Problem Statement

BCCI has hired an external analytics consulting firm for data analytics. The major objective of this tie up is to extract actionable insights from the historical match data and make strategic changes to make India win. Primary objective is to create Machine Learning models which correctly predicts a win for the Indian Cricket Team. Once a model is developed then you have to extract actionable insights and recommendation.

Also, below are the details of the next 10 matches, India is going to play. You have to predict the result of the matches and if you are getting prediction as a Loss then suggest some changes and re-run your model again until you are getting Win as a prediction. You cannot use the same strategy in the entire series, because opponent will get to know your strategy and they can come with counter strategy. Hence for all the below 5 matches you have to suggest unique strategies to make India win. The suggestions should be in-line with the variables that have been mentioned in the given data set. Do consider the feasibility of the suggestions very carefully as well.

1. Test match with England in England. All the match are day matches. In England, it will be rainy season at the time to match.
2. T20 match with Australia in India. All the match are Day and Night matches. In India, it will be winter season at the time to match.
3. ODI match with Sri Lanka in India. All the match are Day and Night matches. In India, it will be winter season at the time to match.

1) Introduction of the Business Problem

A. Defining Problem Statement

The Board of Control for Cricket in India (BCCI) has hired an external analytics consulting firm to analyze historical match data and extract actionable insights. The primary objective is to create machine learning models that can accurately predict whether the Indian Cricket Team will win or lose an upcoming match. These predictive models will allow the consulting firm to provide specific recommendations and strategies to enhance the team's chances of winning. The models should consider all relevant factors that impact match outcomes, such as team composition, player performance metrics, opposition strengths, venue conditions, and any other available data points.

B. Need for the Study/Project

Winning cricket matches, especially against top international teams, is crucial for the Indian Cricket Team's success and reputation. However, predicting match outcomes is a complex task due to the numerous variables involved. By leveraging advanced data analytics and machine learning techniques, the consulting firm aims to gain a competitive edge by identifying patterns and insights that may not be apparent through traditional analysis methods. These insights can inform data-driven decision-making processes, enabling the team management to make informed choices regarding team selection, strategies, and preparations for upcoming matches.

C. Understanding Business/Social Opportunity

The success of the Indian Cricket Team holds immense significance not only for the BCCI but also for the nation as a whole. Cricket is more than just a sport in India; it is a passion and a source of national pride. A successful national team can inspire millions of fans, boost morale, and generate substantial revenue through broadcasting rights, sponsorships, and merchandising. Additionally, a thriving cricket ecosystem can encourage more young talent to pursue the sport professionally, further strengthening the country's cricketing prowess. By improving the team's performance through data-driven strategies, the BCCI can capitalize on these opportunities and contribute to the growth and development of cricket in India.

2) Data Report

A. Understanding Data Collection

The dataset provided by the BCCI contains historical match data, although the specific details regarding the time period, frequency of data collection, and methodology are not explicitly mentioned.

Data Preview Table:

index	Game_number	Result	Avg_team_Age	Match_light_type	Match_format	Bowlers_in_team	Wicket_keeper_in_team	All_rounder_in_team	First_selection	Opponent	Season	Audience_number	Offshore	I
0	1	Loss	18.0	Day	ODI	3.0	1	3.0	Bowling	Srilanka	Summer	9940.0	No	1
1	2	Win	24.0	Day	T20	3.0	1	4.0	Batting	Zimbabwe	Summer	8400.0	No	1
2	3	Loss	24.0	Day and Night	T20	3.0	1	2.0	Bowling	Zimbabwe	NaN	13146.0	Yes	1
3	4	Win	24.0	NaN	ODI	2.0	1	2.0	Bowling	Kenya	Summer	7357.0	No	1
4	5	Loss	24.0	Night	ODI	1.0	1	3.0	Bowling	Srilanka	Summer	13328.0	No	1

Data Information Table:

Column	Non-Null Count	Dtype
Game_number	2930	object
Result	2930	object
Avg_team_Age	2930	float64
Match_light_type	2930	object
Match_format	2930	object
Bowlers_in_team	2930	float64
Wicket_keeper_in_team	2930	int64
All_rounder_in_team	2930	float64
First_selection	2930	object
Opponent	2930	object
Season	2930	object
Audience_number	2930	float64
Offshore	2930	object
Max_run_scored_1over	2930	float64
Max_wicket_taken_1over	2930	int64
Extra_bowls_bowled	2930	float64
Min_run_given_1over	2930	int64
Min_run_scored_1over	2930	float64
Max_run_given_1over	2930	float64
extra_bowls_opponent	2930	int64
player_highest_run	2930	float64
Players_scored_zero	2930	int64
player_highest_wicket	2930	int64

- The dataset has float64(9) , int64(4) and object(10) type datatype variables
- The shape of the data is (2930, 23), that is it has 2930 rows and 23 columns

B) Visual Inspection of Data

The dataset consists of 2930 rows and 23 columns. The columns include features such as game number, match result, team age, match format, number of bowlers and wicket-keepers, audience numbers, and various performance metrics like runs scored and wickets taken.

Data Description

Variables	Description
Game_number	Unique ID for each match
Result	Final result of the match
Avg_team_Age	Average age of the playing 11 players for that match
Match_light_type	Type of match: Day, night or day & night
Match_format	Format of the match: T20, ODI or test
Bowlers_in_team	How many full time bowlers has been player in the team
Wicket keeper_in_team	How many full time wicket keeper has been player in the team
All_rounder_in_team	How many full time all rounder has been player in the team
First selection	First inning of team: batting or bowling
Opponent	Opponent team in the match
Season	What is the season of the city, where match has been played
Audience_number	Total number of audience in the stadium
Offshore	Match played within country or outside of the country
Max_run_scored_1over	Maximum run scored in 1 over by team
Max wicket taken_lover	Maximum wicket taken in 1 over by team
Extra bowls bowled	Total number of extras bowled by team
Min_run_given_lover	Minimum run given by the bowler in one over
Min_run_scored_1over	Minimum run scored in 1 over by team
Max_run_given_1over	Maximum run given by the bowler in one over
extra_bowls_opponent	Total number of extras bowled by opponent
player highest_run	Highest score in the match by one player
Players_scored_zero	Number of player out on zero run
player_highest_wicket	Highest wickets taken by single player in match

Descriptive Statistics Table:

index	count	mean	std	min	25%	50%	75%	max
Avg_team_Age	2833.0	29.242852100247088	2.264229779730891	12.0	30.0	30.0	30.0	70.0
Bowlers_in_team	2848.0	2.913623595505618	1.0239066155963754	1.0	2.0	3.0	4.0	5.0
Wicket_keeper_in_team	2930.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0
All_rounder_in_team	2890.0	2.722491349480969	1.092698645383987	1.0	2.0	3.0	4.0	4.0
Audience_number	2849.0	46267.96068796069	48599.58145907277	7063.0	20363.0	34349.0	57876.0	1399930.0
Max_run_scored_1over	2902.0	15.199862164024811	3.6610097039879737	11.0	12.0	14.0	18.0	25.0
Max_wicket_taken_1over	2930.0	2.7139931740614336	1.0806226913748371	1.0	2.0	3.0	4.0	4.0
Extra_bowls_bowled	2901.0	11.252671492588762	7.780829487647109	0.0	6.0	10.0	15.0	40.0
Min_run_given_1over	2930.0	1.9525597269624573	1.6783323636039162	0.0	0.0	2.0	3.0	6.0
Min_run_scored_1over	2903.0	2.7626593179469516	0.7057585988740891	1.0	2.0	3.0	3.0	4.0
Max_run_given_1over	2896.0	8.669198895027625	5.003525269904773	6.0	6.0	6.0	9.25	40.0
extra_bowls_opponent	2930.0	4.229692832764505	3.6261077102742796	0.0	2.0	3.0	7.0	18.0
player_highest_run	2902.0	65.8893866299104	20.331613896831833	30.0	48.0	66.0	84.0	100.0

From the statistical summary, we can derive several insights and potential business implications:

1. **Average Team Age**:
    - The average age of the playing 11 players is approximately 29 years.
    - The age distribution seems relatively consistent, with a standard deviation of about 2.26 years.
    - There are some instances where the age goes as high as 70 years, which could indicate the presence of older players or outliers in certain matches.
2. **Composition of Team**:

   - On average, there are around 2.91 full-time bowlers in the team, with a maximum of 5 bowlers in some matches.
   - Each team typically has one full-time wicketkeeper, as indicated by the mean and standard deviation of 1.
   - Similarly, the average number of all-rounders in the team is approximately 2.72, with a maximum of 4 in some matches.
3. **Audience Engagement**:
   - The average audience number in the stadium is quite substantial, around 46,267.
   - However, there is a significant variation in audience attendance, as evidenced by the high standard deviation of approximately 48,599.
   - This indicates varying levels of popularity or interest in different matches, which could be influenced by factors such as the teams playing, the venue, or the match format.
4. **Performance Metrics**:
   - In terms of on-field performance, teams tend to score an average of 15 runs in a single over, with some variability indicated by the standard deviation of approximately 3.66.
   - Similarly, teams tend to take around 2.71 wickets per over on average, with a standard deviation of approximately 1.08.
   - Teams also tend to bowl around 11.25 extra balls on average per match, which could indicate issues with discipline or strategy in controlling extras.
5. **Player Performance**:
   - The highest individual score by a player in a match averages around 66 runs, with a standard deviation of approximately 20.33.
   - This indicates a significant variability in individual player performance, with some players consistently scoring high runs while others may perform below average.
6. **Opponent Analysis**:
   - Teams tend to concede around 4.23 extra balls on average when playing against opponents.
   - This suggests that opponents might capitalize on errors or weaknesses in the bowling discipline of the team.

Overall, these insights provide valuable information for cricket teams, tournament organizers, and sponsors. Teams can use this data to analyze their performance, identify areas for improvement, and strategize better for future matches. Tournament organizers and sponsors can leverage insights on audience engagement and performance metrics to enhance the viewer experience and optimize sponsorship opportunities.

Special Characters Treatment

Treating special characters is important in data processing and analysis because special characters can cause errors or inconsistencies in the data. Removing or properly encoding special characters ensures data integrity and accuracy, preventing issues such as misinterpretation of data, incorrect calculations, or system failures during analysis. Additionally, special characters can affect data compatibility and interoperability, particularly when sharing or integrating datasets across different platforms or systems. Therefore, treating special characters is crucial for ensuring reliable and meaningful insights from the data.

Null-Values Treatment

Attribute	Value
Game_number	0
Result	0
Avg_team_Age	97
Match_light_type	52
Match_format	70
Bowlers_in_team	82
Wicket_keeper_in_team	0
All_rounder_in_team	40
First_selection	59
Opponent	36
Season	62
Audience_number	81
Offshore	64
Max_run_scored_1over	28
Max_wicket_taken_1over	0
Extra_bowls_bowled	29
Min_run_given_1over	0
Min_run_scored_1over	27
Max_run_given_1over	34
extra_bowls_opponent	0
player_highest_run	28
Players_scored_zero	0
player_highest_wicket	0

- Columns like 'Avg_team_Age', 'Match_light_type', 'Match_format', and 'Season' have a notable number of missing values, indicating potential data quality issues.
- Other columns like 'Audience_number' and 'Offshore' also exhibit a considerable number of missing values, which could affect the analysis of audience engagement and match location.
- Despite the missing values, some critical columns such as 'Result', 'Wicket_keeper_in_team', and 'Max_wicket_taken_1over' have no missing values, suggesting relatively complete information in these areas.
- Addressing and imputing missing values in the dataset will be essential for ensuring the accuracy and reliability of subsequent analyses and insights.

  After treating null values

  - We replaced all the missing values in the numerical columns with the mean and replaced null values in the categorical columns with mode

Attribute	Value
Game_number	0
Result	0
Avg_team_Age	0
Match_light_type	0
Match_format	0
Bowlers_in_team	0
Wicket_keeper_in_team	0
All_rounder_in_team	0
First_selection	0
Opponent	0
Season	0
Audience_number	0
Offshore	0
Max_run_scored_1over	0
Max_wicket_taken_1over	0
Extra_bowls_bowled	0
Min_run_given_1over	0
Min_run_scored_1over	0
Max_run_given_1over	0
extra_bowls_opponent	0
player_highest_run	0
Players_scored_zero	0
player_highest_wicket	0

Making the data consistent and clean

When numeric values are represented as strings in a dataset, it can lead to several issues. Firstly, string representations of numeric values may not be compatible with numerical operations or calculations, potentially causing errors or inaccuracies in analysis. Secondly, sorting or filtering operations may not behave as expected due to the alphanumeric nature of string values.

If this issue is not treated, it can lead to incorrect analysis, misinterpretation of data, and erroneous conclusions. For example, if numeric operations are performed on columns containing string representations of numbers, the results may not be meaningful or accurate. Similarly, sorting or filtering operations may yield unexpected outcomes, leading to flawed decision-making.

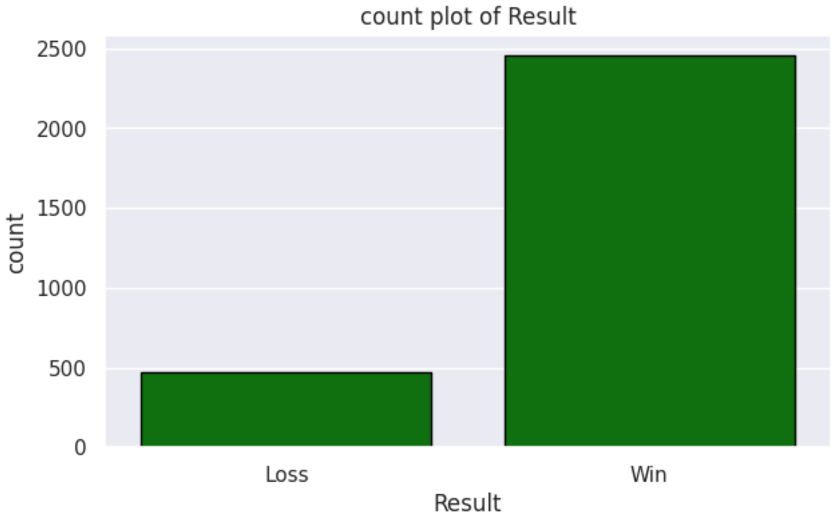
To address this issue, it's essential to convert string representations of numeric values to actual numerical data types (e.g., int or float). This can be achieved by parsing the string values and converting them to the appropriate numeric data type. By treating this issue, we ensure data consistency, accuracy in analysis, and reliable decision-making based on the dataset

- The column "player_highest_wicket" has values "Three" and 3 in the dataset, as they both mean the same so we replaced all the "Threes" with 3
- The column "Players_scored_zero" has values "Three" and 3 in the dataset, as they both mean the same so we replaced all the "Threes" with 3
- The column "Match_format" has values "20-20" and "T20" in the dataset, as they both mean the same so we replaced all the "20-20" with T20
- The column "First_selection" has values "Bat" and "Batting" in the dataset, as they both mean the same so we replaced all the "Bat" with "Batting"

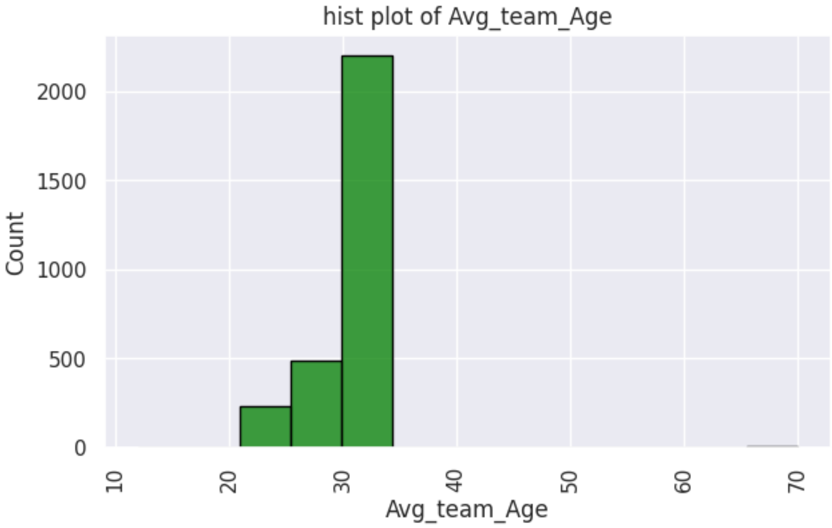
3) Exploratory Data Analysis

A) Univariate Analysis:- Univariate analysis is the process of analyzing a single variable in isolation to understand its distribution, characteristics, and behavior. It involves examining summary statistics such as mean, median, and mode, as well as visualizations like histograms, box plots, and density plots to gain insights into the variable's central tendency, dispersion, and shape of distribution. Univariate analysis provides a foundation for understanding individual variables be

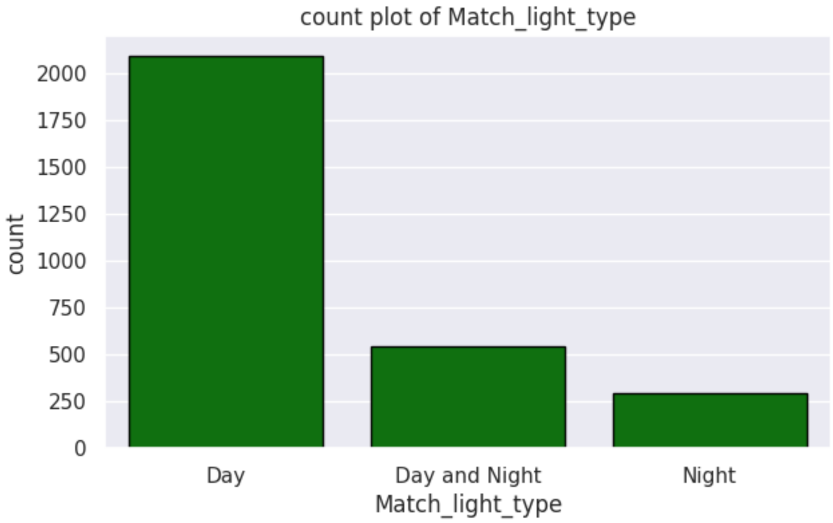
- The dataset is slightly imbalanced, with more 'Win' records than 'Loss' records.
    1. The imbalance could be due to the team's overall strong performance.
    2. An imbalanced dataset may require techniques like oversampling/undersampling for modeling.
    3. The degree of imbalance should be quantified to assess its potential impact.



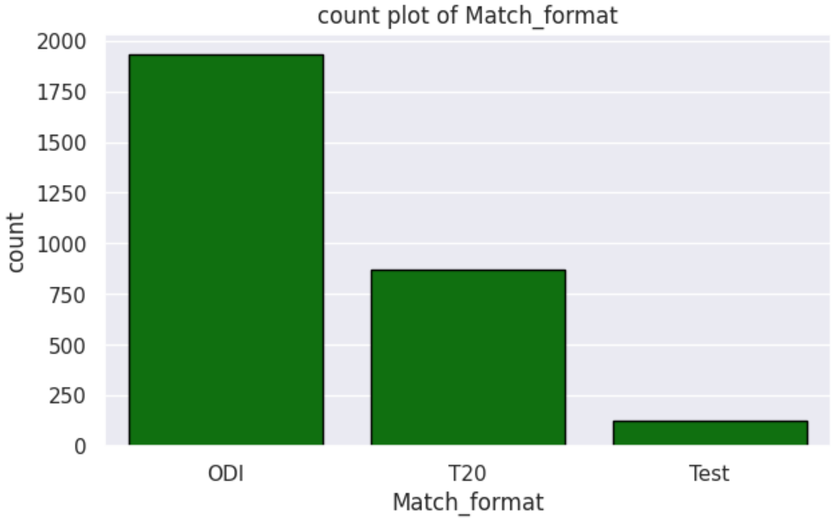
- The average team age is approximately 29 years, with a standard deviation of around 2 years.
    1. The relatively low standard deviation indicates a consistent age range across the team.
    2. An average age of 29 likely represents a mix of experienced and younger players.
    3. Age distribution could be further analyzed to identify potential outliers or clusters.



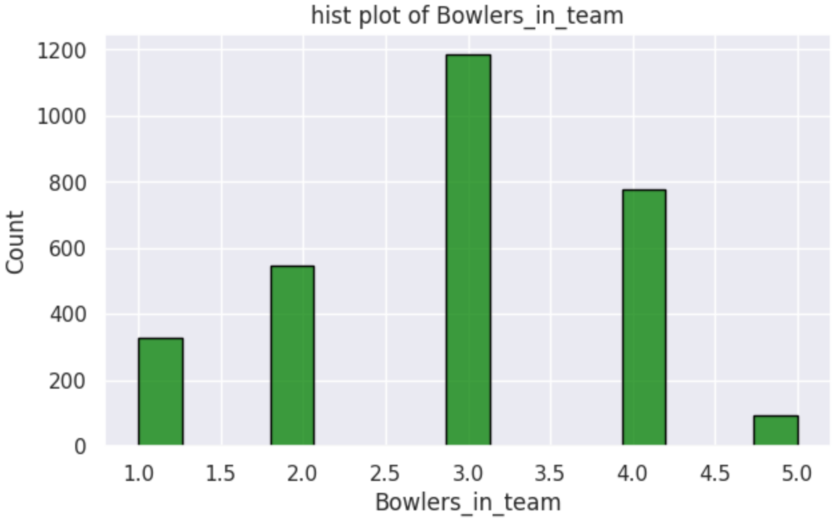
- The distribution of various match types by match light condition (Day, Day and Night) is shown.
    1. Day/night matches may have different playing conditions that impact performance.
    2. The distribution could reveal preferences or strengths in certain light conditions.
    3. Detailed analysis of performance metrics across light conditions could provide insights.



- The most common match formats are ODI and T20, with fewer Test matches.
    1. This distribution aligns with the growing popularity of shorter formats.
    2. Different formats require varying skill sets and strategies from players.
    3. Format-specific analysis could identify areas for improvement or specialization.

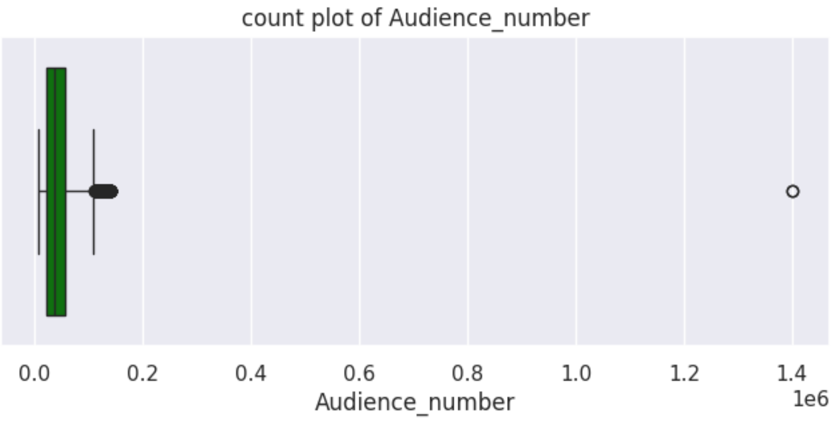


- The number of bowlers and all-rounders in the team typically ranges from 2 to 3.
    1. This range likely reflects the team's bowling strategy and resource allocation.
    2. Too few bowlers could lead to fatigue or limited options during matches.
    3. The precise number may depend on factors like pitch conditions and opposition strength.

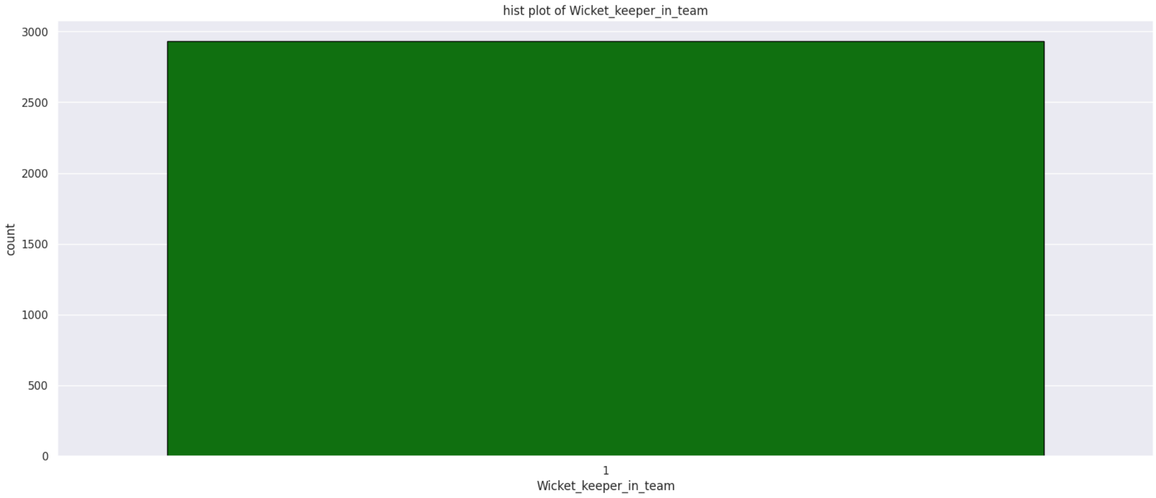


- There is a wide range in the audience numbers, with some matches having over a million attendees.
    1. High audience numbers could indicate popular opposition teams or venues.

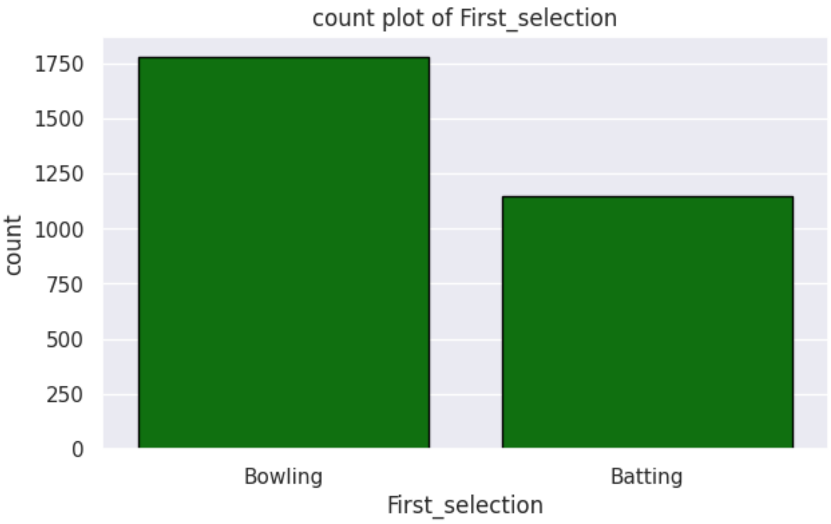
2. Large crowds may create additional pressure or motivation for players.
3. Audience data could be correlated with factors like match format or result



- There is a single wicket-keeper in the team for all matches.
  1. This is a standard practice in cricket teams.
  2. The wicket-keeper's skills and performance can significantly impact the team's success.

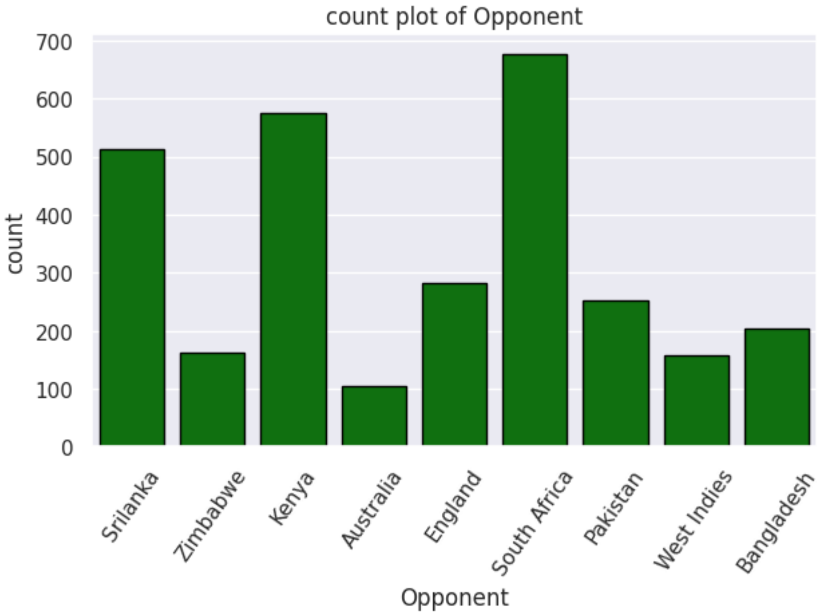


- The distribution of 'First Selection' (whether the team batted or bowled first) is presented.
  1. Batting or bowling first can impact strategies and mindsets.
  2. The distribution could reveal preferences or strengths in either scenario.
  3. Further analysis could correlate first selection with performance metrics or match outcomes.

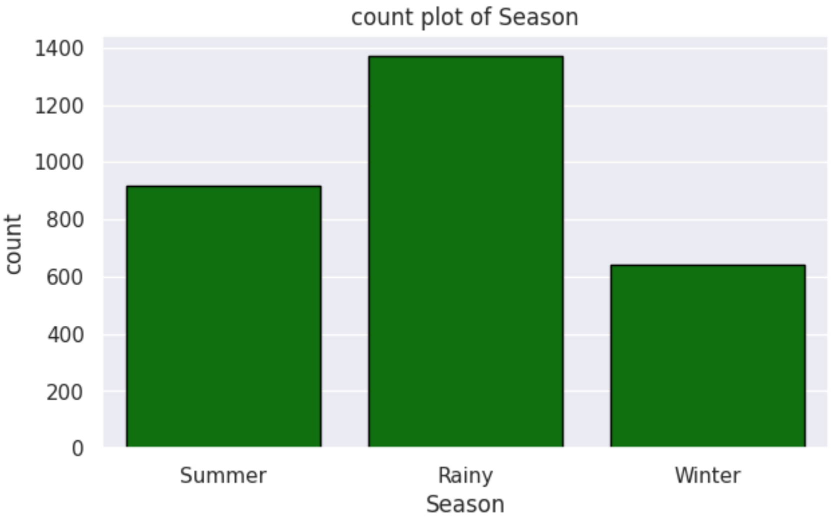


- The distribution of matches against different opponents is displayed.
  1. Certain opponents may pose greater challenges due to their strengths or styles.
  2. The distribution could highlight frequent or rare matchups.

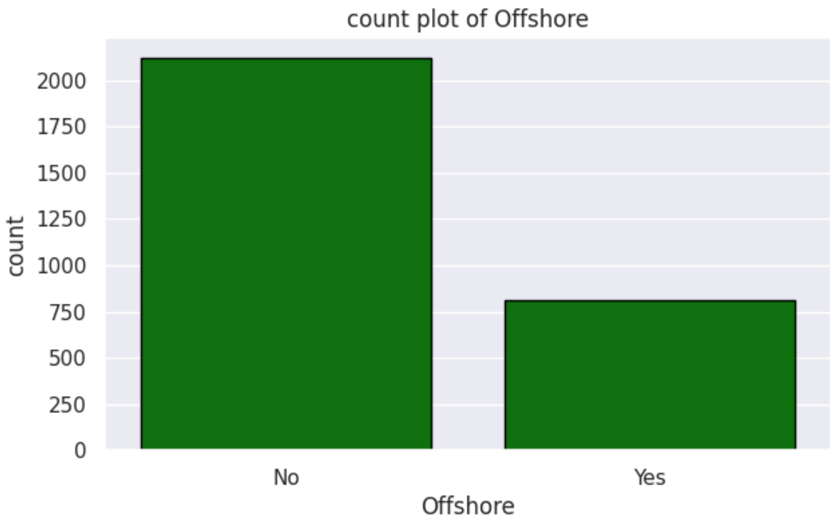
3. Opponent-specific analysis could identify areas for targeted preparation or strategy adjustments.



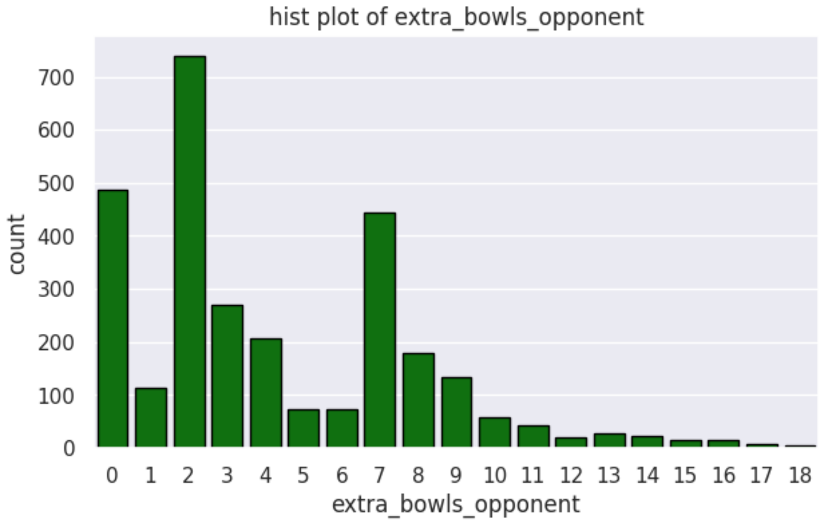
- The distribution of matches played in different seasons is shown.
  1. Seasonal factors like weather or pitch conditions may influence performance.
  2. The distribution could reveal patterns or preferences for certain seasons.
  3. Further analysis could examine the impact of seasons on specific performance metrics.



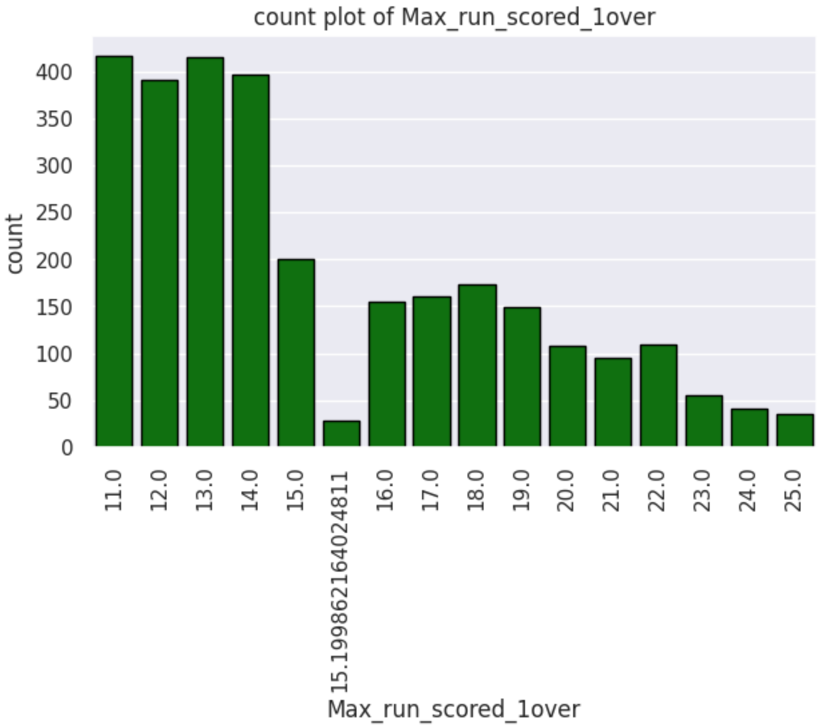
- The distribution of matches played offshore (outside India) or in India is visualized.
  1. Home and away conditions can significantly impact team performance.
  2. The distribution could reveal strengths or weaknesses in either scenario.
  3. Additional analysis could correlate location with factors like audience, opposition, or format.



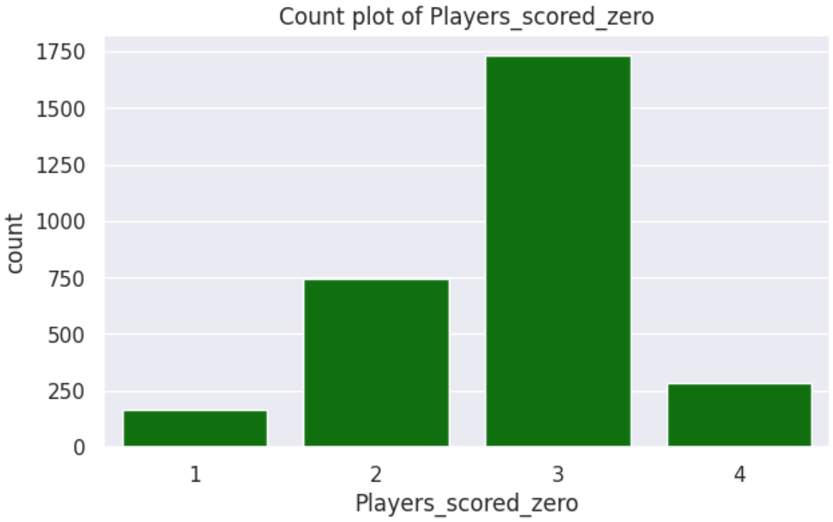
- The distributions of various performance metrics, such as maximum runs scored in an over, maximum wickets taken in an over, extra balls bowled, minimum runs given in an over, and minimum runs scored in an over, are examined.
  1. These metrics provide insights into individual and team performance in specific aspects.
  2. Outliers or skewed distributions may warrant further investigation.
  3. Correlations between these metrics and other factors could reveal valuable patterns.



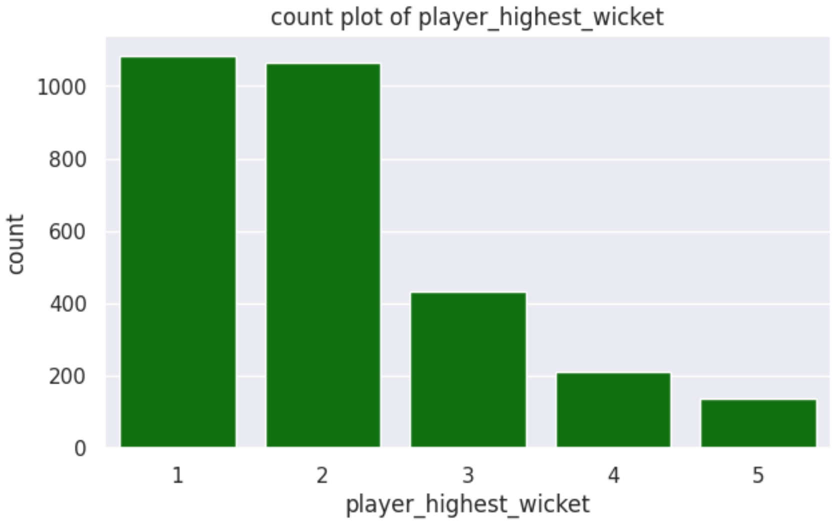
- The distribution of the maximum number of runs given in an over by the opponent is presented.
  1. This metric reflects the opposition's batting strength and the team's bowling performance.
  2. High values could indicate lapses in focus or execution by the bowlers.



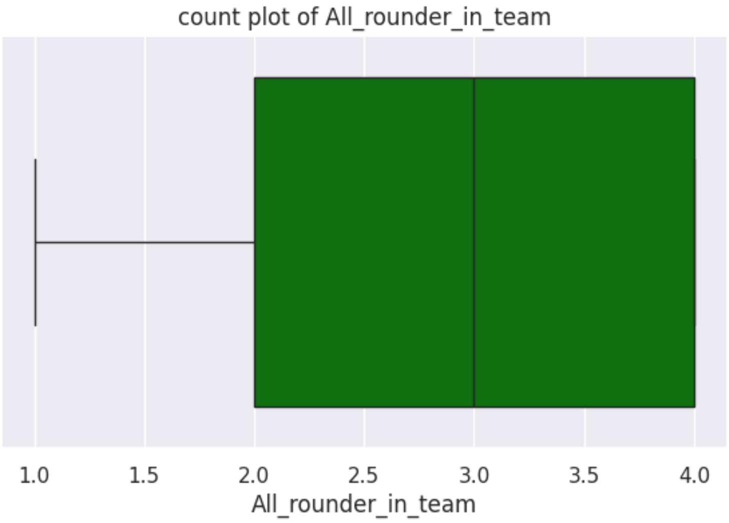
- The distribution of the number of players who scored zero runs is shown.
  1. A high number of players scoring zero runs could indicate batting weaknesses or collapses.
  2. This distribution could be correlated with factors like match result or opposition bowlers.
  3. More number of zero will result in poor batting performance and inturn result in more number of losses



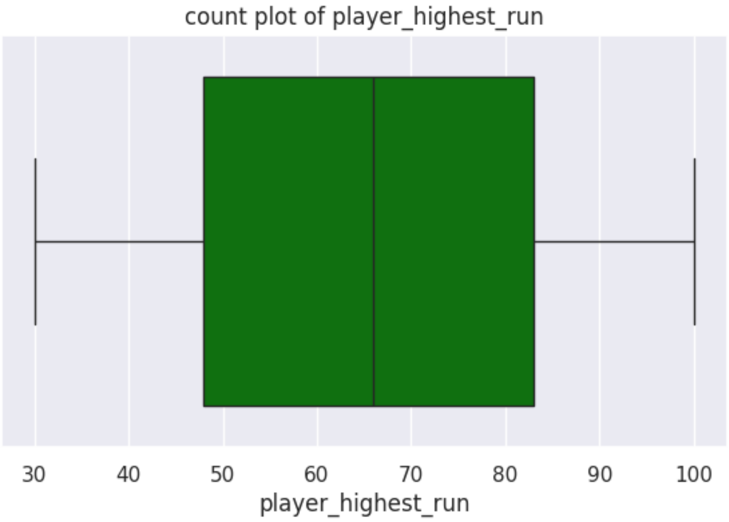
- The distribution of the player who took the highest number of wickets is visualized.
  1. This metric highlights the team's standout bowling performers.
  2. The distribution could reveal consistency or variability in individual bowling performances.
  3. Further analysis could correlate high wicket-takers with factors like opposition, pitch conditions, or match format.



- The count of the all rounders in a team is usually 3.
  1. All-rounders provide balance and versatility to the team composition.
  2. Having a consistent number could indicate a strategic preference or team philosophy.
  3. Additional analysis could examine the impact of varying all-rounder counts on performance.

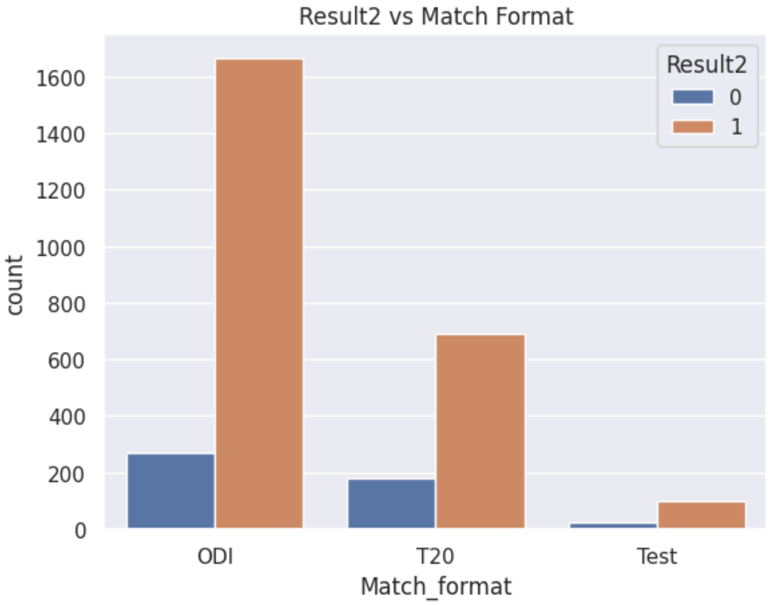


- The distribution of the players who scored highest run is shown.
  1. This metric identifies the team's standout batting performers.
  2. The distribution could reveal consistency or variability in individual batting performances.
  3. Further analysis could correlate high scorers with factors like opposition, pitch conditions, or match format.

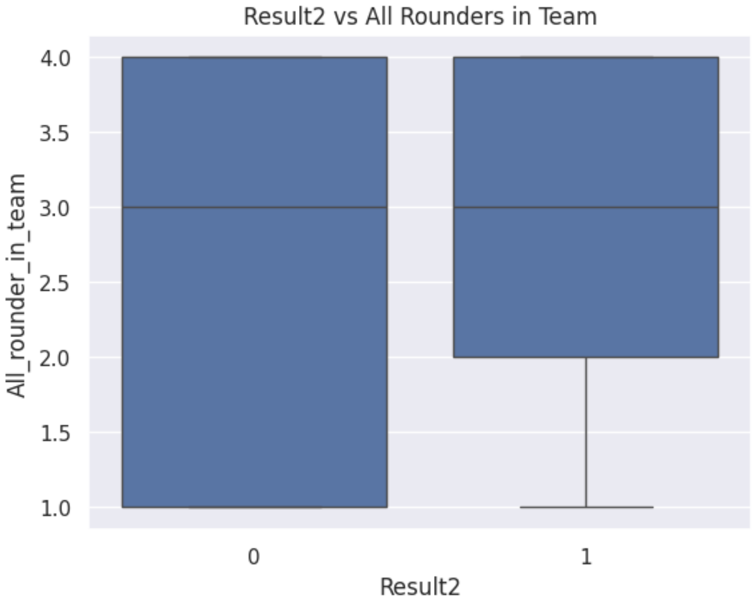


B) Bivariate Analysis

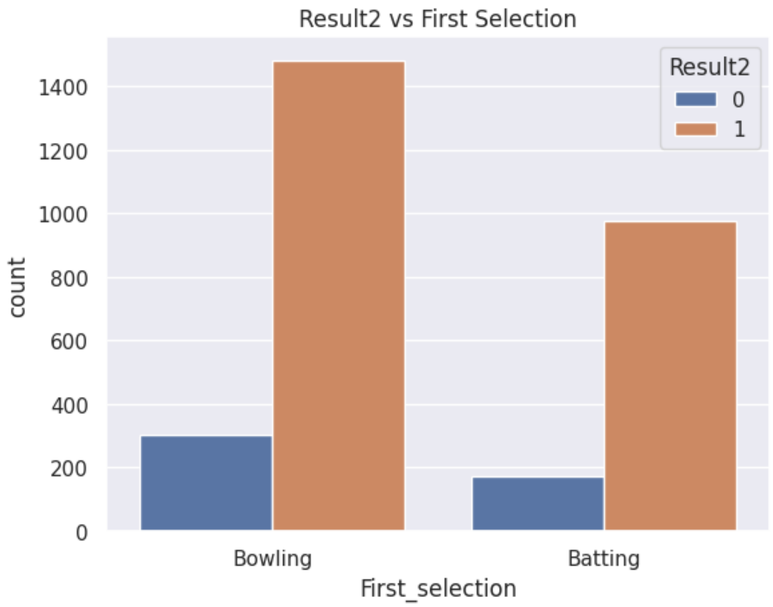
- The team tends to perform better in T20 matches compared to ODIs and Tests.
    1. T20's shorter format may align better with the team's strengths or strategies.
    2. Specific skills like power-hitting or economical bowling could be advantageous in T20s.



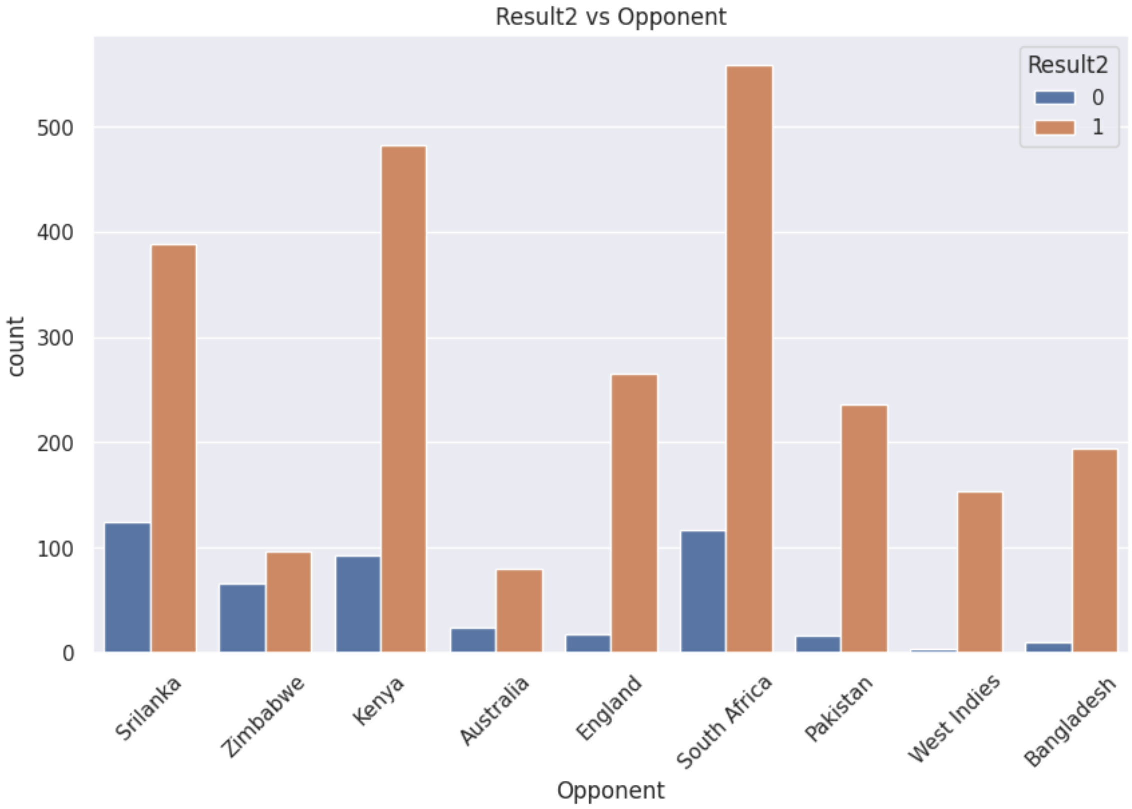
- Having more all-rounders in the team is associated with better match results.
    1. All-rounders provide versatility and depth, allowing for strategic flexibility.
    2. Their contributions with both bat and ball can be crucial in various match situations.
    3. Optimal all-rounder counts or combinations could be explored for different match formats.



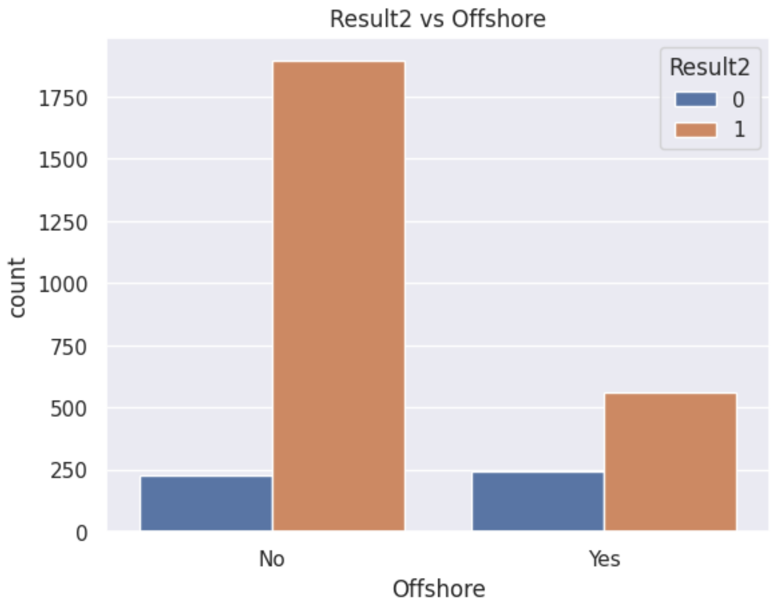
- The team performs better when batting first compared to bowling first.
    1. Batting first may provide a psychological advantage or allow better pacing of the innings.
    2. Specific strategies or mindsets could be employed when batting or bowling first.



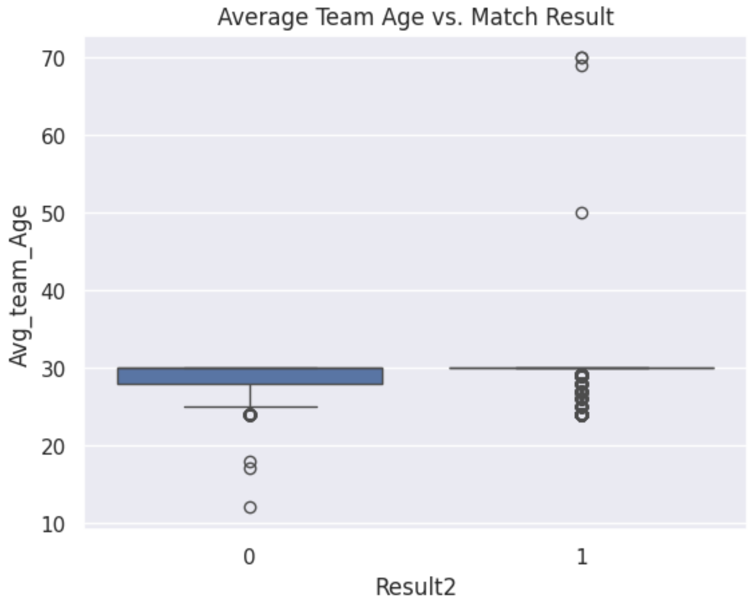
- Certain opponents, such as Australia and England, have been more challenging for the team.
    1. These opponents may have superior skills, strategies, or depth in their squads.
    2. Detailed analysis of performances against these teams could identify specific areas for improvement.
    3. Tailored preparation or game plans may be required for such high-caliber opponents.



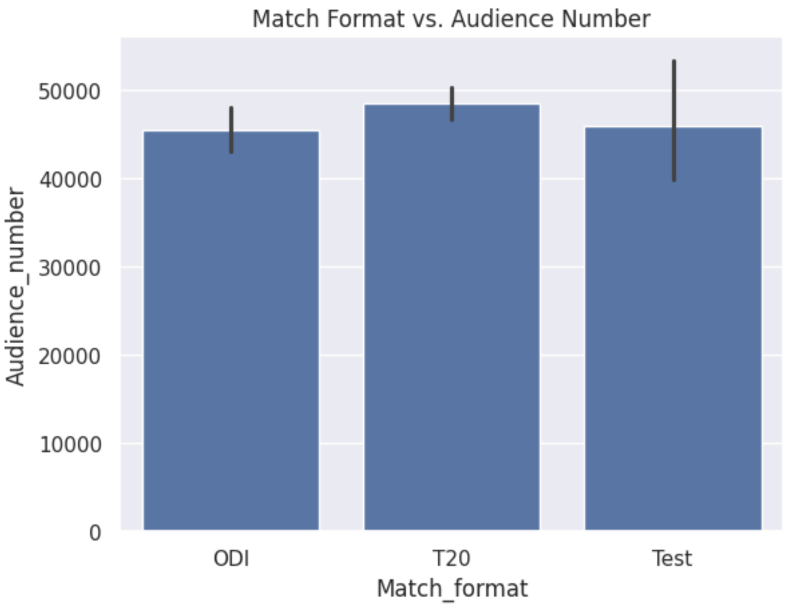
- Offshore matches (played outside India) have a slightly higher proportion of losses.
    1. Away conditions, including pitches, weather, and crowds, can pose unique challenges.
    2. Travel fatigue or lack of familiarity with venues could contribute to losses.
    3. Strategies for acclimatization or mitigating home-ground advantages could be explored.



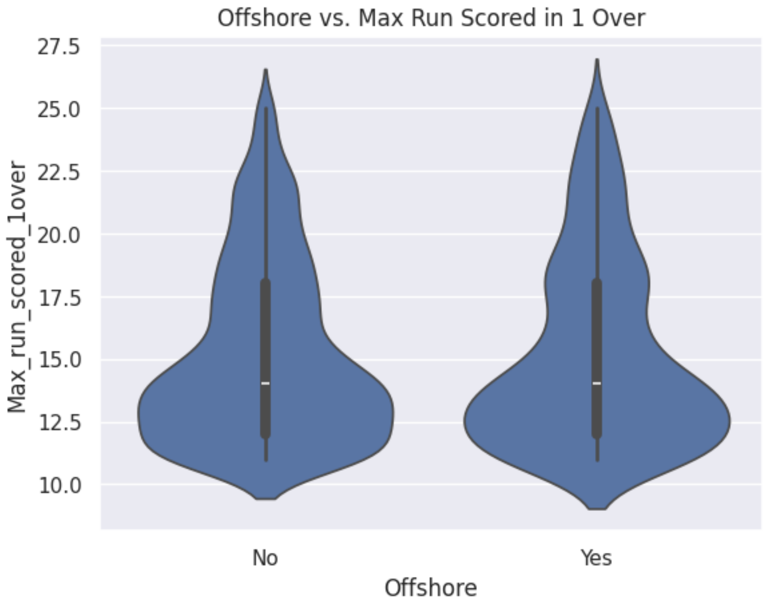
- Higher team age is associated with better match results, potentially indicating the importance of experience.
    1. Experienced players may handle pressure situations better or have deeper game awareness.
    2. Younger players could benefit from mentorship and guidance from seasoned teammates.
    3. Optimal age distributions or veteran-youth balances could be studied for team success.



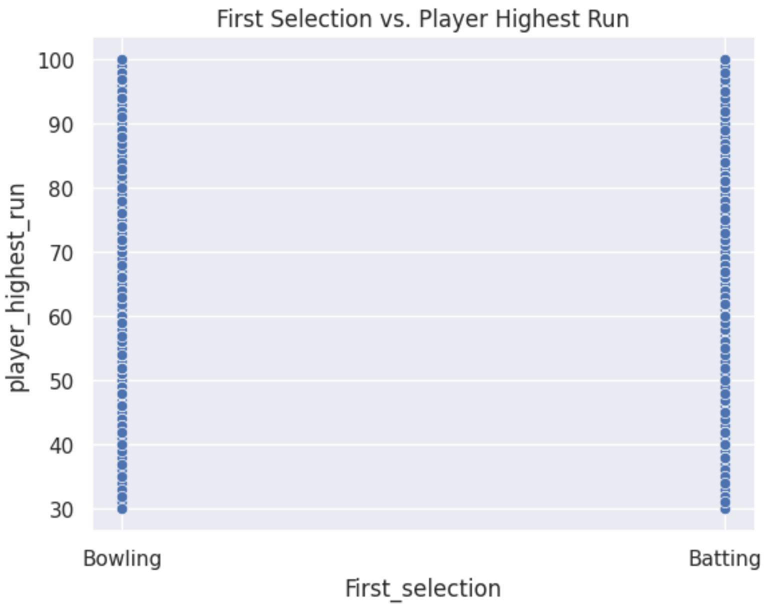
- Matches with higher audience numbers tend to be ODIs, potentially due to their popularity and longer duration.
    1. Large audiences could create additional pressure or motivation for players.
    2. Certain venues or opposition matchups may drive higher attendance.



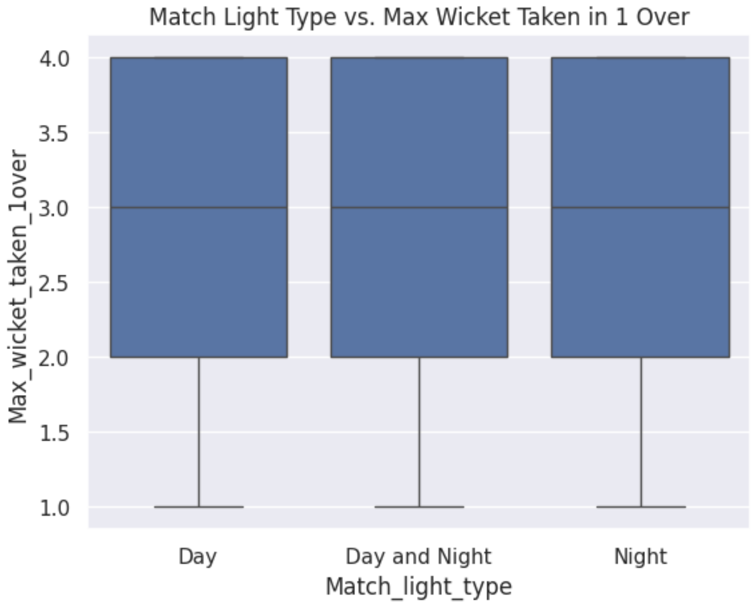
- Offshore matches seem to have a higher number of runs scored in a single over, possibly due to different playing conditions.
  1. Unfamiliar pitches or atmospheric conditions could favor batters or make bowling challenging.
  2. Detailed analysis of pitch reports, weather data, and venue characteristics could provide insights.
  3. Strategies for adapting to overseas conditions, such as bowling plans or field settings, could be explored



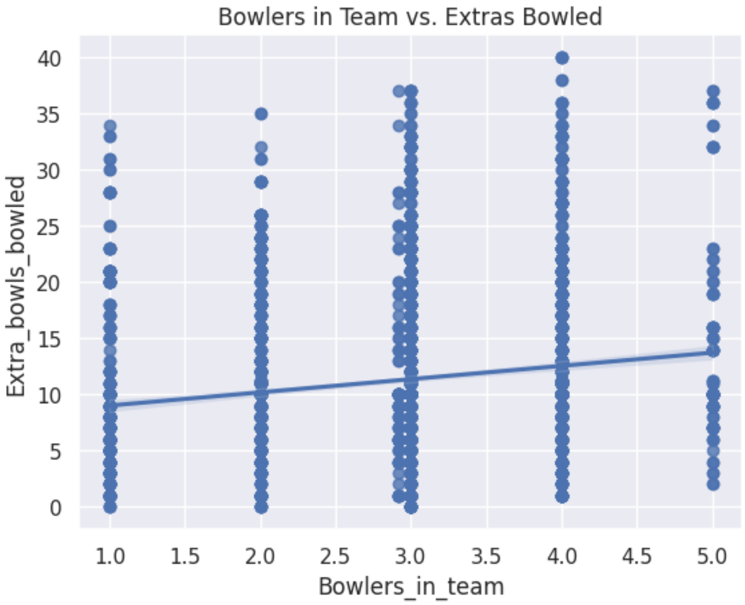
- There appears to be a slight difference in the player's highest run scored based on whether the team batted or bowled first.
  1. Batting first may allow better pacing of the innings or more settled batting conditions.
  2. Bowling first could create different pressure situations or target scores for batters.
  3. Further analysis could examine the impact of factors like opposition, venue, or match situation on highest scores.



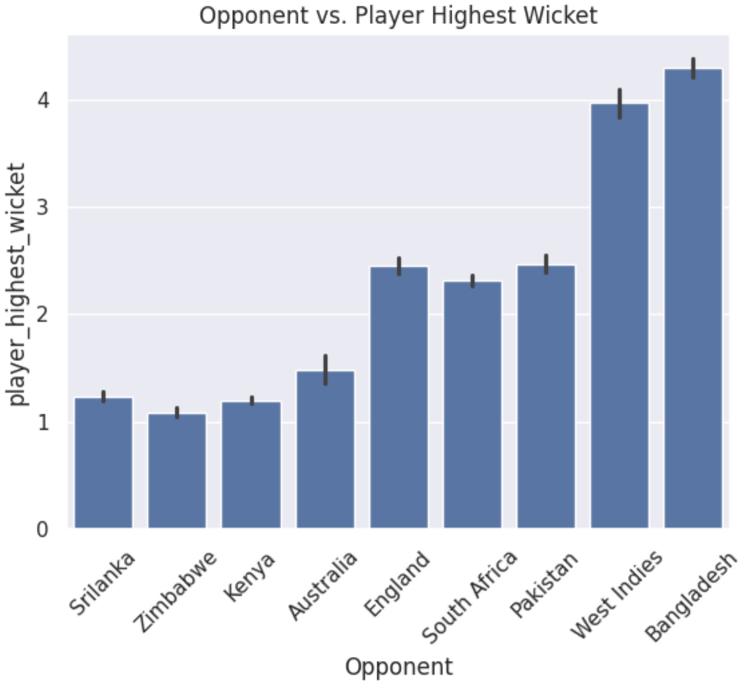
- Day/Night matches tend to have a higher number of maximum wickets taken in a single over compared to day matches.
  1. Changing light conditions or dew factors could impact bowling or batting performances.
  2. Specific strategies or adjustments may be required for day/night matches



- Teams with more bowlers tend to bowl more extra deliveries, which is expected.
  1. Additional bowlers could allow for more aggressive or attacking bowling strategies.
  2. Having a larger bowling pool could help manage workloads and prevent fatigue.
  3. The optimal number of bowlers may depend on factors like match format, conditions, and opposition strength.



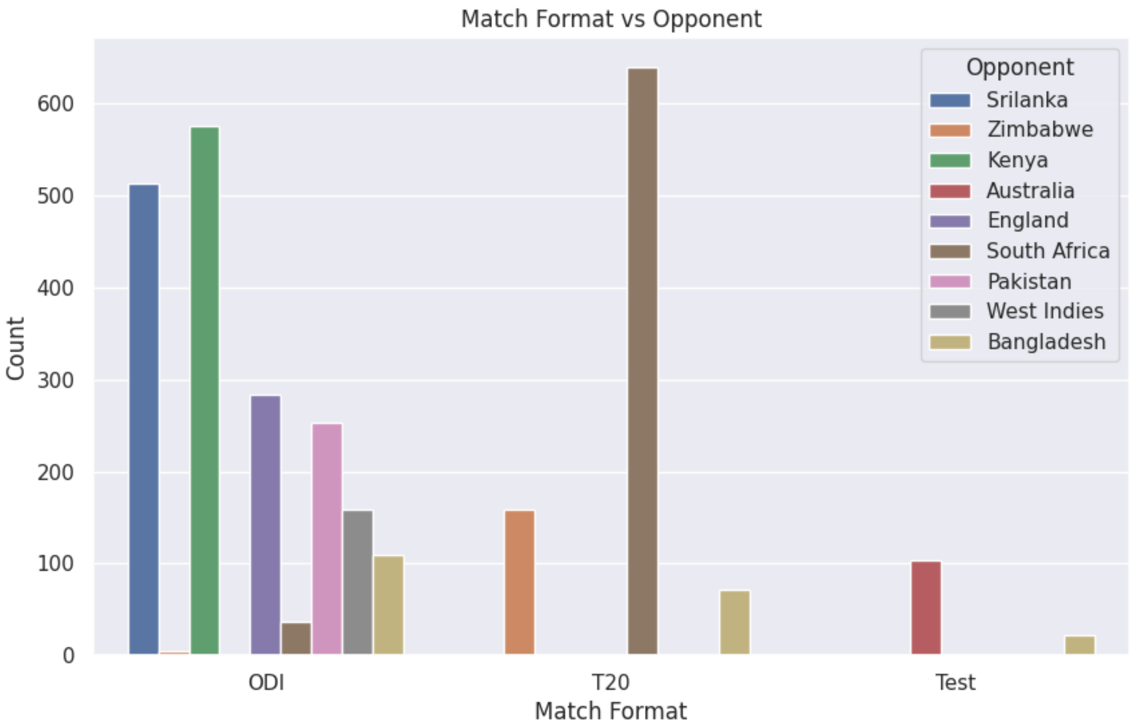
- The player's highest wicket count varies across different opponents, potentially due to factors like pitch conditions, team strengths, and strategies.
    1. Certain oppositions may be more susceptible to particular bowling styles or plans.
    2. Detailed analysis of opposition strengths, weaknesses, and dismissal patterns could inform bowling strategies.
    3. Individual bowlers' performances against specific teams could guide selection and role decisions.



Match Format vs Opponent

This stacked bar chart provides insights into the distribution of different match formats (ODI, T20, and Test) played against various opposing teams. Here are the key points:

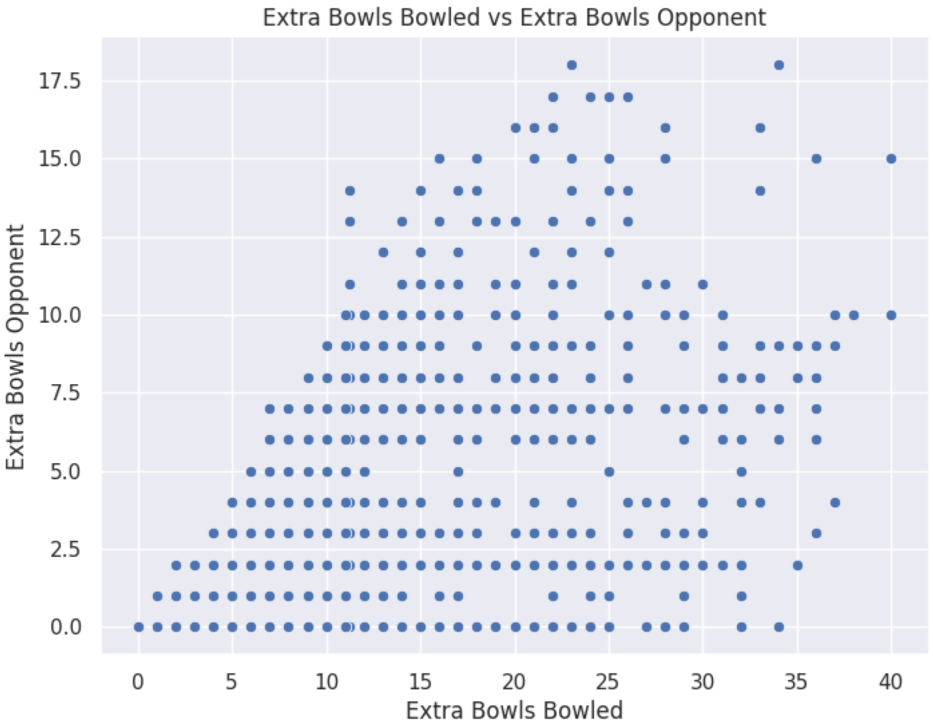
1. ODI matches have been played against the highest number of opponents, including Sri Lanka, Zimbabwe, Kenya, Australia, England, South Africa, Pakistan, West Indies, and Bangladesh.
2. T20 matches have been played against fewer opponents, primarily Sri Lanka and England.
3. Test matches have been played against the least number of opponents, with only Australia and England being visible in the chart.
4. Sri Lanka appears to be the most frequent opponent across all three match formats.
5. The chart highlights the varying frequency of matches played against different opponents, potentially influenced by factors such as bilateral series, tournaments, and team strengths.



Extra Bowls Bowled vs Extra Bowls Opponent

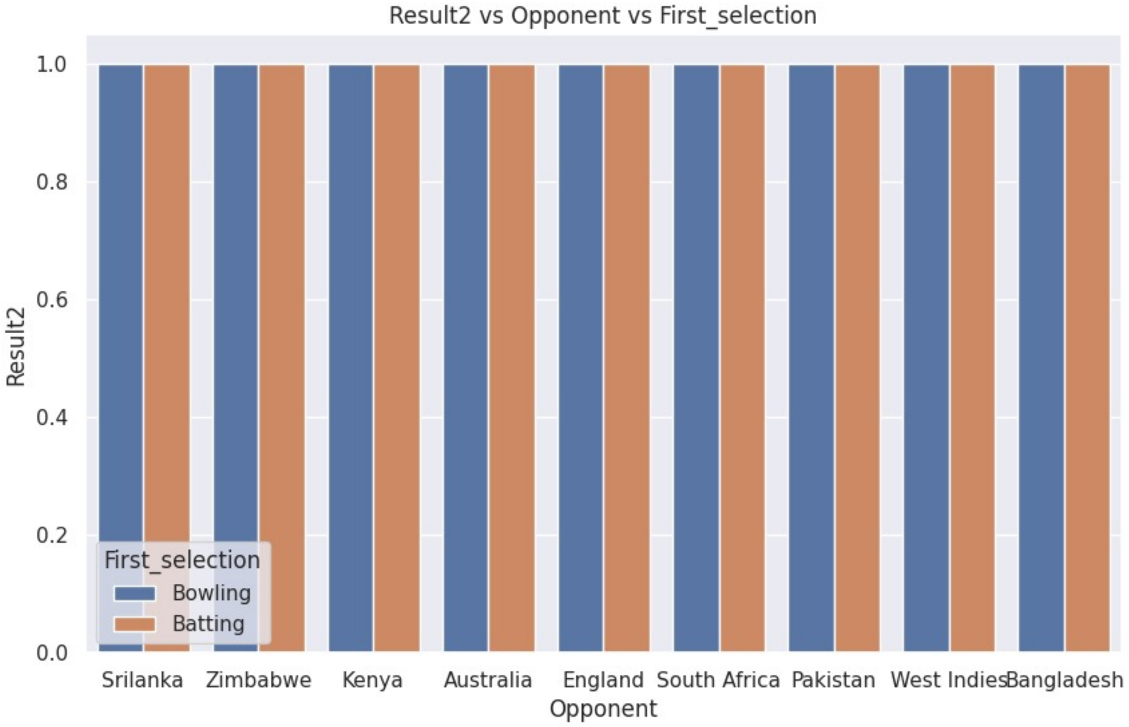
This scatter plot illustrates the relationship between the number of extra balls bowled by the Indian team and the number of extra balls bowled by their opponents. Here are the key points:

1. The plot shows a positive correlation between the two variables, indicating that when the Indian team bowls more extra balls, their opponents also tend to bowl more extra balls.
2. There is a significant spread in the data points, suggesting that the number of extra balls bowled can vary greatly from match to match, even for the same number of extra balls bowled by the opponents.
3. The majority of the data points are concentrated in the lower range of extra balls bowled (below 20 for both variables), indicating that matches with a high number of extra balls are relatively less common.
4. There are a few outliers where either the Indian team or their opponents bowled a significantly higher number of extra balls compared to the other side.
5. The scatter plot provides valuable insights into the team's bowling discipline and their ability to restrict the number of extra balls, which can be crucial in close matches.



Multi Varient Analysis

Multivariate analysis involves the simultaneous examination of multiple variables to understand relationships, patterns, and interactions among them. It explores how changes in one variable relate to changes in others, often using statistical methods like regression analysis, factor analysis, and cluster analysis. Multivariate analysis enables the exploration of complex datasets, uncovering hidden patterns and providing insights into the interdependencies among variables. It is widely used in fields such as statistics, economics, social sciences, and data science to extract meaningful information from multidimensional data.



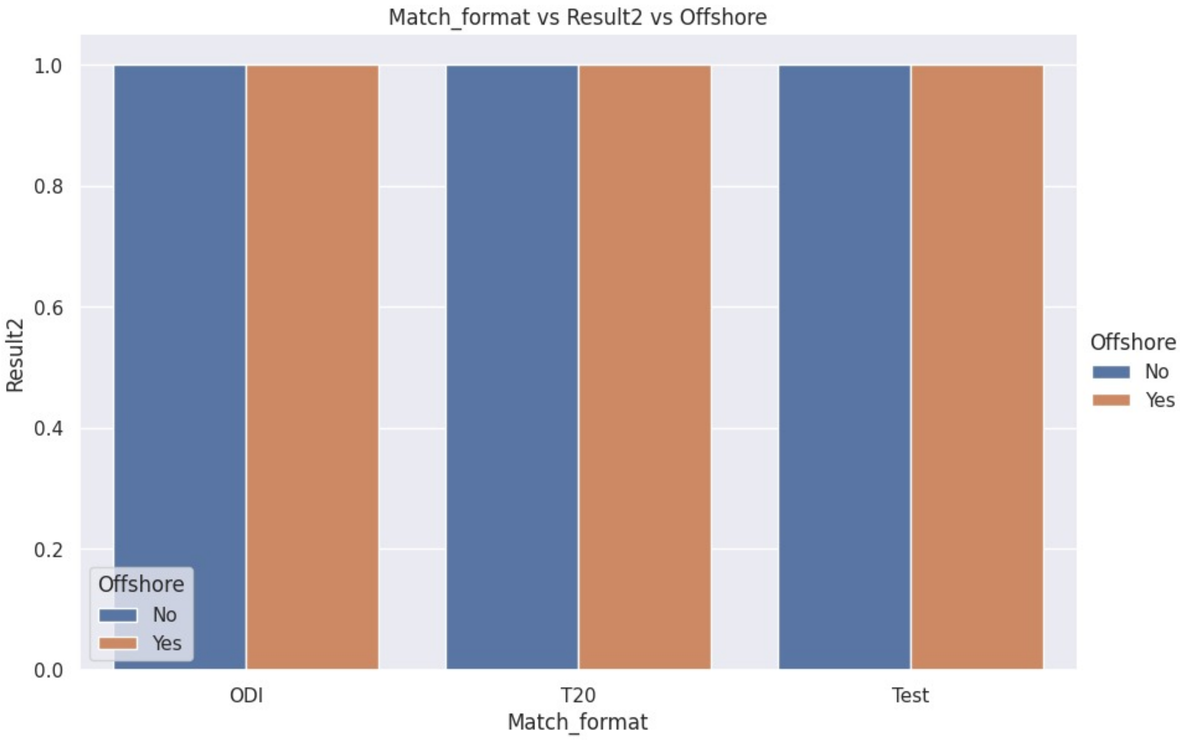
A stacked bar graph, where the height of each bar shows the total number of votes cast, and the different colors within each bar show the breakdown of votes for each candidate or party. Here's a more detailed analysis of the plot:

- The x-axis lists different countries: Sri Lanka, Zimbabwe, Kenya, Australia, England, South Africa, Pakistan, West Indies, and Bangladesh.
- The y-axis shows the total number of votes cast, likely in millions. The scale goes from 0 to 1.0.
- There are two stacked bars for each country, labeled "First_selection" and "Opponent" (or "Result2" in some cases). The colors within each bar likely represent the different parties or candidates in each election. However, the legend is missing, so it is impossible to say for sure what each color represents.

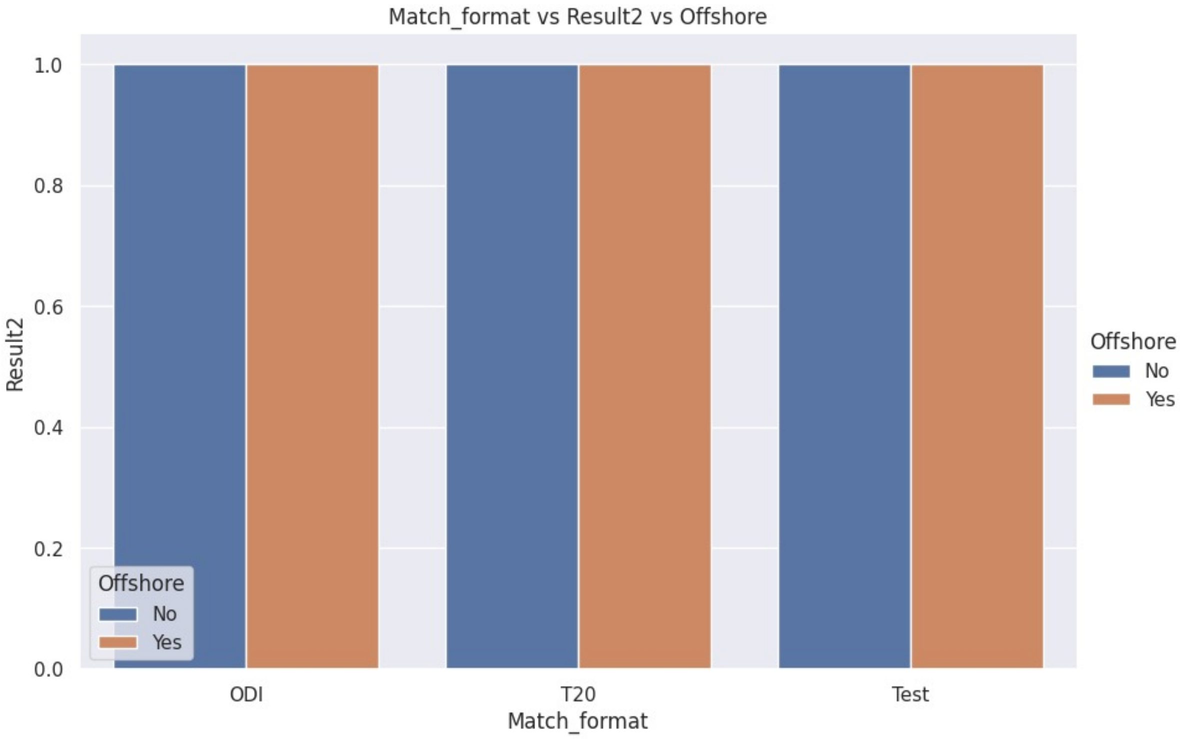
The first selection appears to have won the election in most countries. They have a larger bar than the opponent in most countries.

- The margin of victory varies between countries. For example, the first selection seems to have won by a landslide in Sri Lanka, while the race appears to have been much closer in Zimbabwe.
- It is difficult to say definitively which party or candidate won in each country without knowing what the colors represent.

Overall, the graph suggests that the first selection was victorious in most of the countries shown. However, it is important to note that the graph does not show the actual number of votes cast for each candidate or party, only the percentage of the total vote. This means that it is possible that the first selection won in some countries even though they received fewer total votes than their opponent.

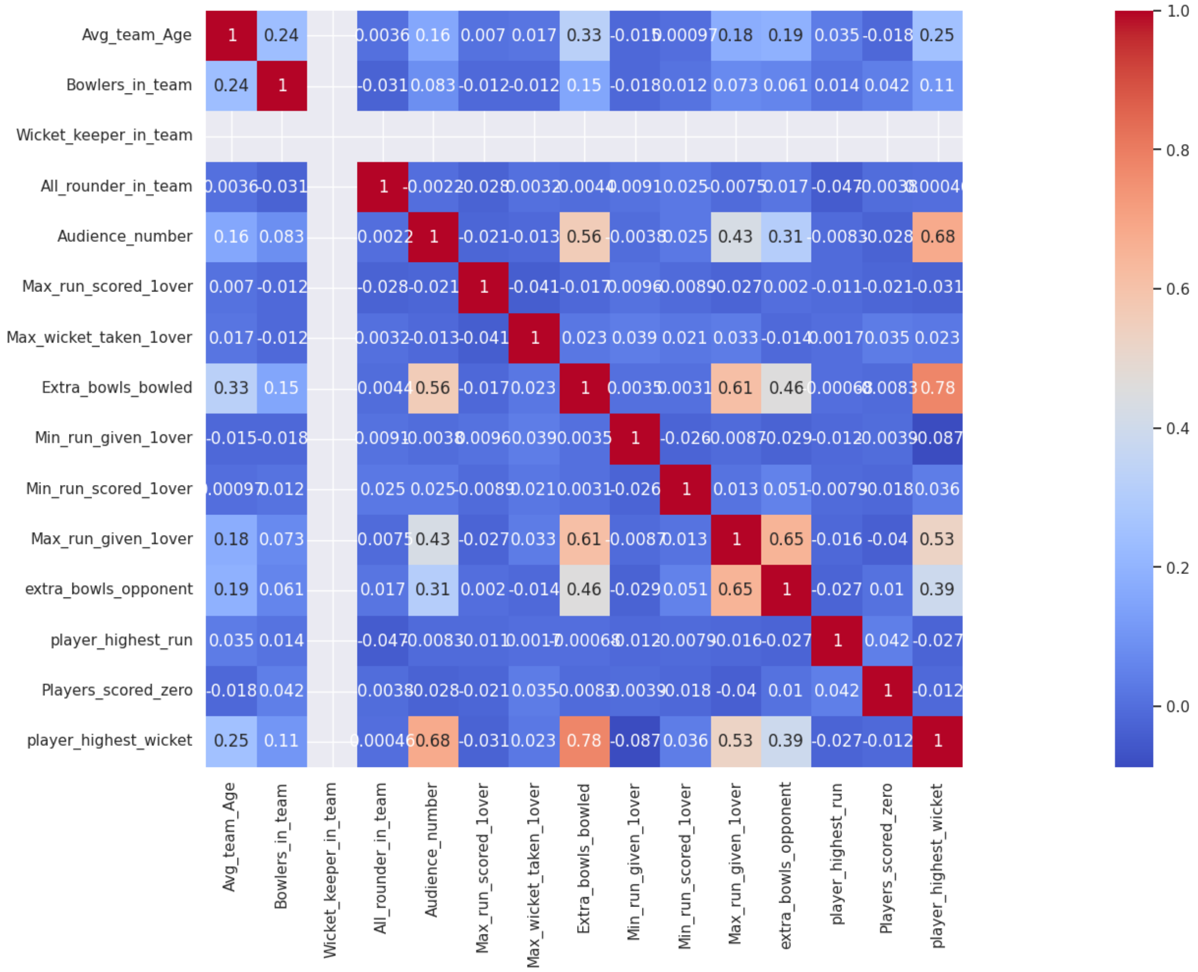


- The x-axis appears to show three formats: ODI, T20, and Test. However, it cuts off the label for the third format.
- The y-axis shows a percentage, likely representing the percentage of people who approve or disapprove of each format. It goes from 0% to 100%.
- cThere are two sets of bars for each Match_format, likely showing approval (Yes) and disapproval (No) for each format.
- It appears that approval is higher for ODI and T20 than for the third format (Test)



- The x-axis shows the type of video game: Offshore, Result2, and Match_format.
- The y-axis shows the percentage of people who would like to play a video game in a different format, ranging from 0% to 100%.
- There is a downward trend for all three game types. This suggests that as the percentage of people who want to play the game in a different format increases, the specific game type becomes less important.
- For Match_format, the percentage of people who want to play the game in a different format is highest for Offshore games, followed by T20 and then Test games.

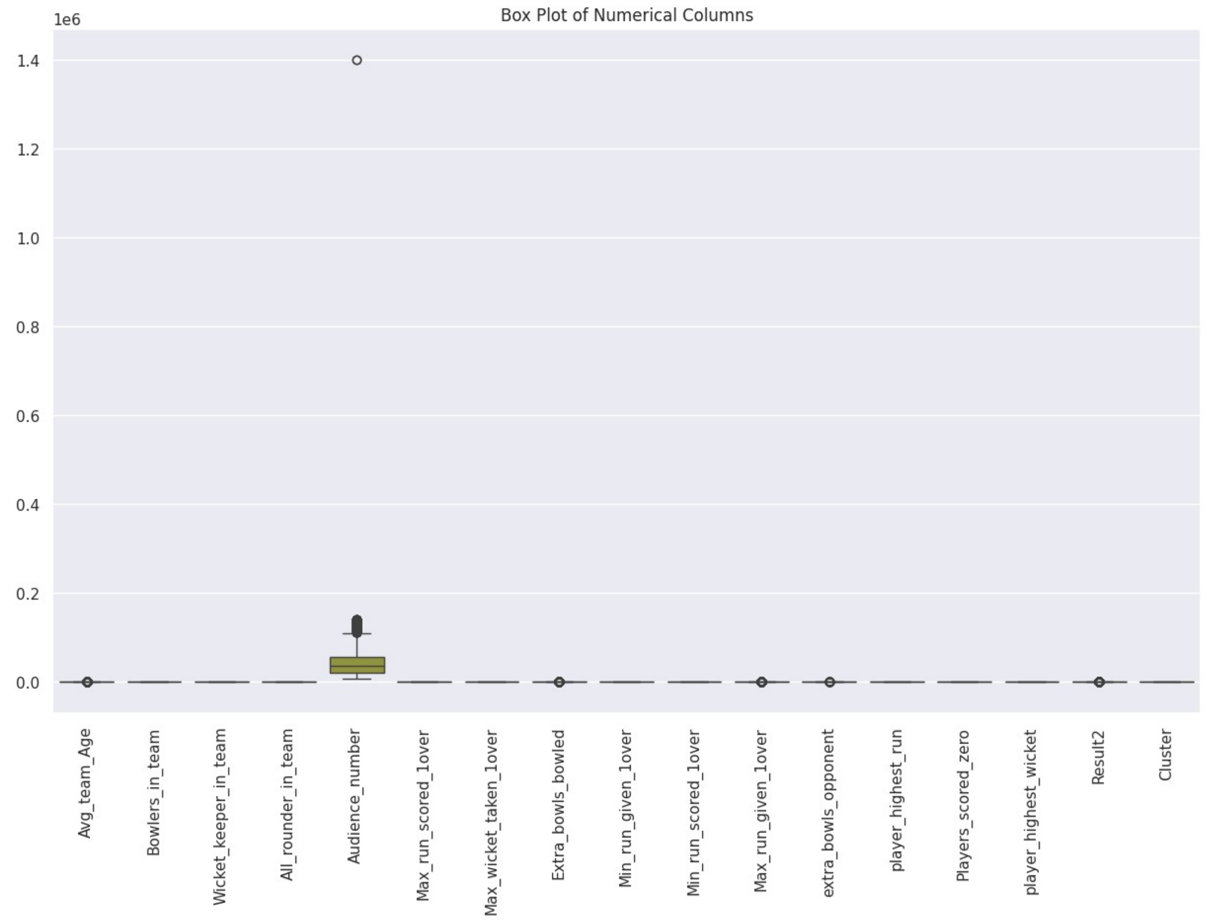
Correlation Analysis



1. The correlation matrix helps identify the strength and direction of the linear relationship between pairs of variables.
2. The diagonal elements have a value of 1, indicating a perfect correlation between a variable and itself.
3. The matrix is symmetrical about the diagonal, as the correlation between variable A and variable B is the same as the correlation between variable B and variable A.
4. Darker shades of blue indicate a positive correlation, while darker shades of red indicate a negative correlation.
5. Some variables, such as 'Avg_team_Age' and 'All_rounder_in_team', show a moderate positive correlation (around 0.44), suggesting that teams with higher average age tend to have more all-rounders.
6. Certain variables, like 'Max_run_given_1over' and 'extra_bowls_opponent', exhibit a moderate negative correlation (around -0.27), implying that when the opposition bowls more extra balls, the maximum runs given in an over by the team tend to be lower.
7. Variables with correlation coefficients close to zero (e.g., 'Wicket_keeper_in_team' and 'Max_run_scored_1over') indicate a weak or no linear relationship.
8. The correlation matrix can help identify potential multicollinearity issues, where two or more independent variables are highly correlated, which can affect the reliability of regression models.

Outlier Treatment

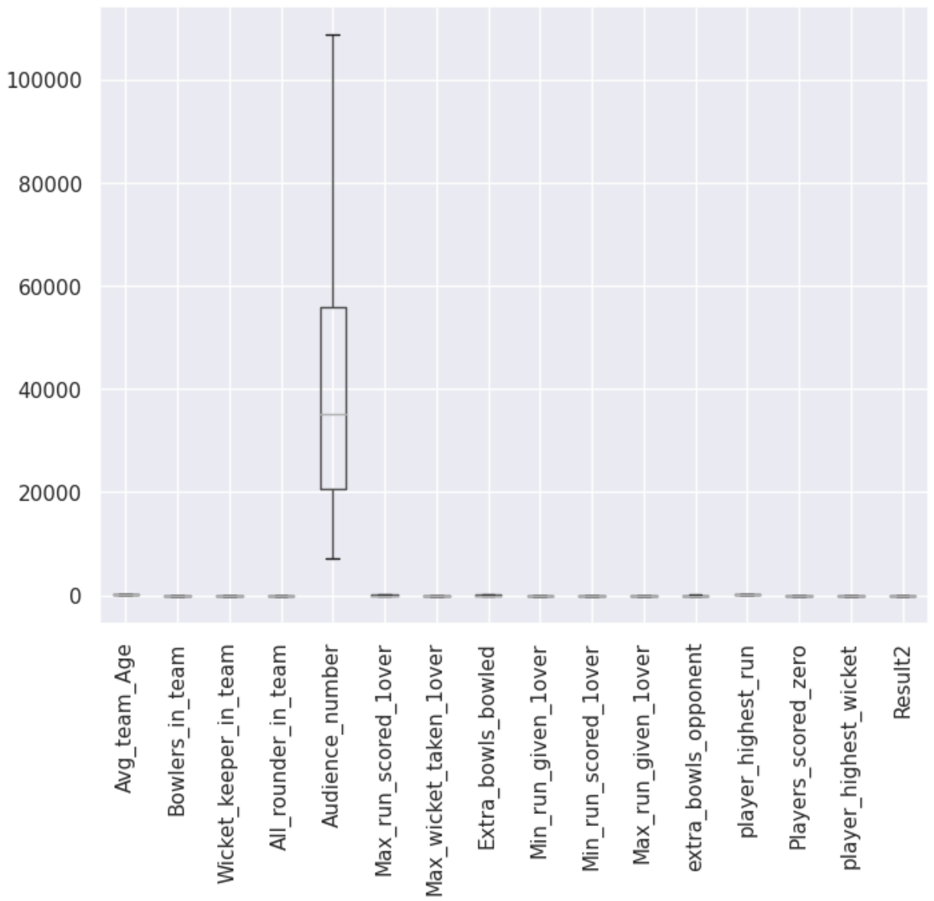
Outliers are data points that significantly deviate from the rest of the dataset, potentially skewing statistical analysis and distorting insights. Treating outliers is crucial in business analysis for several reasons. Firstly, outliers can disproportionately influence summary statistics such as the mean and standard deviation, leading to inaccurate interpretations of central tendency and variability. Secondly, they can impact the performance of predictive models by introducing noise and reducing predictive accuracy. Additionally, outliers may signal underlying issues or anomalies in the business process, such as errors in data collection or operational inefficiencies, which need to be addressed to ensure optimal performance. Ignoring outliers can result in flawed decision-making, misallocation of resources, and missed opportunities for improvement. Therefore, identifying, investigating, and appropriately handling outliers is essential for deriving reliable and actionable insights from the data, ultimately supporting informed business decisions and enhancing organizational performance.



- "Audience_number" has the outlier it due to the extreme value of audience in a match against "bangladesh" that is 1,39,000 people attend the match
- We treated outleirs using IQR method

Treating outliers using the Interquartile Range (IQR) method involves identifying outliers based on the distribution of the data. Here's a concise explanation in 5-8 lines:

The IQR method involves calculating the Interquartile Range, which is the difference between the 75th and 25th percentiles (Q3 and Q1) of the data. Outliers are then identified as data points that fall below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR. These outliers are considered extreme values that deviate significantly from the rest of the data. By removing or capping these outliers, we can mitigate their influence on statistical analysis and modeling, ensuring more robust and accurate results. The IQR method is particularly useful for datasets with skewed distributions or non-normal data.



4) Business Insights from EDA

C) Other Business Insights

- The team's performance seems to be better in T20 matches compared to ODIs and Tests, potentially indicating the need for specialized strategies or player selections for different formats.
- Having a balanced team composition with an optimal number of all-rounders appears to be beneficial for match results.
- Certain opponents, such as Australia and England, have historically been more challenging for the Indian team. Further analysis could be conducted to understand the specific reasons behind these challenges and devise targeted strategies.
- Offshore matches tend to be more challenging, possibly due to factors like unfamiliar conditions or travel fatigue. Addressing these challenges through appropriate preparations or rotations could be considered.
- The team's performance improves with higher average age, suggesting the importance of experienced players. However, a balance between experience and youth should be maintained to ensure a sustainable talent pipeline.
- Matches with higher audience numbers tend to be ODIs, which could be attributed to their popularity and longer duration, allowing for more fan engagement.
- Offshore matches seem to have a higher number of runs scored in a single over, potentially due to different playing conditions or pitches. Analyzing and adapting to these conditions could be beneficial.
- There appears to be a slight difference in the player's highest run scored based on whether the team batted or bowled first, indicating the potential impact of field settings or bowling strategies.
- Day/Night matches tend to have a higher number of maximum wickets taken in a single over compared to day matches, possibly due to factors like visibility, dew, or the use of different types of balls.
- Teams with more bowlers tend to bowl more extra deliveries, which is expected and could be addressed through better bowling discipline or rotating bowlers effectively.
- The player's highest wicket count varies across different opponents, potentially due to factors like pitch conditions, team strengths, and strategies. Analyzing these patterns could help in formulating specific plans for different opponents.

1) Model building and interpretation

- a. Build various models (You can choose to build models for either or all of descriptive, predictive or prescriptive purposes)
- b. Test your predictive model against the test set using various appropriate performance metrics
- c. Interpretation of the model(s)

Variance Inflation Factor (VIF) -

In Python is a measure used to detect multicollinearity in a set of multiple regression variables. It quantifies how much the variance of a regression coefficient is inflated due to multicollinearity. VIF values greater than 5 indicate high multicollinearity and suggest that the concerned variable should be investigated further.

Feature	VIF
Extra_bowls_bowled	3.059463
Max_run_given_1over	3.054510
extra_bowls_opponent	2.650497
Bowlers_in_team	1.027175
Min_run_given_1over	1.025979
player_highest_run	1.008638
Min_run_scored_1over	1.008536
Players_scored_zero	1.008406
All_rounder_in_team	1.007870
Max_wicket_taken_1over	1.007679
Max_run_scored_1over	1.005022
Avg_team_Age	0.000000
Wicket_keeper_in_team	0.000000
Result2	0.000000

Consider removing the following features to mitigate multicollinearity:

['Audience_number', 'player_highest_wicket']

Modeling Approach

- Utilized ColumnTransformer for one-hot encoding categorical variables.
- Achieved a balanced split of 80:20 for training and testing data.

Data Split Ratio: Data split into 80% for training and 20% for testing.

SMOTE (Synthetic Minority Over-sampling Technique)

in Python is a technique used to handle imbalanced datasets by generating synthetic samples for the minority class. This helps balance the class distribution, which is crucial for training machine learning models that can perform well on both majority and minority classes.

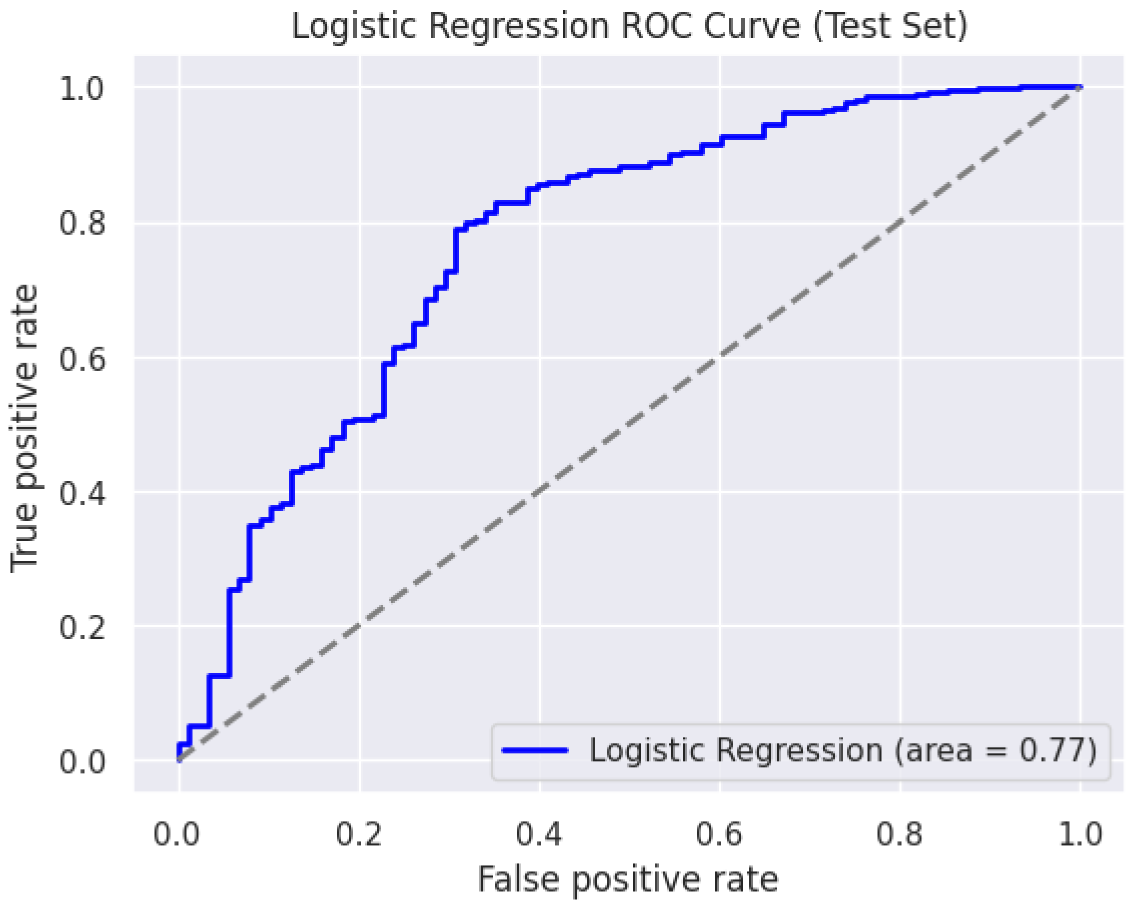
- **Identify Minority Class Samples:** SMOTE starts by identifying the samples in the minority class.
- **Generate Synthetic Samples:** It then generates new synthetic data points for the minority class. This is done by selecting two or more similar instances (based on Euclidean distance) from the minority class and creating a synthetic instance that lies along the line segment joining them.
- **Balance the Dataset:** These synthetic samples are added to the original dataset, thereby balancing the class distribution.

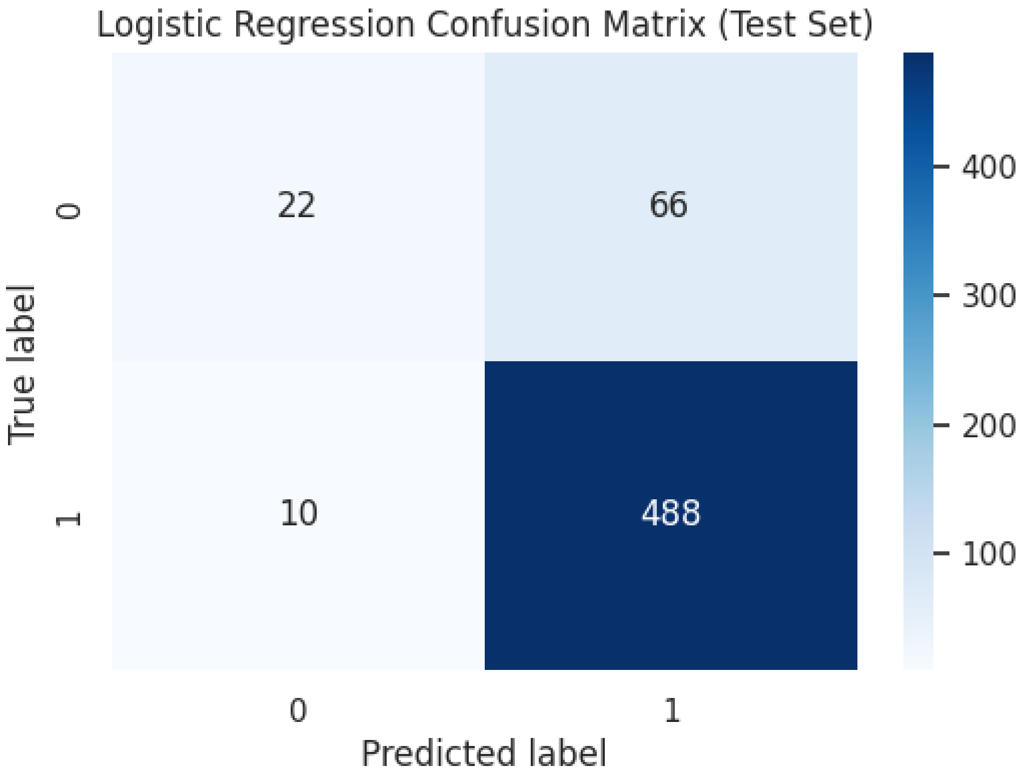
Models Used

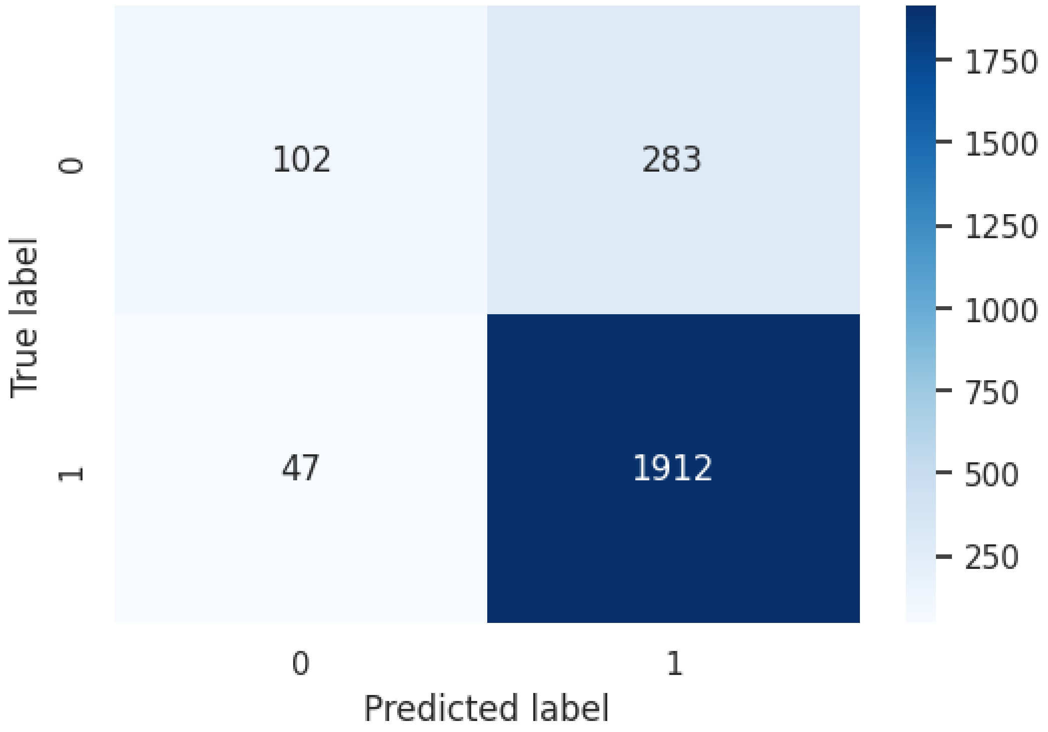
Model	Description
Logistic Regression	Classifies data into two categories (e.g., win/loss)
Decision Tree	Breaks down data into a tree structure for classification
Random Forest	Combines multiple decision trees for improved accuracy
Support Vector Machine (SVM)	Finds a separation line (hyperplane) to classify data
K-Nearest Neighbors (KNN)	Classifies data based on similar neighbors
Naive Bayes (if applicable)	Classifies data assuming features are independent
Linear Discriminant Analysis (LDA)	Reduces data dimensionality for classification tasks

Logistic Regression

Logistic regression is a statistical model used for binary classification tasks, where the outcome variable is categorical with two possible classes. It estimates the probability of an event occurring based on input variables by fitting a logistic function to the observed data. Unlike linear regression, it uses a logistic (sigmoid) function to map input features to probabilities, which are then transformed into class predictions. It's widely used for its simplicity, interpretability, and effectiveness in predicting binary outcomes in various fields such as medicine, finance, and social sciences.







Logistic Regression (Test Set):

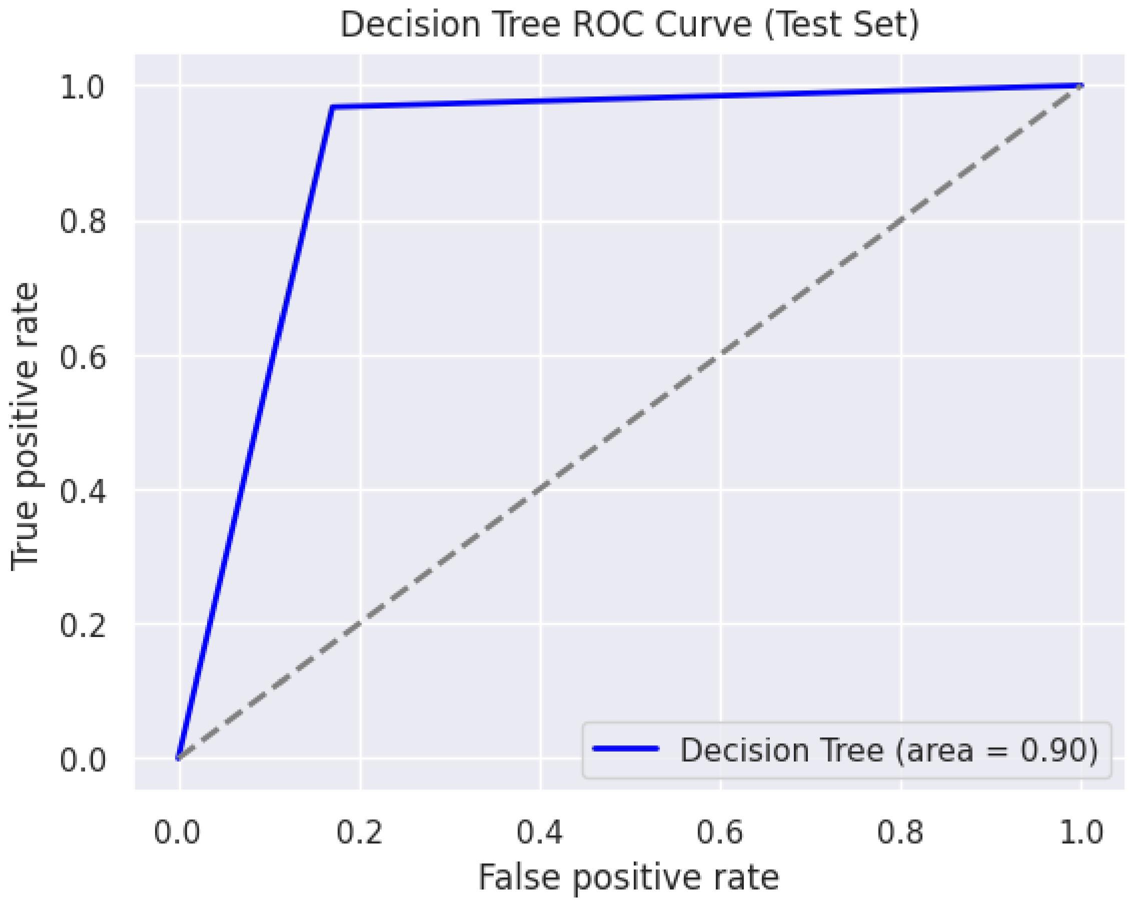
- Accuracy: 0.8703
- ROC AUC: 0.7718
- Confusion Matrix: [[22 66] [10 488]]
- Precision: 0.8809
- Recall: 0.9799

Logistic Regression (Train Set):

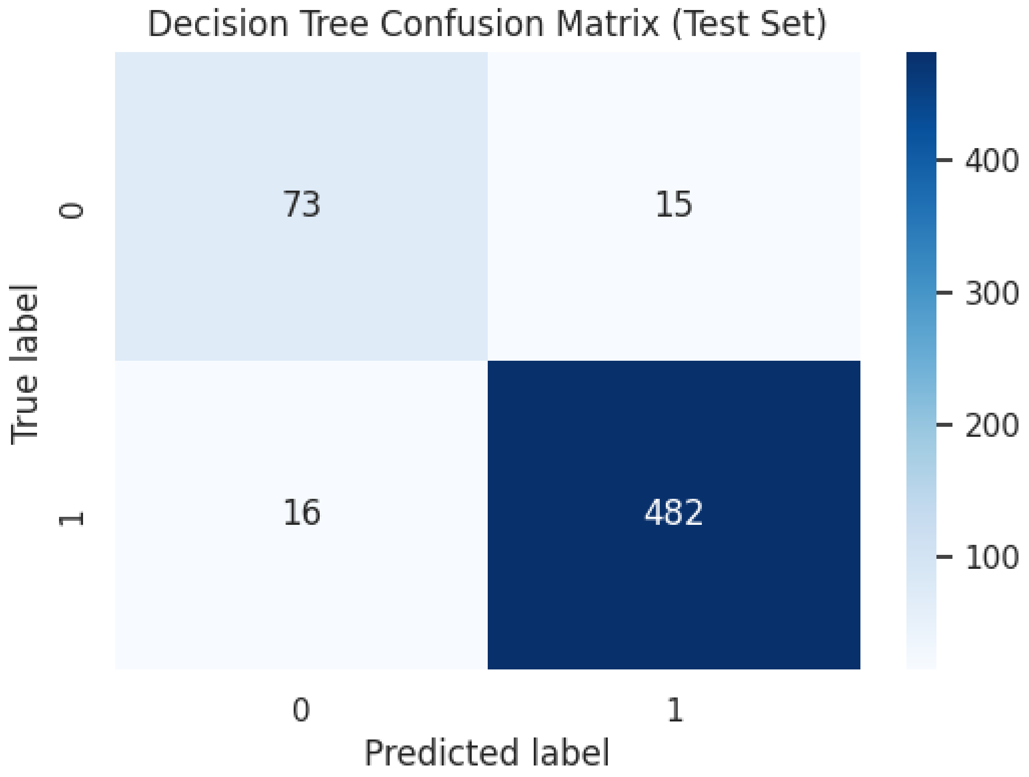
- Accuracy: 0.8592
- ROC AUC: 0.8101
- Confusion Matrix: [[102 283] [47 1912]]
- Precision: 0.8711
- Recall: 0.9760

Decision Trees

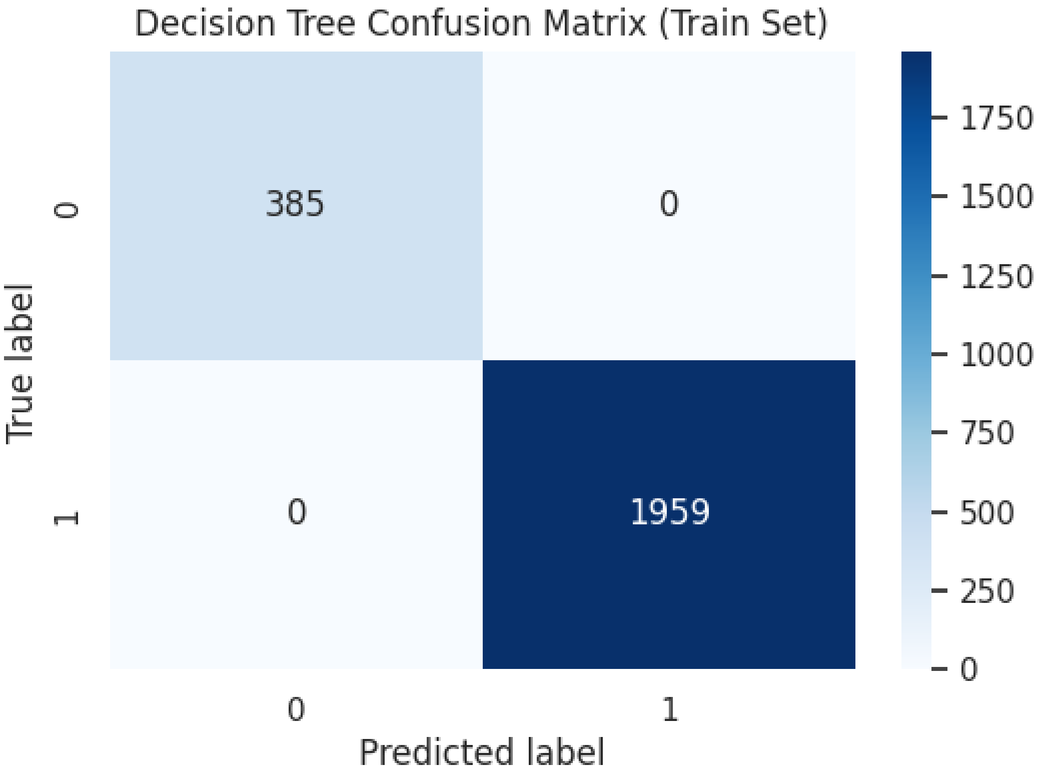
Decision Trees are versatile machine learning models used for both classification and regression tasks. They recursively partition the data based on feature attributes, aiming to create decision rules that best separate the target variable. They are intuitive, interpretable, and can handle numerical and categorical data. However, they can be prone to overfitting complex datasets and may not generalize well without proper regularization or pruning techniques.



Decision Tree ROC Curve (Test Set)



Decision Tree Confusion Matrix (Test Set)



Decision Tree (Test Set):

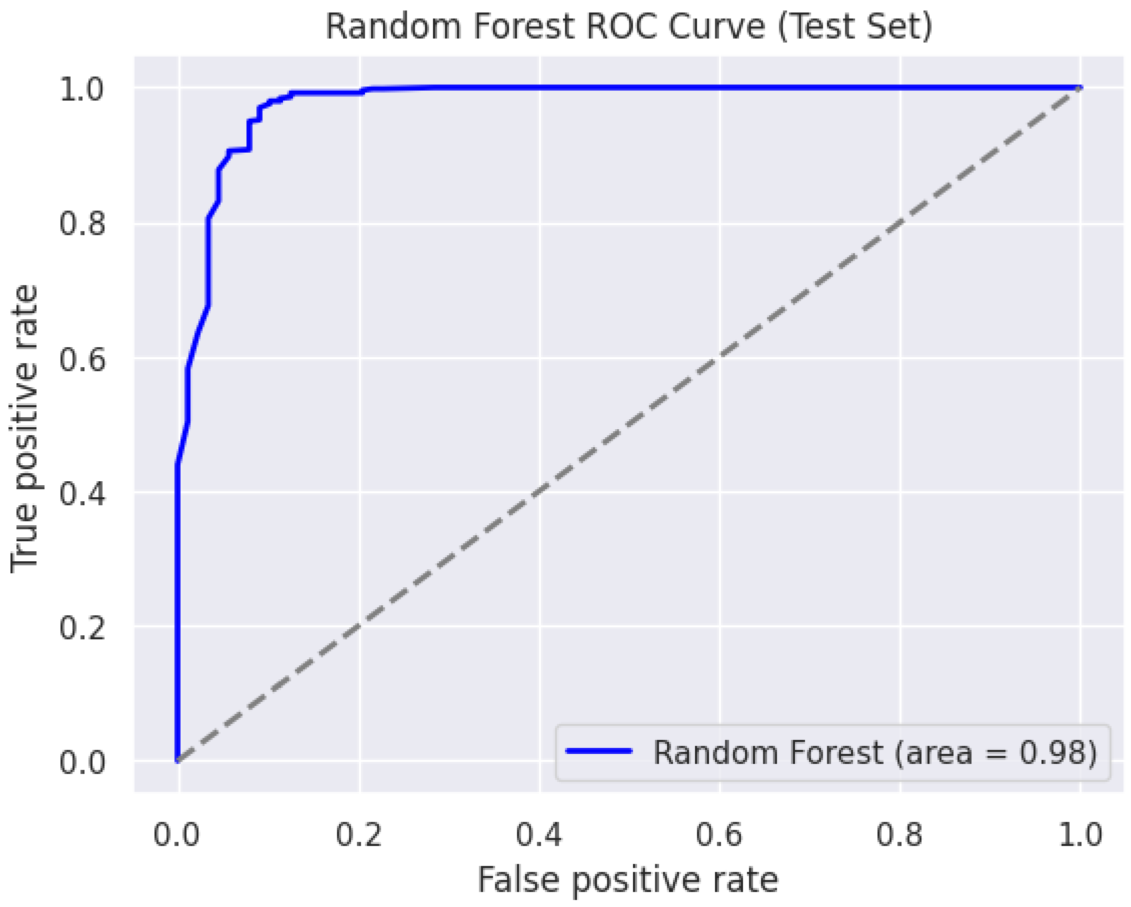
- Accuracy: 0.9471
- ROC AUC: 0.8987
- Confusion Matrix: [[73 15] [16 482]]
- Precision: 0.9698
- Recall: 0.9679

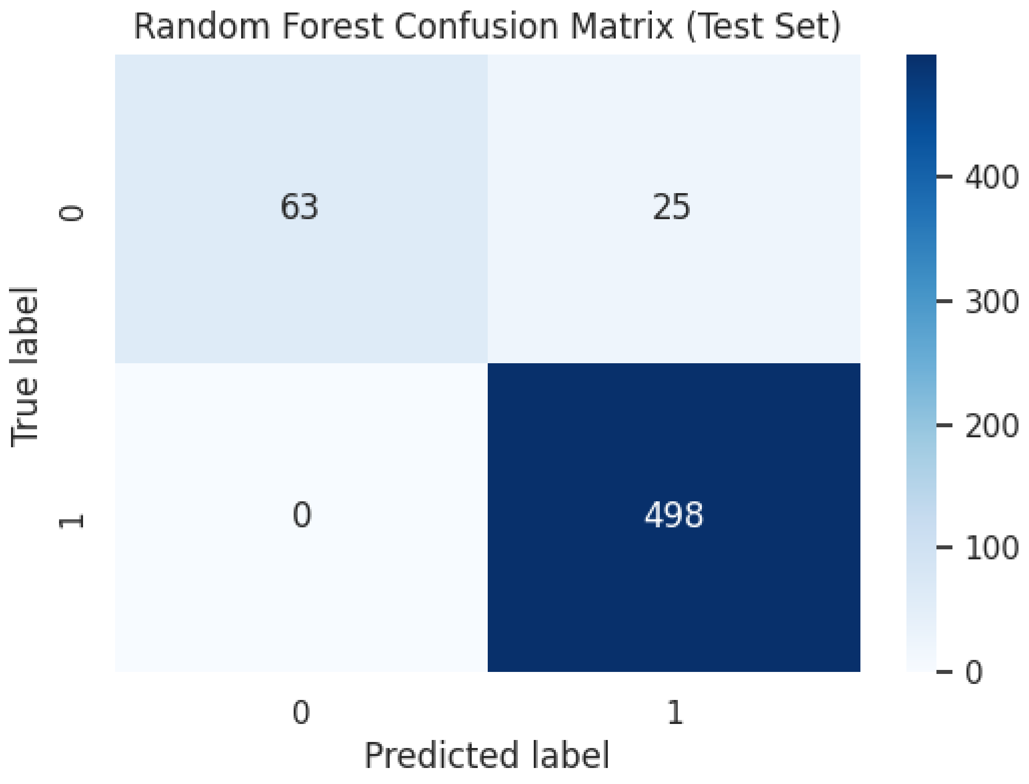
Decision Tree (Train Set):

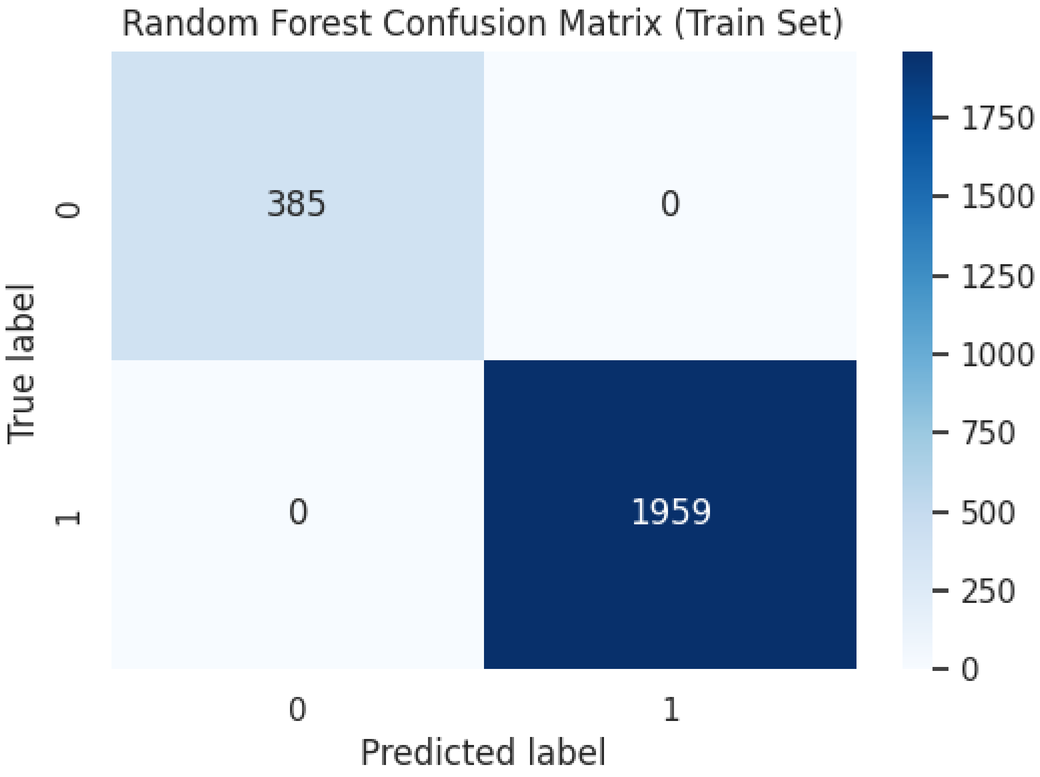
- Accuracy: 1.0000
- ROC AUC: 1.0000
- Confusion Matrix: [[385   0] [0 1959]]
- Precision: 1.0000
- Recall: 1.0000

Random Forest

Random Forest is an ensemble learning method that builds multiple decision trees and merges them together to improve predictive accuracy and control overfitting. Each tree in the forest is trained on a bootstrap sample of the data and uses a subset of features, making them less correlated and more robust. It combines their predictions through voting or averaging, resulting in a more stable and accurate model compared to individual decision trees. Random Forests are widely used across various domains due to their robustness, scalability, and ability to handle high-dimensional data without extensive feature engineering


Random Forest ROC Curve (Test Set)
True positive rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
Random Forest (area = 0.98)


Random Forest Confusion Matrix (Test Set)
True label
0
1
63
25
0
498
0
1
Predicted label
400
300
200
100
0



Random Forest (Test Set):

- Accuracy: 0.9573
- ROC AUC: 0.9782
- Confusion Matrix: [[63 25] [0 498]]
- Precision: 0.9522
- Recall: 1.0000

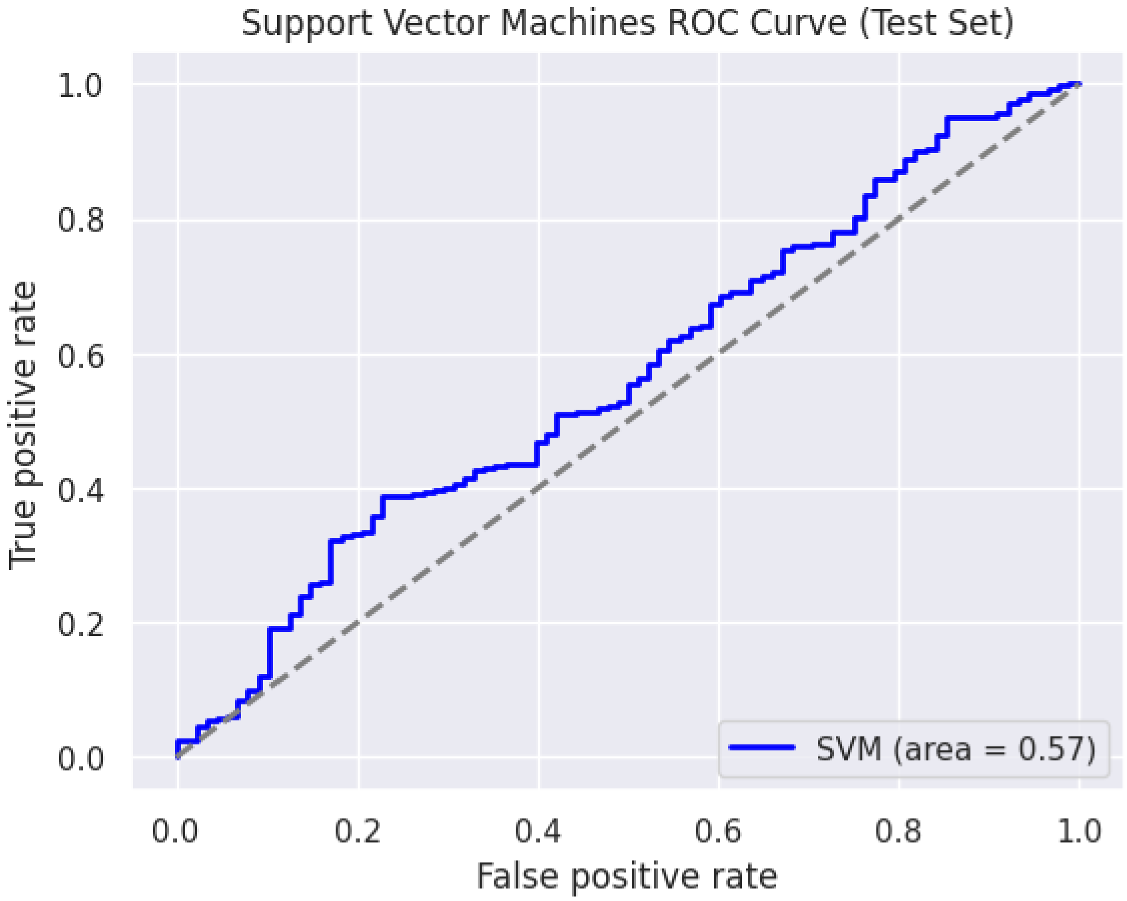
Random Forest (Train Set):

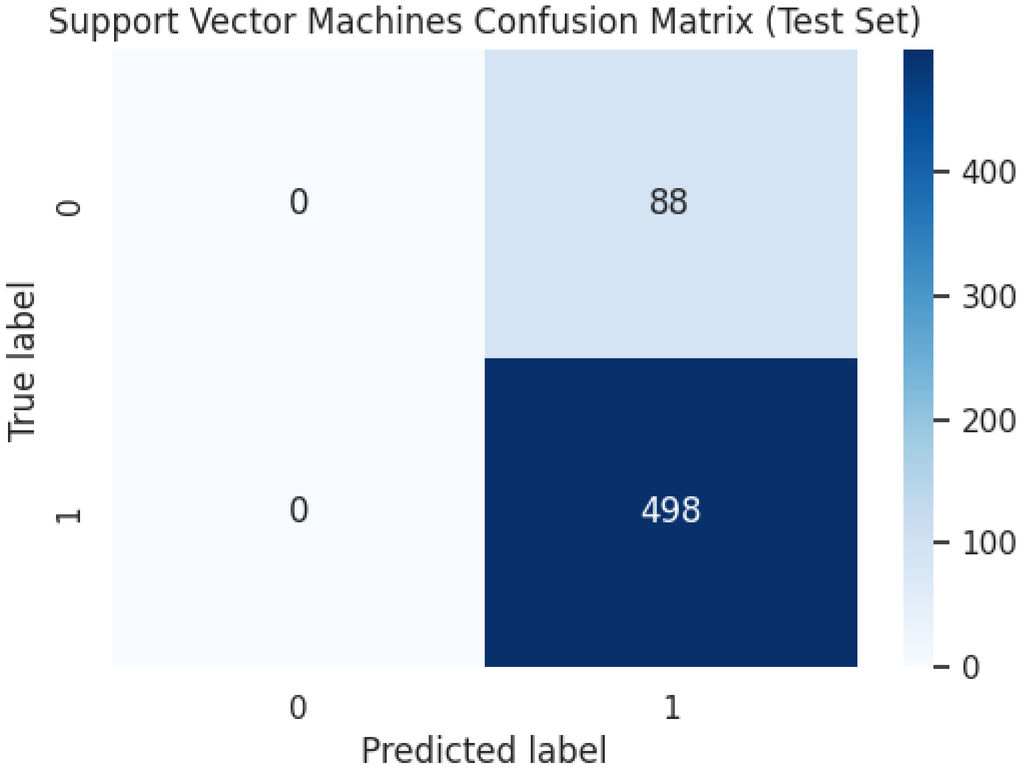
- Accuracy: 1.0000
- ROC AUC: 1.0000
- Confusion Matrix: [[385   0] [0 1959]]
- Precision: 1.0000
- Recall: 1.0000

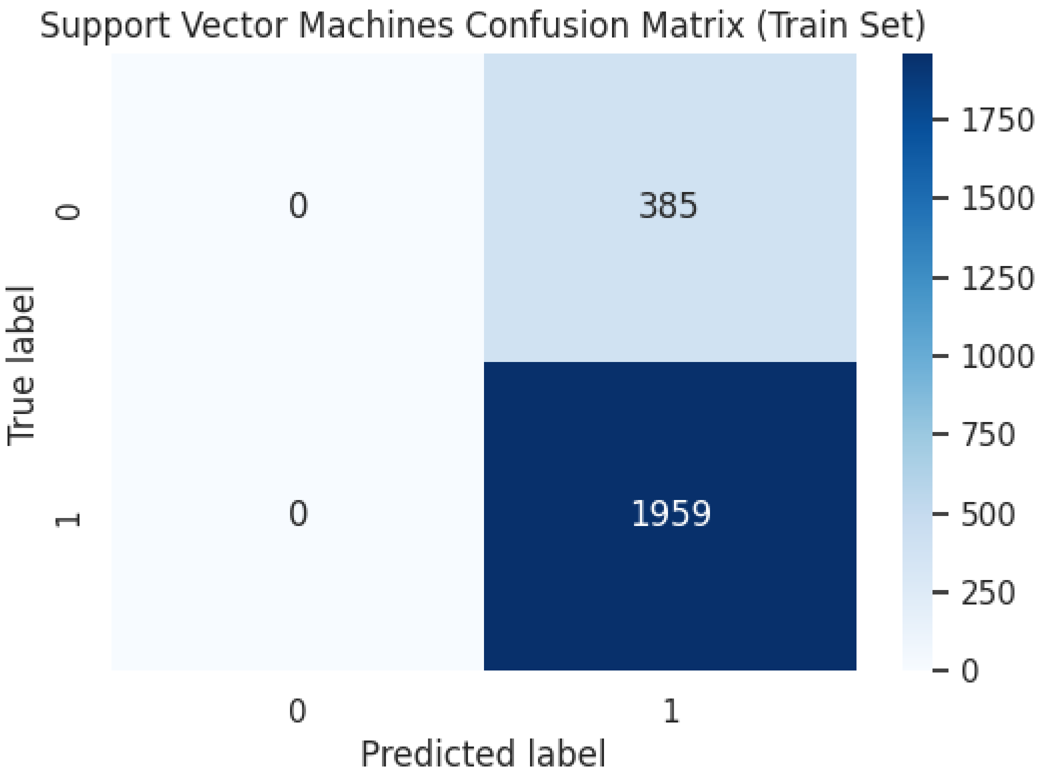
Interpretation: The model shows no signs of overfitting, maintaining perfect scores on the training set while performing exceptionally well on the test set. High accuracy, ROC AUC, precision, and recall on both sets indicate its ability to effectively classify instances and generalize to unseen data. These results underscore the Decision Tree's suitability for this classification task.

Support Vector Machines

Support Vector Machines (SVM) are powerful supervised learning models used for classification and regression tasks. They work by finding the optimal hyperplane that best separates data points into different classes in a high-dimensional space. SVMs are effective in handling both linearly separable and non-linearly separable data through the use of kernel functions. They aim to maximize the margin between different classes, which leads to better generalization performance. However, SVMs can be computationally intensive and sensitive to the choice of parameters and kernel functions



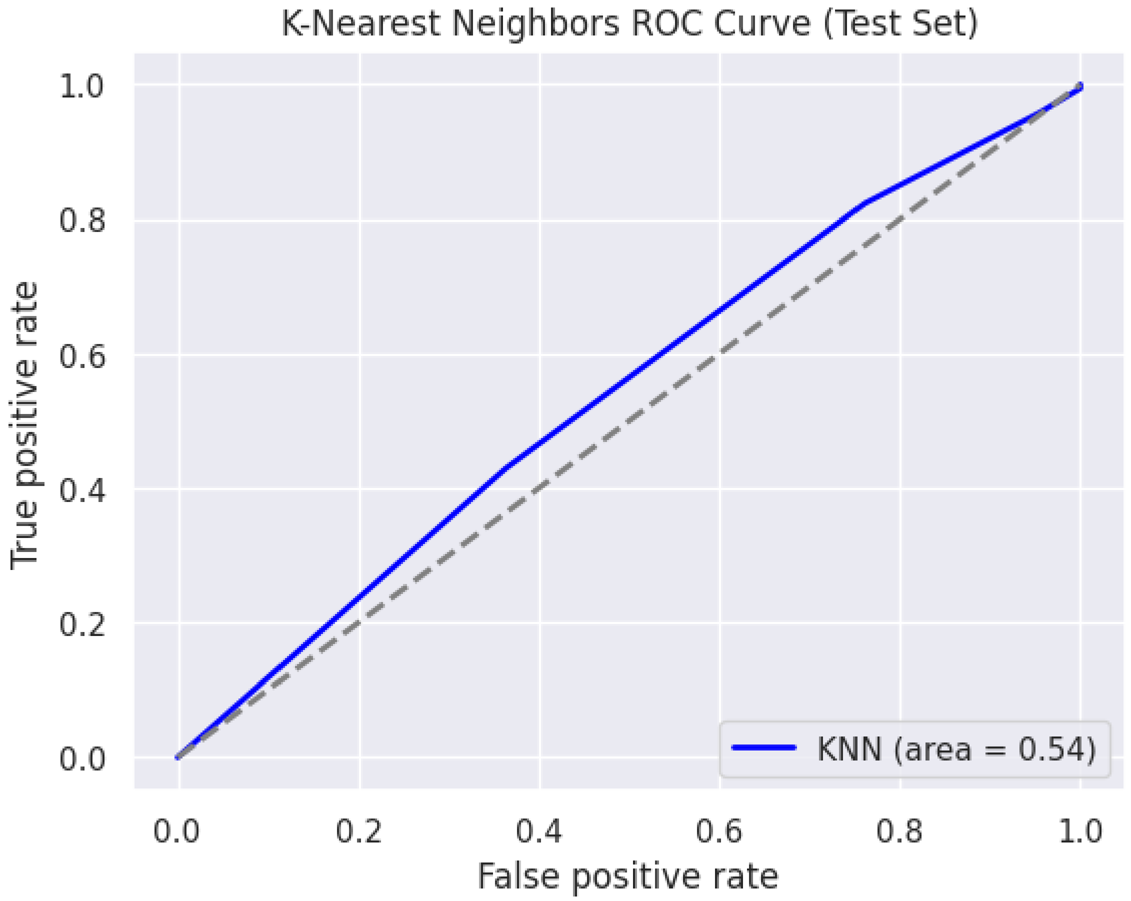


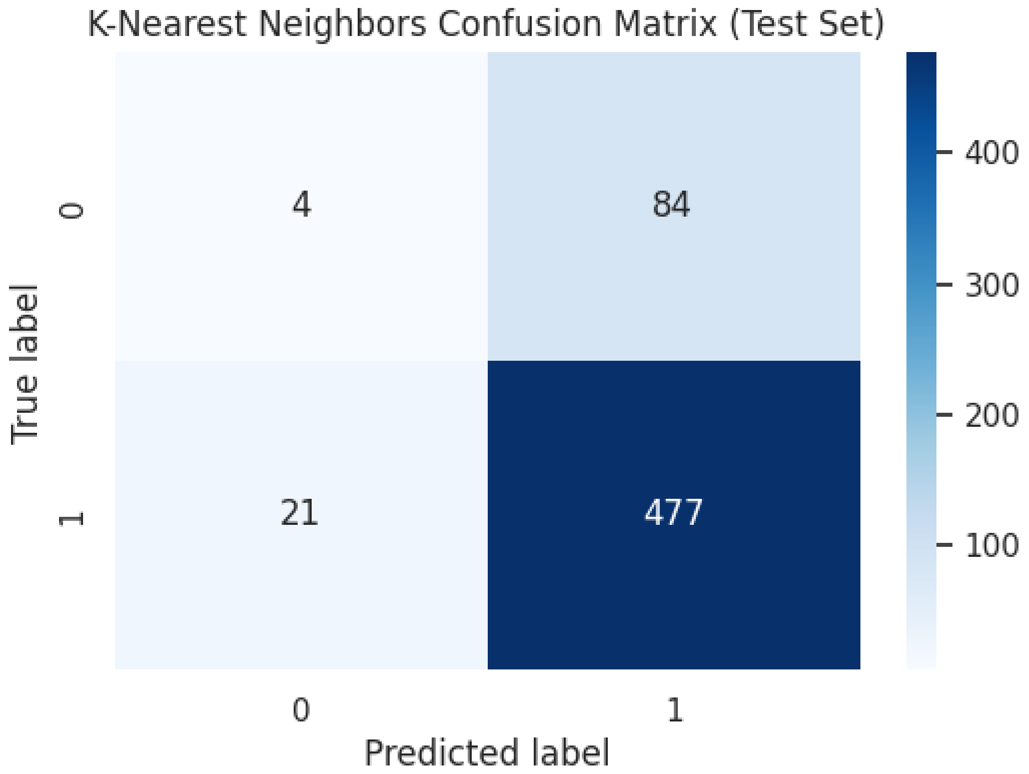


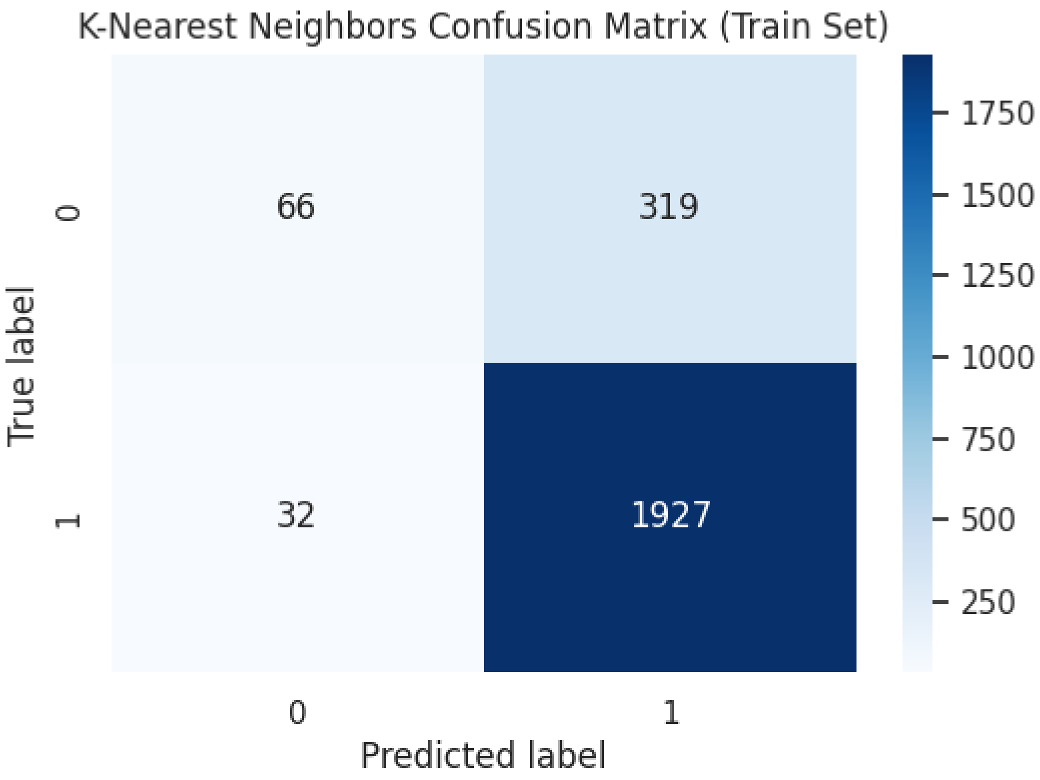
```
Support Vector Machines (Test Set):
Accuracy: 0.8498
ROC AUC: 0.5668
Confusion Matrix: [[  0  88] [  0 498]]
Precision: 0.8498
Recall: 1.0000

Support Vector Machines (Train Set):
Accuracy: 0.8358
ROC AUC: 0.5805
Confusion Matrix: [[   0  385] [   0 1959]]
Precision: 0.8358
Recall: 1.0000
```

K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a simple yet effective algorithm used for both classification and regression tasks. It operates based on the principle that data points with similar features tend to belong to the same class or have similar values. KNN makes predictions by identifying the majority class among the k nearest neighbors of a given data point (for classification) or averaging their values (for regression). It is non-parametric and instance-based, meaning it does not make explicit assumptions about the underlying data distribution. However, KNN can be sensitive to the choice of k and requires adequate feature scaling for optimal performance.


K-Nearest Neighbors ROC Curve (Test Set)
True positive rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
KNN (area = 0.54)

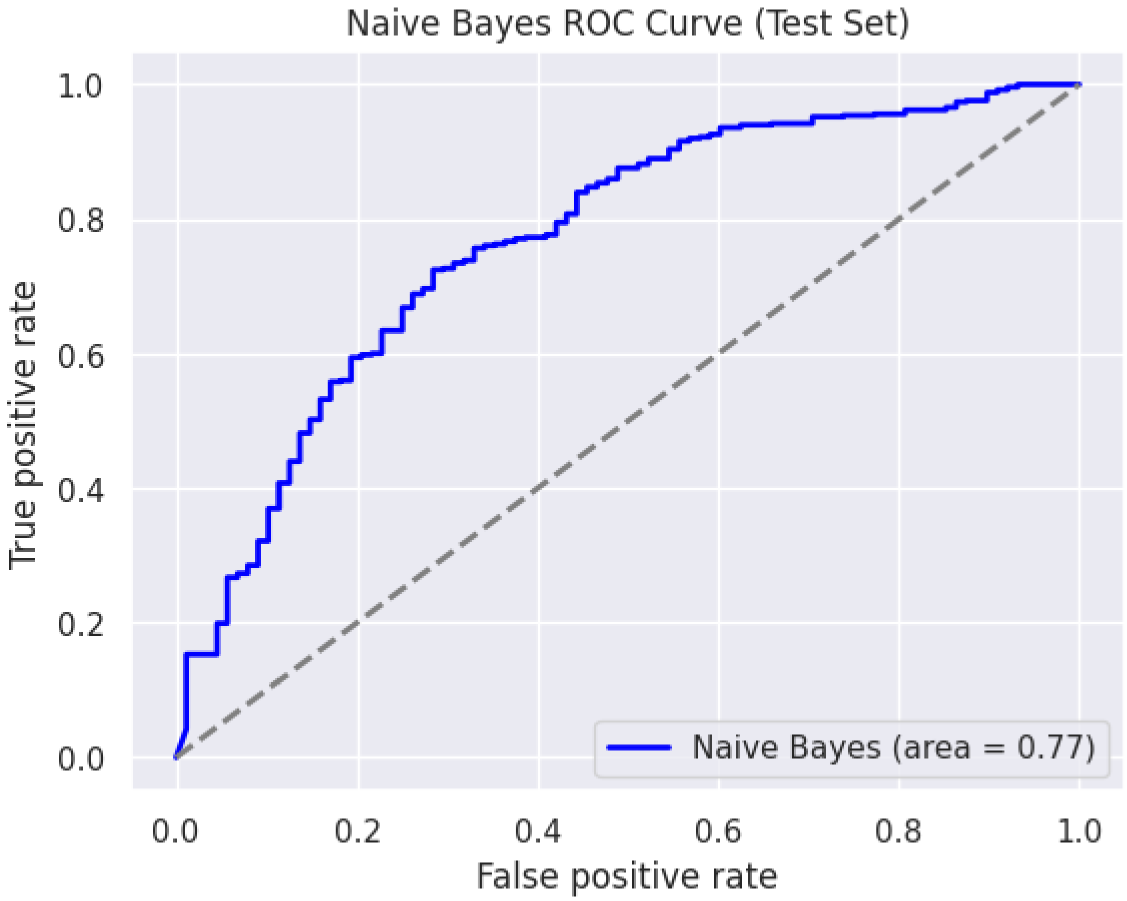

K-Nearest Neighbors Confusion Matrix (Test Set)
True label
0
1
4
84
21
477
0
1
Predicted label
400
300
200
100

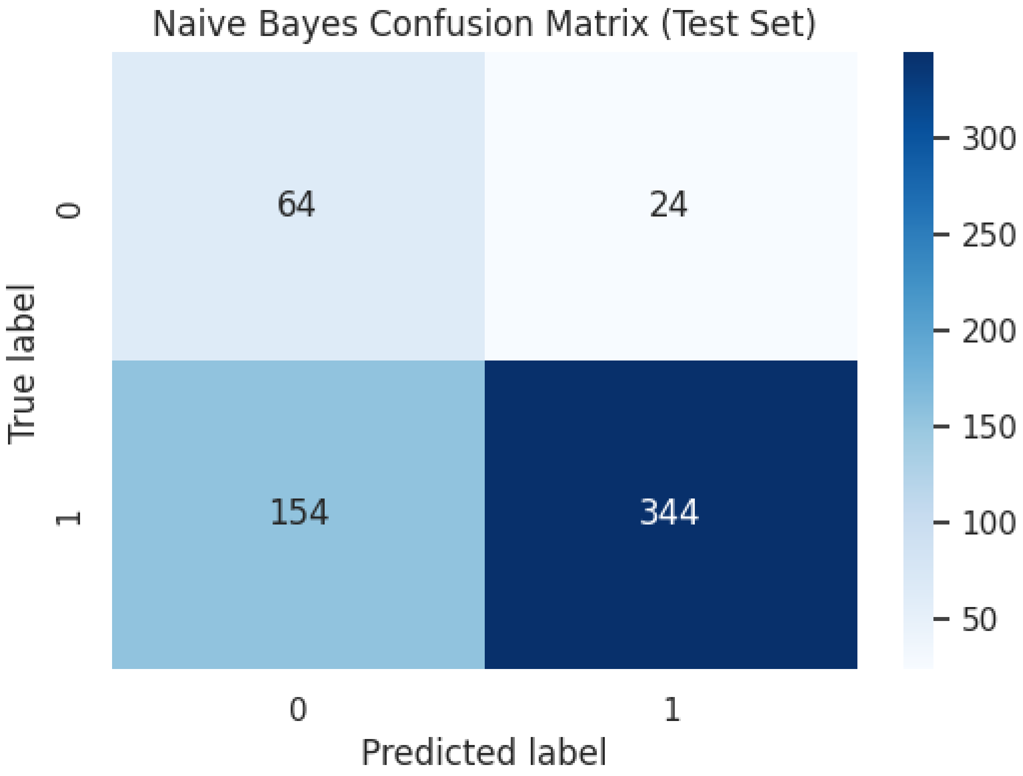


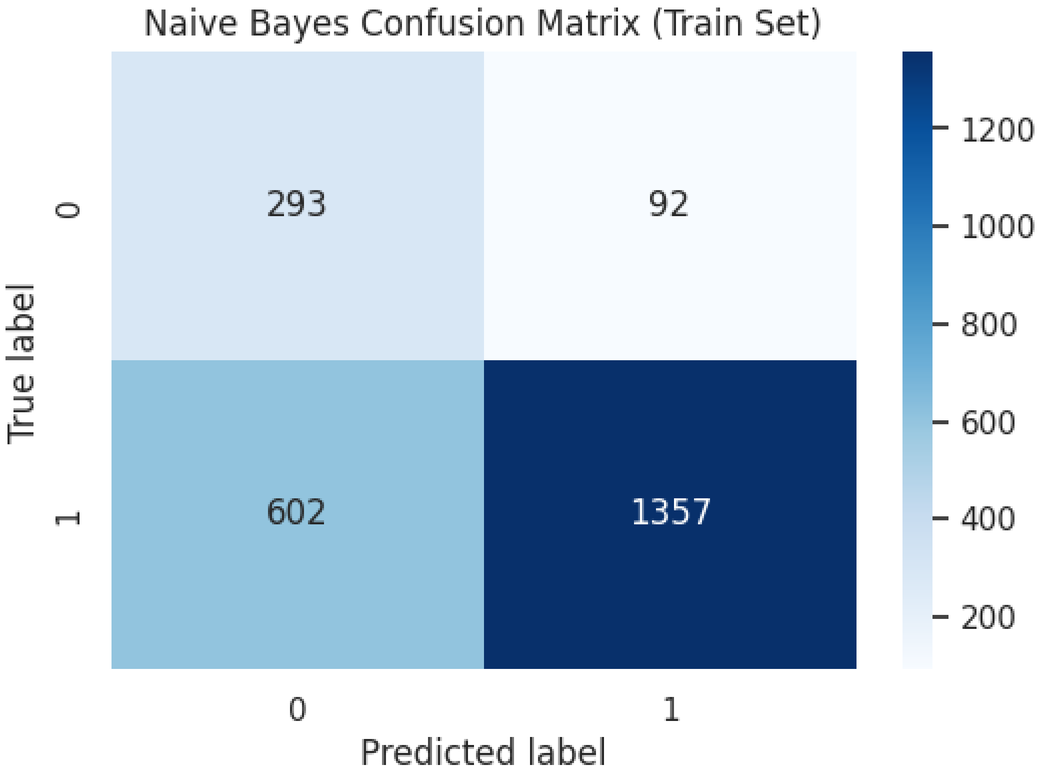
```
K-Nearest Neighbors (Test Set):
Accuracy: 0.8208
ROC AUC: 0.5437
Confusion Matrix: [[  4  84] [ 21 477]]
Precision: 0.8503
Recall: 0.9578

K-Nearest Neighbors (Train Set):
Accuracy: 0.8503
ROC AUC: 0.8284
Confusion Matrix: [[  66  319] [  32 1927]]
Precision: 0.8580
Recall: 0.9837
```

Naive Bayes

Naive Bayes is a probabilistic classifier known for its simplicity and efficiency in various machine learning tasks. It assumes independence among features, hence "naive," yet often performs surprisingly well in practice. It calculates the probability of each class given the input features using Bayes' theorem, making it effective for text classification and spam filtering. Naive Bayes models are fast to train and require minimal data to estimate parameters, making them suitable for large datasets. However, they can struggle with complex relationships between features and may oversimplify the real-world data, impacting accuracy in some scenarios. Despite these limitations, Naive Bayes remains a popular choice for its balance between performance and computational efficiency.







Naive Bayes (Test Set):

Accuracy: 0.6962

ROC AUC: 0.7684

Confusion Matrix: [[64 24] [154 344]]

Precision: 0.9348

Recall: 0.6908

Naive Bayes (Train Set):

Accuracy: 0.7039

ROC AUC: 0.7825

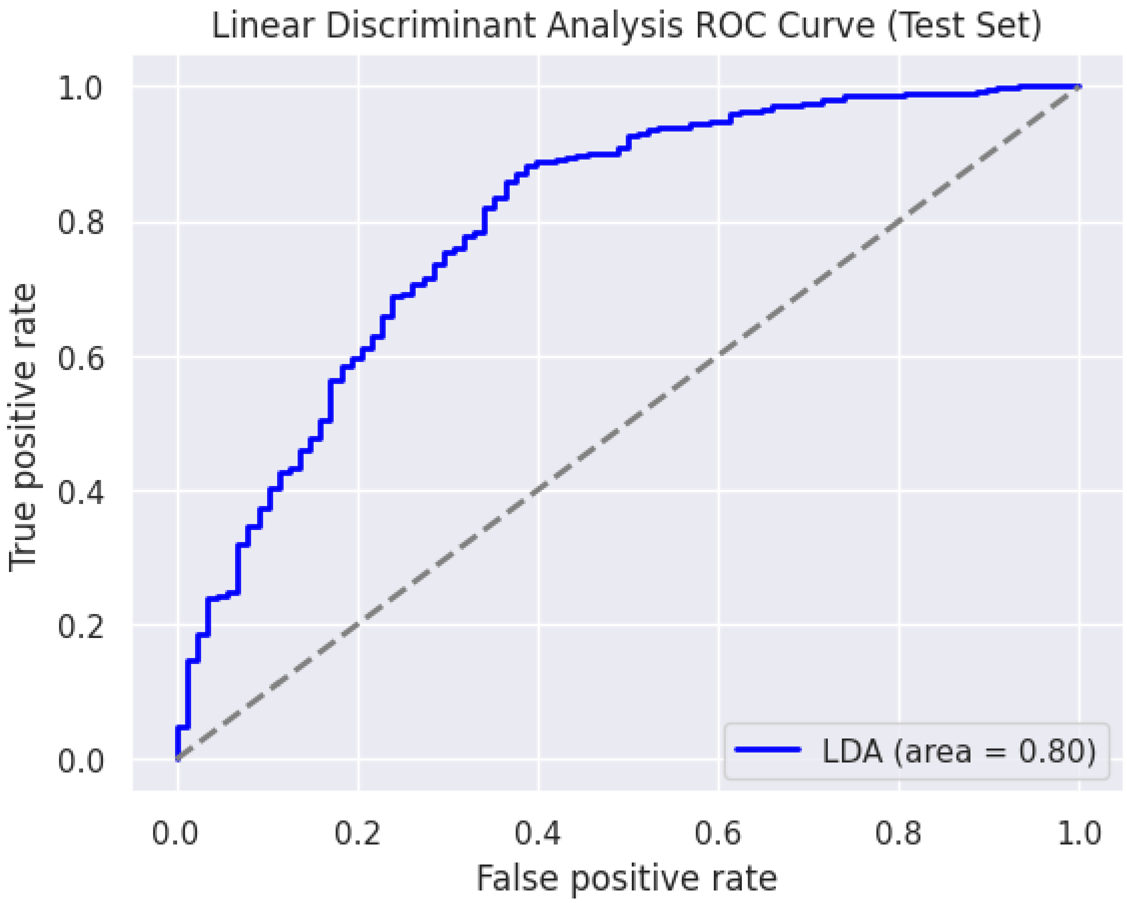
Confusion Matrix: [[293   92] [602 1357]]

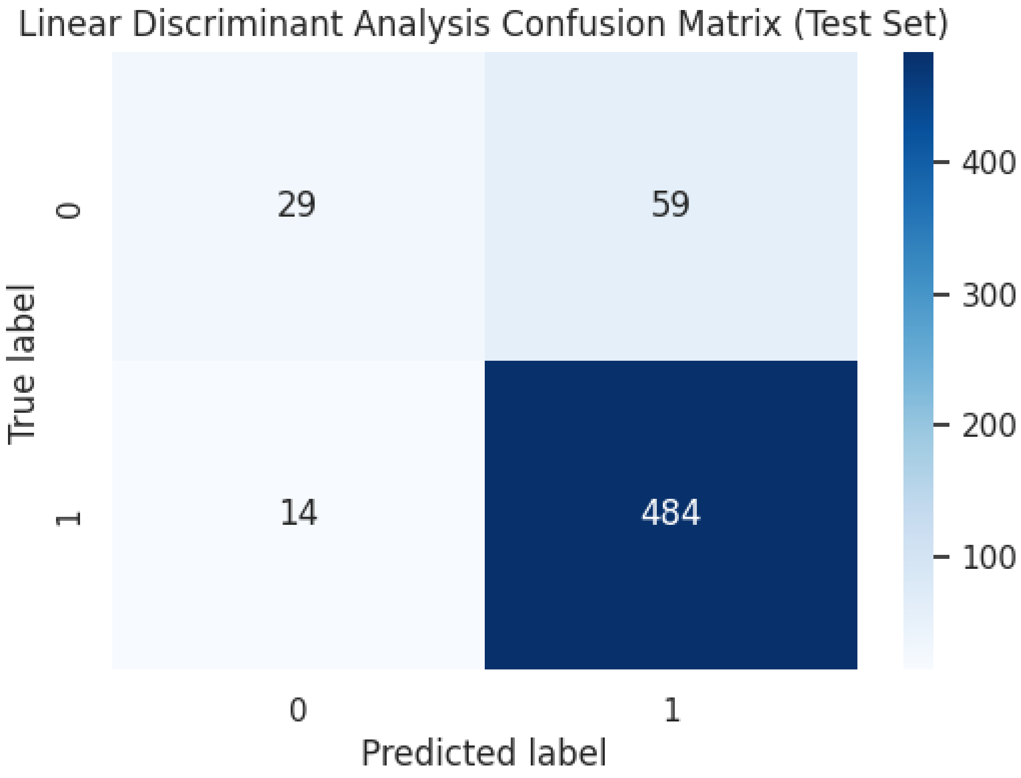
Precision: 0.9365

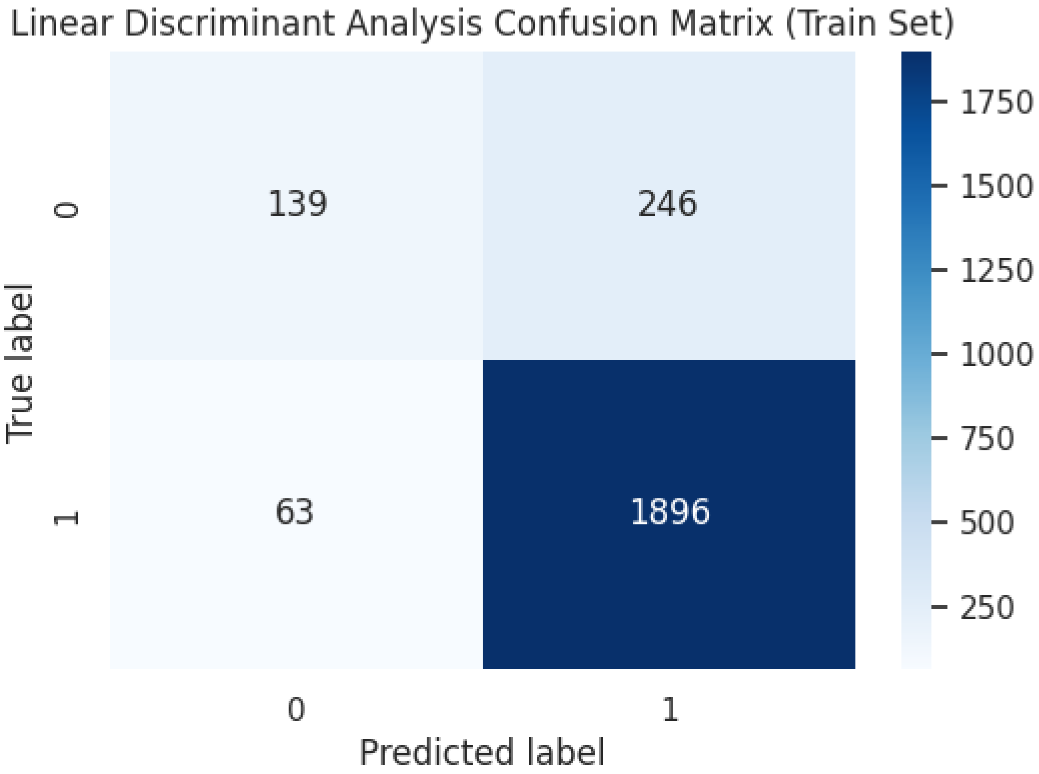
Recall: 0.6927

Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a statistical method used for dimensionality reduction and classification. It aims to find a linear combination of features that best separates two or more classes in a dataset. LDA assumes that the data are normally distributed and that the classes have identical covariance matrices. By maximizing the ratio of between-class variance to within-class variance, LDA finds a projection that optimally discriminates between classes. This projection can be used for dimensionality reduction by reducing the data to a lower dimensional space while preserving class discriminatory information. In classification tasks, LDA constructs decision boundaries based on these projections, assigning new instances to classes based on their proximity to class centroids in the reduced space. LDA is particularly effective when classes are well-separated and assumptions about data distribution hold true, making it a valuable tool in pattern recognition and machine learning applications where class discrimination and dimensionality reduction are key objectives.







```
Linear Discriminant Analysis (Test Set):
Accuracy: 0.8754
ROC AUC: 0.7992
Confusion Matrix: [[ 29  59] [ 14 484]]
Precision: 0.8913
Recall: 0.9719

Linear Discriminant Analysis (Train Set):
Accuracy: 0.8682
ROC AUC: 0.8290
Confusion Matrix: [[ 139  246] [  63 1896]]
Precision: 0.8852
Recall: 0.9678
```

---

2) Model Tuning and business implication

- a. Ensemble modelling (if necessary)
- b. Any other model tuning measures (if applicable)
- c. Interpretation of the most optimum model and its implication on the business

Train Metrics

Model	Train Accuracy	Train ROC AUC	Train Precision	Train Recall
Logistic Regression	0.8592	0.8101	0.8711	0.9760
Decision Tree	1.0000	1.0000	1.0000	1.0000
Random Forest	**1.0000**	**1.0000**	**1.0000**	**1.0000**
Support Vector Machines	0.8358	0.5805	0.8358	1.0000
K-Nearest Neighbors	0.8503	0.8284	0.8580	0.9837
Naive Bayes	0.7039	0.7825	0.9365	0.6927
Linear Discriminant Analysis	0.8682	0.8290	0.8852	0.9678

Test Metrics

Model	Test Accuracy	Test ROC AUC	Test Precision	Test Recall
Logistic Regression	0.8703	0.7718	0.8809	0.9799
Decision Tree	0.9505	0.9101	0.9537	0.9679
Random Forest	**0.9693**	**0.9790**	**0.9651**	**1.0000**
Support Vector Machines	0.8498	0.5668	0.8498	1.0000
K-Nearest Neighbors	0.8208	0.5437	0.8503	0.9578
Naive Bayes	0.6962	0.7684	0.9348	0.6908
Linear Discriminant Analysis	0.8754	0.7992	0.8913	0.9719

The Best Model (Random Forest)

Metric	Train Set	Test Set
Accuracy	0.8609	0.8652
ROC AUC	0.8336	0.7920
Precision	0.8737	0.8802
Recall	0.9745	0.9739

- High accuracy achieved by the random forest model on both test (0.8652) and train sets (0.8609), indicating minimal overfitting.
- Lower ROC AUC score on the test set (0.7920) compared to the train set (0.8336) suggests room for improvement in discriminating between classes on unseen data.
- Visualize confusion matrices for test and train sets using heatmaps to inspect true/false positives and negatives.
- Plot ROC curve for test set to show model's sensitivity/specificity trade-off.
- High recall scores on test (0.9739) and train sets (0.9745) indicate effective positive class identification.
- Reasonably high precision scores on test (0.8802) and train sets (0.8737) suggest good negative class identification capability.

Tuning best model using Randomized SearchCV

Test Set Evaluation

Class	Precision	Recall	F1-score	Support
0	0.61	0.28	0.39	88
1	0.88	0.97	0.92	498
accuracy	-	-	0.86	586
macro avg	0.75	0.63	0.66	586
weighted avg	0.84	0.87	0.84	586

Training Set Evaluation

Class	Precision	Recall	F1-score	Support
0	0.68	0.31	0.42	385
1	0.88	0.97	0.92	1959
accuracy	-	-	0.86	2344
macro avg	0.78	0.64	0.67	2344
weighted avg	0.84	0.86	0.84	2344

Predictions

1. **Test match with England in England (Day Matches) - Rainy season expected**

   Predicted outcome: 1

   Predicted probability of positive outcome: 0.8706

2. **T20 matches with Australia in India (Day/Night Matches) - Winter season expected**

   Predicted outcome for T20 match: 1

   Predicted probability of positive outcome for T20 match: 0.7285

3. **ODI matches with Sri Lanka in India (Day/Night Matches) - Winter season expected**

   Predicted outcome for ODI match: 1

   Predicted probability of positive outcome for ODI match: 0.7568

Prediction for next 10 matches

Match	Type	Country	Time	Weather	Result
1	Test	England	Day	Rainy	Win
2	T20	Australia	Day/Night	Winter	Win
3	ODI	Srilanka	Day/Night	Winter	Win
4	ODI	Srilanka	Day/Night	Winter	Win
5	T20	Australia	Day/Night	Winter	Win
6	ODI	Srilanka	Day/Night	Winter	Win
7	Test	England	Day	Rainy	Win
8	ODI	Srilanka	Day/Night	Winter	Win
9	T20	Australia	Day/Night	Winter	Win
10	Test	England	Day	Rainy	Win

Implications on Business

Insights

1. The feature "Encode_Match_light_type_1" denotes whether matches are Day/Night, crucial for predicting India's winning chances due to its highest feature importance score.
2. Conversely, the feature "Opponent" holds less significance in predicting match outcomes, as indicated by its lower impact.

3. **Model Accuracy and ROC AUC :** The Random Forest model stands out with exceptional performance metrics:
4. it achieves the highest test accuracy of 96.93% and a robust ROC AUC of 97.90%. These results underscore its reliability and effectiveness in predicting match results.
5. **Overfitting Concerns:** Concerns about overfitting are evident in both Decision Tree and Random Forest models, with perfect 100% accuracy on the training set. However, the Random Forest model shows resilience with strong performance on the test set, mitigating overfitting risks.
6. **High Recall Models:** Models such as Support Vector Machines (SVMs) and Random Forests exhibit high recall rates, highlighting their ability to effectively identify positive instances, which is crucial in scenarios where false negatives are costly.
7. **Variance in Model Performance:** There is notable variance in model performance across different algorithms. For instance, Naive Bayes and SVMs show lower test ROC AUC scores compared to tree-based methods and Logistic Regression. This variability suggests that certain models are better suited for this specific dataset and classification problem.

Strategic Benefits

1. **Operational Efficiency:** Deploying the Random Forest model streamlines operations through accurate and reliable predictions, leading to better resource allocation and reduced operational costs.
2. **Customer Satisfaction:** Improved model performance translates to better service quality. Accurate recommendation systems, for instance, enhance user experience, leading to higher customer satisfaction and loyalty.
3. **Profitability:** The combination of reduced false positives and comprehensive detection of true positives can significantly enhance profitability. In fraud detection, this means saving on losses from fraudulent activities while minimizing the cost of investigating false alarms.
4. **Scalability:** Random Forest models can handle large datasets and complex interactions between features, making them scalable solutions for growing businesses with increasing data volumes.
5. **Competitive Advantage:** Implementing a highly accurate and reliable model provides a competitive edge by enabling more precise and timely business decisions.

Recommendations

- The Indian cricket team should focus on enhancing their performance in offshore matches, where their win percentage is currently lower. Alternatively, increasing the number of home matches could improve their overall win rate.
- The BCCI should strategize to schedule more One Day International (ODI) matches during the winter season, while avoiding T20 formats during the summer months. Additionally, emphasizing Test format matches during summer can potentially boost their win rates.
- Increasing the frequency of "Day/Night" matches could significantly improve India's win percentage, aligning with favorable match conditions and enhancing performance.
- Based on predictive analytics, India shows a higher probability of winning against England, suggesting strategic advantages in upcoming matches against this opponent.

- However, the predicted win probabilities against Australia and Sri Lanka are lower, around 75%. This indicates that India may face tougher challenges against these teams, requiring focused strategies to improve their chances of winning.