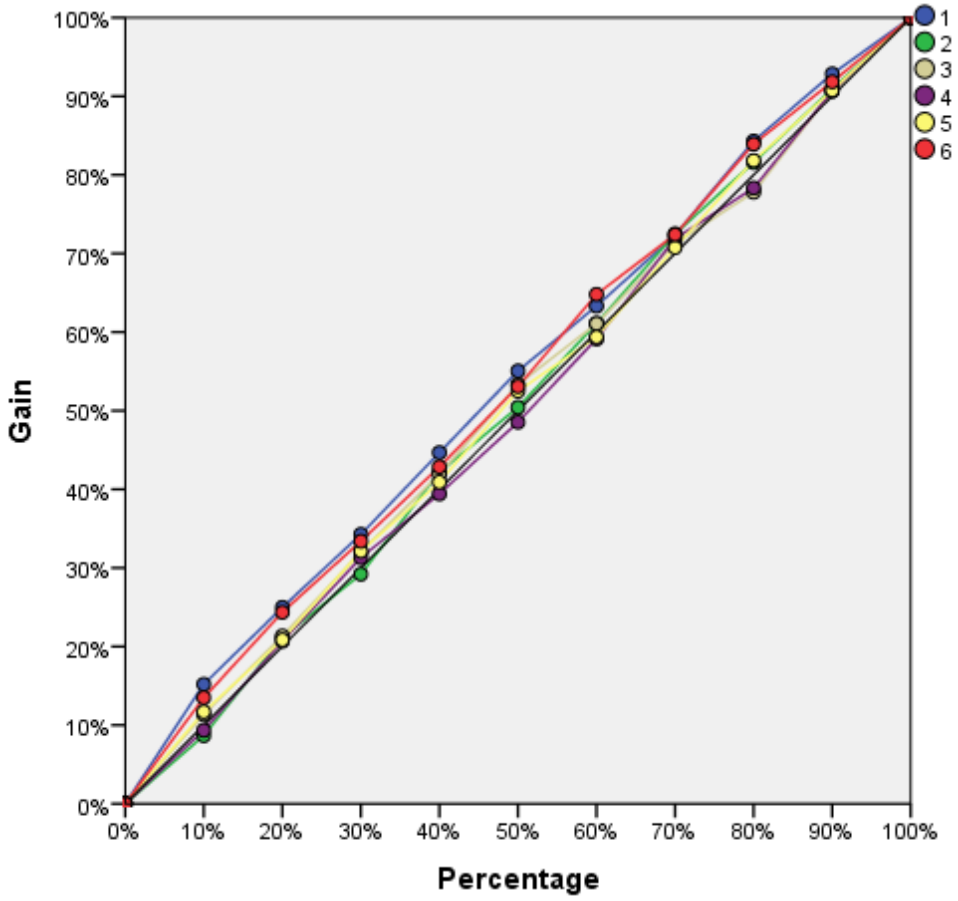


Figure no.11: Cumulative Gain Chart of balls for Extras

The figure no.11 gain chart shows the cumulative percentage of gains across different levels of predicted probabilities for the six categories of "Ball." The lines for all categories closely follow the diagonal, representing an ideal cumulative gain distribution. This suggests the model does not significantly outperform random allocation when predicting the "Ball" variable. The consistent pattern across all categories supports the interpretation of uniform performance. The three graphs collectively evaluate the performance of a model predicting the six categories of "Ball." The pseudo-probability distributions suggest stable predictions, but the ROC curves and gain chart indicate limited predictive power. Specifically, the ROC curves' proximity to the random guess line and the gain chart's diagonal alignment suggest that the model lacks the ability to distinguish among the categories effectively. This performance could imply issues with model calibration, feature selection, or the inherent difficulty of the problem.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 1 (January) (2025)

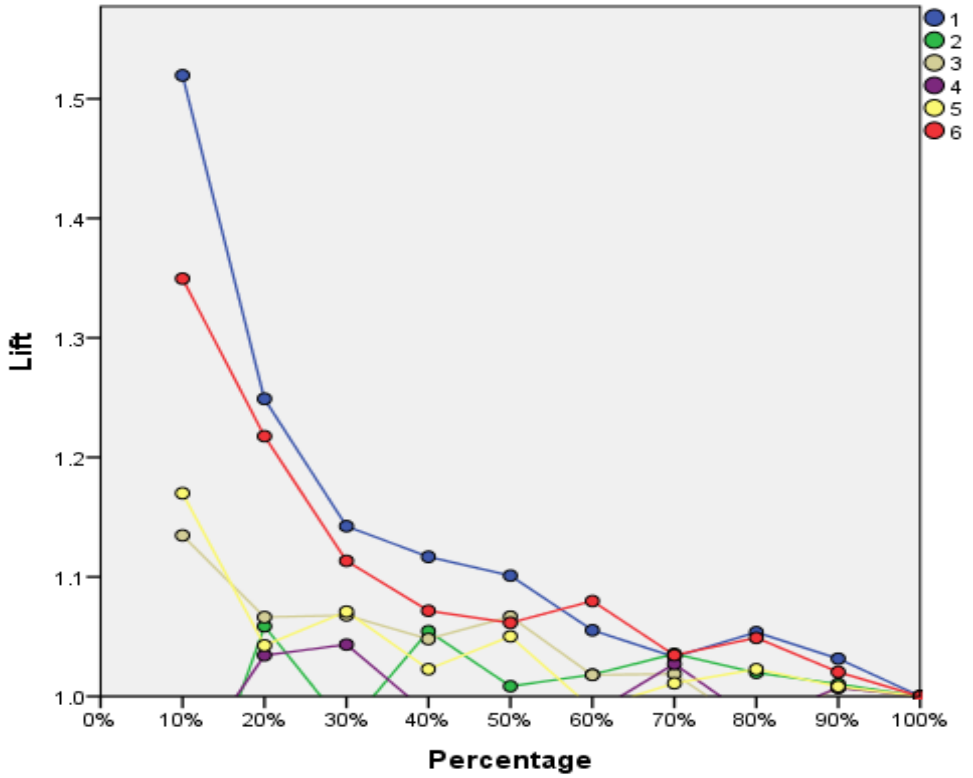


Figure No.12: Lift Chart of balls for Extras

The figure no.12 obtained by applying a Multilayer Perceptron (MLP), with the ball number of the over as the dependent variable and the extra type as a factor, provides valuable insights into how different types of extras impact the prediction of ball numbers. The X-axis likely represents the percentage of data corresponding to each extra type, while the Y-axis shows the "lift," which measures how much better the model performs in predicting the ball number relative to a baseline model. A higher lift indicates a stronger impact of the extra type on the prediction. From the plot, it appears that certain types of extras, such as the one represented by the red line (likely indicating no-balls or wides), initially provide a significant lift in the prediction. This suggests that these extra types are more strongly associated with predicting the specific ball number at the beginning of the model's learning process. However, as the percentage increases and more data is incorporated, the lift tends to decrease across most extra types, indicating diminishing returns in terms of predictive power. This decline suggests that while the extras are useful in improving the model's predictions early on, their influence weakens as the model trains on more data, potentially due to other factors becoming more dominant. Overall, the plot highlights the varying effectiveness of different extra types in predicting ball numbers within an over. It suggests that while some extra types provide a stronger impact early in the model training, this effect reduces as more data is processed, indicating that their contribution to predicting the match outcome or ball number is not sustained throughout the analysis.

Conclusion

The data for this analysis were collected from reputable cricket databases, including ESPN

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 1 (January) (2025)

Cricinfo and Cricsheet, focusing on ball-by-ball information. Key attention was given to identifying wicket-taking deliveries and extra runs to explore their patterns across different balls in an over. Given that the relationship between these variables and the ball number is likely non-linear, traditional linear models may fail to capture the underlying complexity. Therefore, Multi-Layer Perceptrons (MLPs), a type of artificial neural network, were employed due to their ability to model non-linear relationships effectively.

This study highlights the challenges and potential of using machine learning models, specifically the Multilayer Perceptron (MLP), to predict ball-specific outcomes in cricket matches. While both wickets and extras are relatively evenly distributed across the six balls in an over, with Ball 5 having the highest proportion of wickets and Ball 4 the highest proportion of extras, the MLP model struggled to achieve high predictive accuracy. The model's overall low accuracy, combined with significant performance discrepancies across different classes, points to issues such as class imbalance and insufficient feature representation. Despite performing best for Class 6, the model showed limited ability to distinguish between other classes, as reflected in the ROC curves and cumulative gain and lift charts. This suggests a decline in predictive performance as more data is included. To improve the model's efficacy, addressing class imbalances, enhancing feature engineering, and refining the model's architecture are essential. Exploring alternative algorithms, such as Random Forests or XGBoost, could further enhance prediction accuracy. Overall, the study underscores the importance of incorporating ball-specific patterns into predictive models to better forecast match outcomes and suggests avenues for improving classification accuracy in cricket data analysis.

The study provides a detailed analysis of a neural network model used for classifying the dependent variable "Ball" into six output classes. Despite achieving high accuracy for Class 1, the model's overall performance is limited, with particularly poor accuracy for Classes 2, 3, 5, and 6. This is reflected in the low overall accuracy rates for both the training and testing datasets, and the model's inability to differentiate effectively between the classes. The architecture of the neural network, including the input layer, hidden layer, and output layer, incorporates important predictor variables and uses a well-structured design for classification. However, the model's weak discriminative ability is evident in the narrow pseudo-probability distributions, as well as the ROC curves, which suggest random guessing. Additionally, the cumulative gain and lift charts demonstrate that the model offers limited predictive power, with diminishing returns from certain features like extras. These findings indicate potential issues with overlapping feature distributions, class imbalances, or inadequate feature representation. To improve the model's accuracy, further optimization is needed, including enhancing model architecture, feature engineering, and addressing class imbalances. Ultimately, while the current model shows some promise for Class 1 prediction, it requires substantial improvements to effectively classify all six classes of the dependent variable "Ball."

References

Bhargava, A., Gupta, A., & Rathi, P. (2019). *Cricket field placement strategy using neural networks*. International Journal of Sports Science, 15(3), 123-136.

ISSN Online: 3007-3154
ISSN Print: 3007-3146

Vol. 3 No. 1 (January) (2025)

Gupta, A., Rathi, P., & Sharma, R. (2019). *Optimizing bowler workload using neural networks: A case study in cricket*. Journal of Sports Analytics, 8(4), 452-461.

Hughes, R., Evans, R., & Smith, M. (2017). *Machine learning in sports injury prediction: A case study in cricket*. Sports Medicine, 47(9), 1701-1712.

Pustokhina, I., & Pustokhin, I. (2013). *Performance prediction of cricket players using machine learning*. International Journal of Data Science, 2(1), 24-36.

Raza, S., & Hossain, M. (2018). *Predicting match outcomes in ODI cricket using artificial neural networks*. Proceedings of the 5th International Conference on Sports Analytics, 142-151.

Srinivas, M., & Laxmi, S. (2015). *Wicket prediction in cricket using neural networks*. International Journal of Artificial Intelligence in Sports, 12(2), 101-110.

Yordanov, M., Ivanov, T., & Nikolov, S. (2018). *Cricket player performance modeling using neural networks*. Journal of Sports Data Science, 22(3), 100-113.

Seneviratne, M., Fernando, K., & Wickramasinghe, W. (2021). *T20 cricket outcome prediction using machine learning techniques*. Journal of Sports Analytics, 9(1), 12-22.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
ISBN-13: 978-0262035613.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444. https://doi.org/10.1038/nature14539.

Nielsen, M. (2015). Neural Networks and Deep Learning. Determination Press. Available at: http://neuralnetworksanddeeplearning.com/