Durham University

Supervised and Unsupervised Learning for Post-Hoc Evaluation of Durham County Cricket's U15 Selection Decisions

Mukund Ranjan Tiwari, MDS

Master of Data Science, Department of Natural Sciences

Van Mildert College

Durham University, Durham, England

5th September 2025

Supervisors: Dr. Amitabh Trehan and Dr. Peter G. Swift

Banner ID - Z0207662

Acknowledgements

I would like to thank my supervisors, Dr. Amitabh Trehan and Dr. Peter Swift, as well as my parents, Dr. Anu Khanna (my mother) and Dr. Heeraman Tiwari (my father).

Abstract

This dissertation explores how supervised and unsupervised machine learning techniques can be used to evaluate Durham County Cricket's Under-15 (U15) player selection process. The study focuses on a group of 50 players who took part in U15 trials in 2024, with 20 of them being selected. By analysing batting, bowling, and fielding data from the 2025 season, a multi-stage analytics pipeline was developed to examine selection patterns and identify potentially overlooked talent.

The analysis starts with exploratory data analysis (EDA), including Principal Component Analysis (PCA) and K-means clustering, to identify performance patterns and classify players into meaningful archetypes. An ensemble of XGBoost classifiers was then employed to estimate selection probabilities and compare algorithmic recommendations with Durham's actual selections. To enhance interpretability, SHAP values were used to explain the model's outputs and emphasise the most influential performance metrics. Finally, an interactive Streamlit dashboard was created to visualise results and offer coaches and analysts an accessible decision-support tool.

Findings indicate that selection decisions are heavily influenced by both overall match involvement and specialist roles, such as wicketkeeping. The models identified several players with strong performance profiles who were not chosen, implying potential gaps in the current process. This study illustrates that machine learning can complement expert judgement, offering a more transparent and evidence-based approach to talent identification. It also provides a reproducible pipeline that can support fairer and more consistent decision-making in grassroots cricket.

Table of Contents

1. Introduction

The identification and development sporting talent, particularly at the junior levels, is a tricky process. Traditionally, coaches, selectors, scouts and analysts use subjective assessments to judge the suitability of young players to progress to the upper echelons of youth programs. These assessments rely on subjective judgements of technique, skill, flair, game awareness, and a variety of physical attributes. These assessments also lead to coaches forming ideas of perceived potential, often relying on "gut feelings" rather than hard, numeric evidence. It is, therefore, reasonable to say that such selection processes are liable to inherent biases and misjudgments, variability between selectors and unavoidable inconsistencies. Whilst such judgements can indeed produce rich contextual insights, they run the risk of overlooking players who may have consistent performances and reliable contributions but have not caught the selectors' eyes.

Over the past decades, the growing influence of data analytics and machine learning (ML) has begun to transform decision-making processes across professional sport. The use of data analytics was largely heralded by the growth of the field of Sabermetrics, coined by Bill James and the Society for American Baseball Research (SABR). All these analytical methods came into the spotlight during the 2002 MLB (Major League Baseball) season, when the Oakland A's, under General Manager Billy Beane, signed a bevy of seemingly undervalued players and built a championship-winning team through the power of analytics alone. Today, major sports across the world utilise large, highly detailed datasets and massive teams of analysts and machine learning engineers to understand the sport at the microscopic level. In cricket, these technologies enable the integration of multiple performance indicators – spanning batting, bowling, and fielding metrics – into systematic, evidence-based evaluations. Data-driven approaches can reveal underlying patterns and relationships between player attributes that may not be immediately visible to human observers, providing new opportunities for enhancing accuracy and fairness in selection decisions.

Despite these advances, there is limited research that explores the applications of ML-based analytical techniques to junior cricket. Much of the work focuses on elite leagues and international games from around the world but rarely talks about the potential impact it can have in seeking out young talent and change the way we think about player development pathways. With smaller and less detailed datasets, fewer resources, and limited opportunities, grassroots cricket analytics offers an interesting challenge. This study aims to fill that gap by exploring how such methodologies can try and decipher the combination of performance metrics that define a desirable player. By using a combined supervised and unsupervised learning framework, this dissertation critically evaluates Durham County's Under-15 (U15) selection process. The study focuses on a cohort of 50 players who attended U15 trials last year, of whom 20 were selected for advancement into the U15 squad. By leveraging batting, bowling, and fielding statistics collected from the 2025 season, the study applies a comprehensive analytics pipeline designed to both replicate and challenge historical selection outcomes.

The primary aim of this study is to investigate the alignment between coach-led selection decisions and algorithmic predictions, exploring where these two approaches converge and diverge. To achieve this, the research is guided by the following objectives:

- To conduct a two-stage Exploratory Data Analysis (EDA) consisting of a basic statistical analysis of the dataset followed by a Principal Component Analysis (PCA) and Clustering stage to understand performance patterns across batting, bowling and fielding metrics whilst identifying player archetypes and player groupings.
- To understand how these performance patterns and player groupings affect selection decision and earmarking potentially overlooked talent.

- To develop an ensemble of XGBoost classifiers to predict selection probabilities based on player performance data and compare model outputs with Durham's actual U15 selections.
- To enhance model interpretability via SHAP (SHapley Additive exPlanations) and indetify performance metrics that most influence selection.
- To build an interactive Streamlit dashboard that visualises analytical findings, enabling coaches and analysts to explore insights dynamically and apply them to future decision-making.

As detailed above, this study employs a multi-stage analytics pipeline that integrates supervised and unsupervised learning techniques. After an initial visual inspection of the dataset via various statistical plots such as a distribution matrix, heatmap, pair plot, etc., the study utilises a PCA and Clustering algorithm. PCA reduces dimensionality and identifies key metrics driving performance differences, while clustering finds natural player groupings to demarcate player archetypes and their value to squad selections. Building on this, a robust XGBoost classifier models complex interactions between batting, bowling, and fielding metrics and predicts selection outcomes. These predictions help generate a list of top 20 players, which is then compared with the top 20 players from last year's trials, allowing for a thorough validation of selection methodologies. To boost stakeholder confidence and aid in the understanding of what drives selections, SHAP quantifies the influence of individual metrics on selection probabilities, thus bridging the gap between model outputs and coaching interpretations. Finally, all the findings from this multi-stage pipeline are presented via a dynamic and interactive Streamlit dashboard, supporting practical integration into the selection process as an accessible tool for coaches and analysts.

By comparing model-derived recommendations with historical U15 selections, this dissertation assesses Durham's current selection framework's efficacy and identifies instances where players with good performance profiles might have gone unnoticed. Through these methods, the study shows how data-driven methods can reinforce and complement expert intuition rather than take its place, providing a more impartial and balanced basis for judgement. The results offer a reproducible framework for using machine learning to identify junior talent while encouraging equity, openness, and accountability in player assessments, which has wider ramifications for grassroots cricket development.

2. Literature Review

The process of identifying and developing sporting talent, particularly at the youth level, has long posed significant challenges for coaches, scouts, and performance analysts. Traditionally reliant on subjective assessments and coach intuition, talent selection has more recently become a focal point for empirical research and data-driven methodologies. In the context of county-level cricket, especially at the U15 level, the stakes are high – with early decisions influencing long-term athlete development pathways. This literature review explores the key themes underpinning these election processes through four main areas - talent identification and development, the role of human decision-making in selection processes as compared to ML models, and the growing contribution of machine learning and statistical modelling in supporting or enhancing these judgements. The review then explores existing supervised and unsupervised learning pipelines to determine the best path forward with the data at hand. Together, these strands provide a critical foundation for the development of an integrated, evidence-based selection pipeline tailored to youth cricket.

2.1 Talent Identification in Youth Sport

The process of identifying athletic talent amongst young players has evolved from basic physical testing to complex, multi-dimensional models. Vaeyens et al. (2008) argue for the integration of physical, technical, tactical, and psychological assessments, noting that one-off trials are insufficient for evaluating long-term potential. Consistent with this, Dugdale et al. (2020) emphasise the limited value of isolated physical tests and advocate for ecologically valid, sport-specific evaluations.

In practice, selection is often influenced by subjective judgments. Larkin & O'Connor (2017) highlight selectors' reliance on perceived decision-making and 'coachability,' while also acknowledging inconsistencies in scouting methods. Meanwhile, Till et al. (2016) demonstrate that early physical maturity can bias selectors in youth rugby, advocating for bio-banding to support late developers. In a cricket-specific context, Phillips et al. (2010) show that technical skills outperform fitness metrics in predicting selection outcomes, reinforcing the need for bespoke testing protocols tailored to the sport. Collectively, these studies converge upon an undeniable notion: effective talent identification in youth sport requires a balanced approach combining objective testing, game-relevant skills, and longitudinal monitoring — all of which inform the data-driven model developed in this dissertation.

2.2 Machine Learning in Sports Performance Analytics

The application of machine learning (ML) techniques within sports performance analytics has gained significant traction in recent years, offering unique approaches to player evaluation, tactical analysis, and outcome prediction. These methodologies enable systematic, data-driven decision-making, complementing or, in some cases, challenging traditional talent identification ideologies that are grounded in expert intuition. Within the context of youth cricket, such as Durham's U15 development pathway, ML provides a means to assess player potential and performance trajectories using a scalable and objective framework.

Bunker and Thabtah (2019) present a generalisable ML framework for predicting sport results, demonstrating the superior performance of ensemble learning methods, such as Random Forest and Gradient Boosted Trees, over traditional statistical calculations. Their findings highlight the importance of algorithm selection and feature engineering in constructing predictive systems capable of capturing non-linear relationships and underlying variable interactions in sports data.

This foundational work sets a precedent for using ML pipelines in varied sporting contexts, including youth talent assessment.

In parallel, Gudmundsson and Horton (2017) offer a thorough review of spatio-temporal analysis techniques in team sports. Their survey illustrates how clustering algorithms, trajectory mining, and temporal sequence modelling can uncover hidden structures in movement and positioning data. While primarily applied to sports such as football and basketball, these approaches hold potential for adaptation in cricket, for example in analysing player fielding patterns or movement during trials. The methodological insights provided are valuable for developing unsupervised learning pipelines in environments where data is constrained.

Injury forecasting has also emerged as a prominent domain for the application of ML-based techniques. Rossi et al. (2018) utilise GPS-derived training data to predict injury risk in professional footballers using supervised learning models. Their emphasis on the importance of long-term monitoring and the inclusion of cumulative load metrics demonstrates the value of high-resolution temporal data in sports-related predictions. Although the direct applicability to cricket may be limited by differences in injury type and measured metrics, their framework shows how ML-based methods can be used for a variety of prediction and classification tasks in professional sport.

In the context of cricket-specific applications, Chakraborty et al. (2024) employ various ML classifiers to predict T20 match outcomes using a range of engineered features, including recent form, venue-specific statistics, and head-to-head performance records. Their modelling workflow – comprising data preprocessing, feature selection, class balancing, and performance evaluation – showcases best practices in predictive modelling. While focused on match-level forecasting, their methods are adaptable for player-centric evaluation, particularly when seeking to quantify contextual performance factors. More directly relevant to individual player assessment (especially at the youth level) is the work of Kiran Babu et al. (2025), who develop supervised models – such as decision trees, support vector machines, and Naïve Bayes classifiers – to predict cricket player performance. Their inclusion of multiple evaluation metrics, including precision, recall, and F1-score, highlights the importance of multidimensional model validation in the context of talent identification. Their study demonstrates the feasibility of implementing data-driven player evaluation frameworks even in data-sparse settings such as youth trials, and makes a strong case for the use of tree-based ensemble methods.

Further expanding the methodological scope, Jianjun et al. (2025) propose a performance modelling framework that integrates biometric and gameplay data using ensemble algorithms and deep learning architectures. Their findings suggest that incorporating physiological variables significantly enhances predictive accuracy. Although the integration of biometric data may be impractical in grassroots settings, their approach offers a solid grounding in the ability of ML algorithms to perform predictive analysis of sport data, and highlights how more detailed datasets allow for the application of more complex ML techniques, leading to greater accuracy in results.

A central tension within ML-based sports analytics lies in balancing model interpretability with predictive power. Highly interpretable models – such as decision trees or linear classifiers – offer transparency and ease of communication with coaches and analysts but may lack the capacity to capture complex, non-linear relationships. In contrast, ensemble methods and neural networks typically achieve superior predictive accuracy at the cost of model transparency, making it difficult to explain their outputs to non-technical users. Furthermore, more complex techniques require large and highly detailed datasets. The selection of ML models must account not only for

performance metrics and the end goals of the analysis, but also for more practical considerations – such as the quality of the limited data at hand and the need for transparency and interpretability for non-technical stakeholders.

Collectively, the reviewed studies illustrate the growing applicability and acceptance of ML in sports performance analytics and its relevance to youth talent evaluation. The convergence of predictive modelling, spatio-temporal analysis, and multimodal data integration presents a robust foundation for developing systems that can support and enhance traditional coach-led decision-making processes. Within the context of Durham's U15 cricket selection, these methodologies offer a means of critically assessing existing practices and exploring how ML-based techniques can add value to existing practices.

2.3 Data and Human Judgement

The increasing application of machine learning and data analytics in sport has led to a growing debate surrounding the value of algorithmic decision-making versus traditional human-centric assessments. This section critically evaluates literature comparing the two approaches, with particular attention on youth sport selection. The focus of this section is to compare these approaches based on their respective strengths and limitations, whilst seeking best practices for the application of such technologies on the field.

As discussed earlier, historically, talent identification has relied heavily on the subjective judgments of coaches and scouts. Larkin and O'Connor (2017), in a qualitative investigation of selection practices in youth football, highlighted that selectors predominantly assess attributes such as game intelligence, psychological resilience, and technical proficiency through physical observation. While such methods provide rich contextual insight, they are also prone to cognitive biases and inconsistencies. For instance, there is evidence to suggest that subjective evaluations may be influenced by physical maturity, prior exposure, or stereotypes around specific positions in the game, particularly in junior age groups where the variability between players is high as young players go through great physical and mental changes at the biological level.

In contrast, recent developments in artificial intelligence (AI) and machine learning promise a more objective, data-driven alternative to selection. Zhou et al. (2025) provide a comprehensive narrative review of the use of AI in sport, throwing light on its potential for enhanced predictive accuracy through the analysis of large-scale multimodal sports data. These models can detect hidden patterns and interactions that may escape human observation, enabling more evidence-based decisions. However, their performance is dependent on the quality of the input data, which can in of itself reflect underlying biases in talent development systems.

Several studies have explored the integration of algorithmic methodologies into the talent identification process. Monsees (2025), drawing on expert interviews at the hight levels of youth sport, noted a growing acceptance towards data-based selection but highlighted the fact that many in the field still had their doubts. Coaches expressed concerns over the perceived "black box" nature of predictive models and their limited capacity to account for unmeasurable quantities such as motivation, adaptability, character and game knowledge. This concern underscores a fundamental trade-off in talent identification: trusting the predictive power of a purely numbers-based system vs understanding how and why it gives certain outputs. While machine learning models can yield superior performance when making predictions, they often lack the transparency and intuition of human decision-makers. This makes many non-technical stakeholders of the talent

identification process wary of trusting these seemingly vague and undecipherable technologies. Furthermore, the interpretability of algorithmic systems is not merely a technical challenge but a sociopolitical one. As talent pathways shape athletes' futures, decisions must be justifiable to athletes, parents, and stakeholders. In this context, hybrid models – where algorithmic outputs are amongst one of the many considerations when making final decisions – may offer a logical compromise. Such approaches have gained traction in high-performance settings, where data is used to complement and reinforce expert assessment, not replace it (Nijenhuis et al., 2024).

Nonetheless, algorithmic methods contain biases of their own. Data-driven models trained on historical selection outcomes risk perpetuating existing imbalances if past decisions were biased themselves. This is especially problematic at the youth level where performance data may be sparse, inconsistent, or skewed due to several factors. When deploying such tools in practice, it is important to keep this context in mind. Emerging discourse also emphasizes the importance of "epistemic humility" for both proponents and detractors of algorithmic methods. Neither coaches nor algorithms possess a monopoly on accuracy; rather, both have their own flaws and pitfalls. As the LinkedIn article "The Role of Big Data in Talent Scouting and Analytics" argues, the future of selection may lie not in choosing between human and machine but in developing pipelines that involve collaborations between the two. Here, human experts can offer domain-specific insights and contextual reasoning, while algorithms can enhance consistency and earmark undervalued and underappreciated players, otherwise overlooking by traditional means.

In the context of this dissertation's focus, the tension between subjective and data-driven selection is of particular importance. Cricket selection is often influenced by technical skills and coach intuition, but early identification can carry long-term consequences. Given the variability in maturation and skill development at this age, the inclusion of data-driven methods, particularly unsupervised and supervised learning pipelines, offers an opportunity to critically evaluate and potentially enhance selection fairness and validate existing selection practices.

In summary, the literature presents a complex view of human versus algorithmic talent identification. Data science offers scalability, consistency, and deeper pattern recognition. On the other hand, expert judgment provides contextual decision-making and a safeguard against purely statistical picks, ensuring that the "human" element of the sport is not replaced by robots chasing numeric milestones. The challenge for modern sport analysts, selectors and coaches, therefore, is not to replace human insight, but to design systems where algorithmic and expert perspectives enrich and reinforce one another.

2.4 Modelling Pipelines and Feature Engineering for Sports Data

Machine learning (ML) techniques in sport talent identification require technical expertise and careful methodological planning. Recent studies offer a solid basis for understanding best practices in developing reliable modelling pipelines for performance prediction and player classification, particularly in the context of U15 cricket.

A strong pipeline starts with thorough data preparation. Chicco et al. (2022) emphasise that feature engineering and data cleaning are some of the most vital yet often neglected stages in ML workflows. Common problems in sports datasets, such as missing values, outliers, inconsistent time sampling, and multicollinearity, can affect both training and validation processes. Imputation methods, such as Multiple Imputation by Chained Equations (MICE), have been successfully applied in sports science applications to address these issues. Furthermore, because performance

datasets are multidimensional – often covering a wide range of performance-related and contextual variables – techniques like Principal Component Analysis (PCA) and t-SNE are frequently employed to mitigate the "curse of dimensionality"[1] and enhance clustering or classification outcomes.

Identifying meaningful, context-specific features is especially difficult in youth talent discovery pathways. Unlike pro-level data, which may include notable events (such as a ball-by-ball record) or tracking information, grassroots data is often limited in both range and detail. Nonetheless, carefully developed calculated metrics – such as bowling averages, economy rates, batting/bowling strike rates, or peer-relative physical scores – can lead to valuable predictive outputs when appropriately contextualised.

Supervised learning models have become a key part of prediction pipelines in athlete performance modelling. Qin et al. (2025) used ensemble methods such as XGBoost and Random Forests to predict performance across multiple physiological areas, showing their ability to understand nonlinear interactions within diverse datasets. These methods are directly useful for selection tasks in youth cricket, where binary classification (such as "selected" vs. "not selected") is usually the main objective. However, model performance must be balanced with interpretability, especially when used by coaches and selectors. Techniques like SHAP values and feature importance rankings are increasingly used to produce explanations that stakeholders can understand.

Unsupervised learning methods provide complementary capabilities, especially when labelled data is scarce or ambiguous. Clustering algorithms such as k-means, DBSCAN, and hierarchical clustering are used to group players based on multi-metric profiles, uncovering hidden patterns within player populations. Ajay (2021) demonstrated how clustering techniques can identify performance-based archetypes within cricket data. These methods are particularly valuable in developmental contexts where long-term outcome labels are either unavailable or unreliable, allowing coaches to discover emerging groupings and underappreciated talent. Dimensionality reduction is often combined with these methods to improve clustering quality or to enhance the visual interpretability of high-dimensional performance data.

Modern pipelines increasingly integrate both supervised and unsupervised techniques in iterative or hybrid workflows. In such systems, clustering can be used to first identify player types, with subsequent supervised models trained to predict outcomes for each cluster, thus tailoring predictions to specific developmental trajectories. Kuhn and Johnson (2016) stress the importance of pipeline tuning, validation on independent cohorts, and automation using tools such as *caret*, *mlr*, or *scikit-learn*. These principles are directly applicable to youth cricket, where dynamic and evolving datasets necessitate flexible yet rigorous model development processes.

The ethical and practical challenges associated with algorithmic talent identification cannot be overlooked. Kim et al. (2025) warn that AI systems might unintentionally reinforce existing selection biases, especially if trained on historical data that reflects deep-rooted inequalities. In youth sport, where selection outcomes can greatly influence future prospects, ensuring transparency, fairness, and contextual understanding is essential. Therefore, modelling pipelines should be designed not only for accuracy but also for accountability – supporting, rather than replacing, expert decision-making based on on-field observations.

[1] The "curse of dimensionality" is a concept in the field of data science, ML and statistics. It is a phenomenon where the efficiency and usability of data and algorithms rapidly deteriorate as the number of dimensions grows, causing sparsity and making meaningful pattern discovery, computation, and generalisation much harder in high-dimensional spaces.

3. Methodology

3.1 Introduction

This study aims to analyse the performance and selection of U15 cricket players based on historical data from the previous year's U15 trialists. The datasets include batting, bowling, and fielding statistics collected during the current season. Initial data cleaning and preprocessing steps were applied to ensure consistency across the datasets, including standardising player names, handling missing values, and aligning records based on player identifiers. The dataset was labelled to mark players who were selected in last year's trials.

Exploratory Data Analysis was then conducted in two stages. For the first stage, scatterplots, heatmaps, boxplots, and line graphs were used to analyse player performance and gain a general understanding of the dataset's structure. Then, for the second stage, Principal Component Analysis (PCA) was used to reduce dimensionality and identify key patterns in player performance. K-Means clustering was applied to group players into distinct performance archetypes.

For predictive modelling, an ensemble XGBoost classifier was trained to predict player selection status based on performance metrics. Model interpretability was enhanced using SHAP (SHapley Additive exPlanations) to identify the most influential features driving selection decisions. The model's predictions were then compared to actual selection outcomes to evaluate alignment and identify potential discrepancies. Evaluation metrics including precision, recall, and F1-score were used to assess model performance, with an emphasis of F1-score due to the small size of the dataset. These findings were then presented in an accessible, dynamic and interactive Streamlit dashboard. A general overview of the pipeline can be seen here:

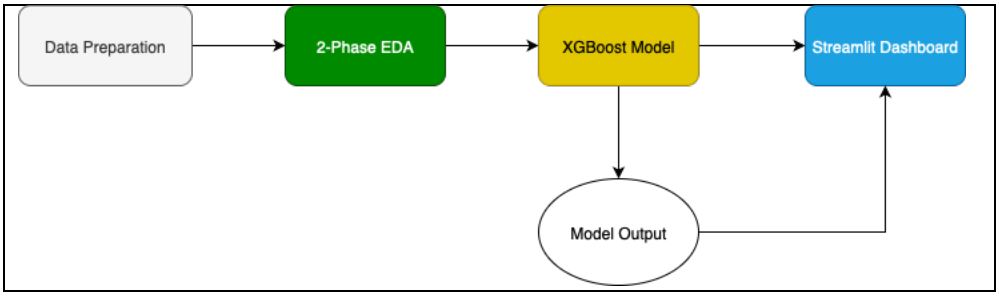

Fig 3.1 – Complete analytics pipeline flowchart

3.2 Data Sources and Data Preparation

The data was sourced from PlayCricket, an online platform licenced by the ECB, which provides statistics for all levels of local competitions involving pro and semi-pro players across England. Three datasets were downloaded from a website, containing batting, bowling, and fielding statistics for players in the Durham Cricket Board Junior League. The original, unmodified dataset included 1 1000 entries for players across various age divisions and teams in the County Durham area for the 2025 season. The data was then filtered to include only the 50 players targeted for this study- those who tried out for the U-15 team last year. The datasets were downloaded in Excel format with the ".xlsx" extension. To manipulate this data using Python, the dataset was converted from an Excel sheet into a Pandas DataFrame and subsequently merged on the "Player" column. This resulted in a new dataset containing all batting, bowling, and fielding metrics for every player. Typically, at higher levels, these datasets would be analysed separately; however, initial inspections revealed that nearly all players had

a significant amount of data for both batting and bowling, with only a few players who exclusively batted or were wicketkeepers. Therefore, comparing all metrics across disciplines was deemed appropriate. This approach also aligns with coaching perspectives, as U14 and U15 players are encouraged to both bat and bowl since most players have not yet chosen a specific discipline (or decided to become all-rounders) at this stage of their development (Brown et al., 2023). This dual focus is regarded as an important element of youth cricket for two main reasons. First, as players learn and refine skills in both batting and bowling, they discover which discipline suits them best, based on both aptitude and enjoyment. Second, even if a player is primarily a batter or bowler, playing the opposite role helps them understand the mindset of their counterparts – a batter, by actively bowling in matches, learns how a bowler plans variations for an over and reads the conditions, while a bowler, by batting longer innings, learns how batsmen counter bowling strategies and how to "read the bowler". Consequently, given the substantial amount of batting and bowling data for most players and the coaching imperative for all players to experience both roles, a combined dataset was a logical choice.

Initial dataset inspection was carried out by looking at the size and shape of the data, scanning for missing values and looking for other faults that could affect the quality of the analysis. The initial dataset had several missing values. Due to the nature of the analysis at hand, imputing values was out of the question, as it would skew the model and affect outputs in the forthcoming stages. Furthermore, as the data was collected over several weeks from multiple club pages on PlayCricket, there were some inconsistencies. The merged dataset was therefore extracted as a CSV (Comma-Separated Values) file, and the tabular data was manually edited by entering fresh values for each of the 50 players in the dataset. It was noted that some inconsistencies remained in the dataset, even after editing the CSV file with fresh data, primarily due to variations in the number of matches played by different players. An attempt was made to account for this by ensuring that each player had data for a minimum of 15 matches played for statistical fairness. This is because a smaller number of matches means that even one or two good performances can heavily bias metrics such as batting/bowling strike rates, batting/bowling averages, and even economy rates. Beyond these concerns, players with more match data will have more raw statistics, such as runs/wickets, which will also impact the analysis outcomes. This was not possible for all players – some players had simply not played that much this season at the point of data collection. As such, there was nothing more to be done. To address this, certain decisions were made regarding model tuning and feature engineering, which are detailed in the further sections of this report.

A new column, "Selected", was introduced, containing a 0 or 1 value – 1 indicating successful selection and 0 indicating unsuccessful tryouts. This resulted in the creation of binary labels for the forthcoming modelling tasks. Finally, the data was scaled for effective model performance in the PCA, clustering, and XGBoost tasks ahead. PCA and clustering algorithms, such as k-means, are highly sensitive to feature scales. Unscaled data can bias results, as variables with larger ranges tend to dominate the analysis. Standardisation techniques (such as z-score normalisation) ensure that each feature contributes equally, allowing PCA to extract meaningful directions of variance and enabling clustering methods to form accurate groupings. Although XGBoost's tree-based nature makes it less sensitive to feature scaling, normalisation can still improve convergence speed and performance, especially in cases with highly skewed or extreme values, as can be seen in that dataset. This was the first step taken to account for the nature of the dataset. As such, the preprocessing done to the dataset by feature scaling facilitates easier model comparison and interpretation, prevents algorithmic bias, and contributes to consistent processing across the data pipeline.

3.3 Exploratory Data Analysis (EDA) Pipeline

Before running a model and extracting results from the dataset, it is important to understand the complex relationships in the dataset. For this, a comprehensive EDA pipeline was designed that incorporated analysis via basic plotting, followed by PCA and Cluster analysis to guide insights from future modelling results. The entire EDA pipeline is illustrated in the visual below.

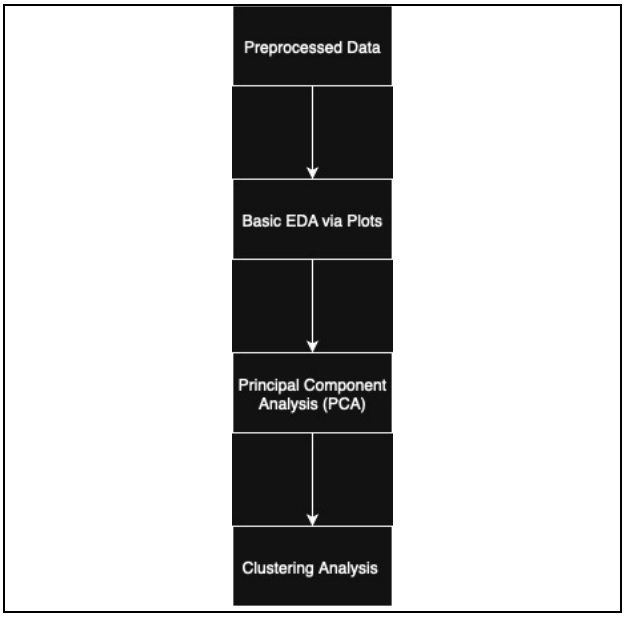

Fig 3.2 – EDA Pipeline Flowchart

3.3.1. Standard Plotting for EDA

To gain an initial understanding of player performance patterns and relationships within the dataset, standard plots were generated using Python based on the numeric data in the dataset. The plots focused on summarising key statistical properties, visualising distributions, and identifying relationships between variables relevant to player selection. First, a feature distribution matrix was generated for all numerical variables to visualise the spread and density of player performance metrics. This provided an overview of data symmetry, skewness, and the presence of potential outliers, which provided insights for subsequent PCA and clustering, as well as further preprocessing and modelling steps for the XGBoost model. Next, a correlation heatmap was created to look at linear relationships between performance metrics. The identification of highly correlated features gave an understanding of potential redundancies and guided dimensionality reduction using PCA in the next stage of the EDA pipeline. The distribution of selected vs non-selected players was then visualised to examine class balance within the dataset. This step was crucial for evaluating model performance later, as imbalanced datasets can bias classification results. This allowed the modelling steps to be modified and fine-tuned as per the inherent imbalances and limitations of the dataset.

Additionally, the top five features with the highest variance were identified, as high-variance features typically contribute more to predictive modelling. To further investigate these influential variables, pair plots were generated for the top four features, comparing their distributions across selected and non-selected players. For additional context, pair plots are a type of data visualisation that display pairwise relationships between multiple numeric variables in a dataset. They create a grid of scatter plots showing how each variable relates to every other, with histograms or density plots on the diagonal to represent the distribution of individual variables. These visualisations provided

preliminary insights into the discernibility between classes based on individual performance metrics. These visualisations provided a broad understanding of the dataset and formed the foundation for the subsequent PCA and clustering analysis performed in R, which aimed to uncover hidden performance patterns and group similarities among players.

3.3.2 Principal Component Analysis (PCA)

In the next stage of the EDA, Principal Component Analysis was employed to reduce the dataset's dimensionality and visualise player distributions using only the most significant attributes that affect selection status. The R language makes it easier to perform and visualise PCA as compared to Python. Using the *prcomp()* function to perform the necessary calculations, along with libraries such as *tidyverse*, *ggplot2,* and *ggrepel* to make clean and easy-to-understand visuals, R proves to be a superior alternative to Python for this type of analysis via unsupervised learning techniques.

Principal Component Analysis (PCA) is a technique used to simplify complex data by reducing its dimensions while preserving as much important information as possible. PCA transforms a set of potentially correlated variables into a smaller number of uncorrelated components, known as principal components, which capture the maximum variance within the dataset. Mathematically, PCA works by finding new directions (the principal components in question), which are combinations of the original variables, by computing the covariance matrix of the data, which is as

$$C = \frac{1}{n-1} X^T X$$

It then calculates its eigenvalues and eigenvectors. The eigenvectors show the directions of maximum variance – attributes (in this case, performance metrics) that have the greatest impact on selection status. The eigenvalues indicate how much variation each direction explains, *i.e.* the importance of each of the eigenvectors. This is done via eigenvalue decomposition, which is as

$$C = PDP^T$$

Where P is a matrix whose columns are the eigenvectors (principal component directions), and D is the diagonal matrix containing the corresponding eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$

The PCs calculated are orthogonal, to ensure that there is no redundancy (as mentioned earlier, the objective is to produce an *uncorrelated* set of ordered values). These are calculated as

$$\mathrm{PC}_i = \mathrm{X}w_i$$

Where w_i is the ith eigenvector. Finally, the cumulative proportion of variance explained by the first k components is calculated via the formula

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{p} \lambda_j}$$

By projecting the original data onto a smaller number of principal components, PCA allows easier visualisation and analysis without losing key patterns or trends (Jolliffe & Cadima, 2016). Essentially, it visualises the distribution of the variables of the dataset in the context of only the most important

dimensions of the dataset, accounting for only the most statistically significant factors influencing selection status.

The PCA was conducted in R as part of the exploratory data analysis (EDA) to uncover underlying patterns within the player performance dataset and identify relationships between playing attributes and selection outcomes. In this study, PCA was not applied as a feature selection mechanism for the predictive models but rather as an interpretive tool to better understand hidden performance structures and the extent to which players could be distinguished based on their performance metrics. Prior to conducting PCA, the dataset was prepared to ensure methodological fairness and accuracy of results. Player identifiers and the binary "SELECTION" variable were removed, as PCA is designed to operate exclusively on continuous numeric variables, and only performance-related metrics were retained for analysis. Variables exhibiting zero variance were excluded as they do not contribute meaningfully to the principal components. Given that the dataset comprised variables measured on different scales, z-score standardisation was applied to ensure all features contributed equally, since PCA is highly sensitive to differences in variable magnitudes. The analysis was performed using the *prcomp()* function in R, which computed the principal components, their loadings, and the proportion of variance explained by each. Visualisations were then generated to aid interpretation - a two-dimensional PCA scatter plot mapped players according to the first two principal components (PC1 and PC2), which collectively captured the most significant proportion of dataset variance, while points were colour-coded by selection status to examine potential associations between underlying performance groupings and player outcomes. Only two principal components were extracted as initial analysis indicated that the greatest variance was captured by the first two PCs alone. To further enhance interpretability, a PCA biplot was produced, displaying both player scores and variable loadings simultaneously, allowing identification of performance metrics exerting the strongest influence on each principal component. Variables with longer loading vectors were observed to have greater explanatory power, and to quantify these contributions, the top ten most influential variables for PC1 and PC2 were extracted and visualised using bar charts, highlighting the attributes that primarily drove separation between players. Finally, a scree plot was examined to determine the proportion of total variance explained by each principal component, ensuring the components retained for interpretation provided a sufficient representation of the data's structure. The scree plot validated the initial assumption of extracting only two principal components (PC1 and PC2) and thus gave robust mathematical backing to the initial decisions to extract PC1 and PC2 alone. This can be seen in the scree plot given below.

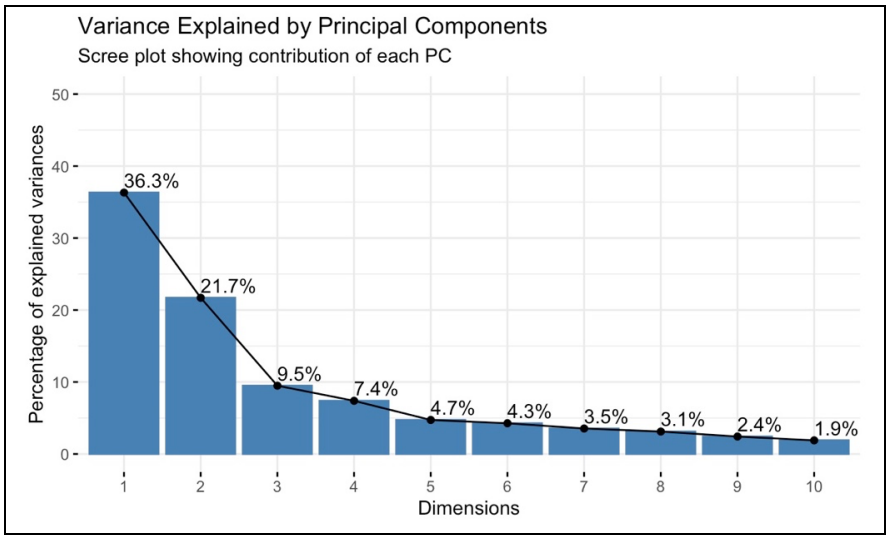

Fig 3.3 – Scree plot to validate the number of PCs extracted

Overall, PCA enabled the identification of performance archetypes within the player cohort, facilitated a deeper understanding of interrelationships between key variables, and provided an interpretable framework upon which clustering analysis and subsequent comparisons with predictive modelling outcomes were built.

3.3.3 Clustering Analysis via K-Means Clustering

Clustering is an unsupervised machine learning technique used to group observations with similar characteristics into distinct clusters, enabling the discovery of underlying patterns within the data. In the context of this dissertation, clustering was applied to analyse cricket players' performance profiles, helping identify player archetypes and uncovering similarities that might influence selection outcomes. R was chosen for this stage of the analysis due to its extensive statistical computing capabilities, superior support for clustering visualisations, and specialised libraries such as *factoextra*, *cluster*, and *ggplot2*, which provide powerful tools for both modelling and interpretation.

Among various clustering algorithms, K-means clustering was selected as the most appropriate approach. Unlike hierarchical clustering or density-based methods such as DBSCAN, K-means is computationally efficient, and produces clearly defined, non-overlapping clusters that are easier to interpret in the context of cricket performance statistics. Furthermore, its integration with PCA allowed for effective two-dimensional visualisation of multi-dimensional data, helping in the interpretation of player groupings and selection patterns. Initially, hierarchical clustering was considered as it does not require a predefined value of k, is slightly better when analysing smaller datasets and produces rich visuals in the form of dendrograms. However, this approach has one major weakness – it struggles with separability when analysing datasets with many overlapping clusters (seen here) and makes no difference to the output (perhaps due to the size of the dataset and how specific and concentrated the objective of the analysis is). As such, due to its simplicity and straightforward implementation, K-Means was the preferred choice of clustering algorithm.

Mathematically, K-means clustering partitions the dataset into k clusters by minimising the within-cluster sum of squares (WSS), defined as:

$$\underset{C}{arg\ min} \sum_{j=1}^{k} \sum_{x_i \in C_j} ||x_i - \mu_j||^2$$

Where:

- x_i represents the ith data point (player's performance vector),
- C_j is the set of points assigned to cluster j,
- μ_j is the centroid of cluster j, calculated as the mean of all points in that cluster.

The algorithm follows an iterative optimisation process consisting of two steps, which are:

1. **Assignment step**: Each observation x_i is assigned to the nearest centroid based on Euclidean distance:

$$C_j = \left\{ x_i : \| x_i - \mu_j \|^2 \leq \| x_i - \mu_l \|^2 \ \ \forall l = 1, \ldots, k \right\}$$

2. **Update step**: After assignment, the centroids are calculated *again* as the mean of all points in each cluster:

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$$

These steps are repeated until convergence, *i.e.*, when the positions of the centroids stabilise or the reduction in WSS falls below a threshold.

To determine the optimal number of clusters, three complementary methods were used - the Elbow method, which identifies the point where adding more clusters produces diminishing returns with respect to variance reduction; the Silhouette method, which measures how well-separated and cohesive the clusters are; and the Gap statistic, which compares clustering performance against random distributions (Tibshirani et al., 2023). Based on these diagnostics, k=4 was chosen as the most suitable value for this dataset. There was some conflict between the output graphs of the three methods; however, as the Elbow curve and the Gap statistic both indicated an optimal k=4, that value of k was selected. The validation graphs for each of these methods can be seen below.

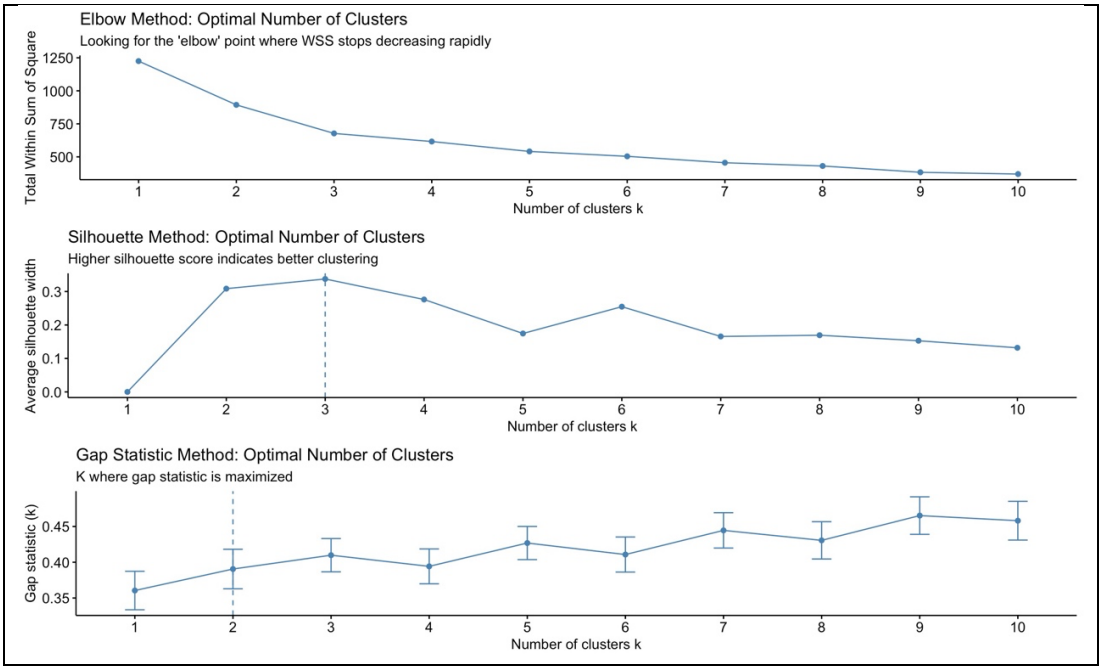

Fig 3.4 – Finding optimal value of k via Elbow, Silhouette and Gap

After clustering, each player was assigned to one of the four groups by the algorithm, and PCA was applied to project the clusters into two-dimensional space for enhanced visualisation. This step made it easier to evaluate the relationship between clusters and player selection status, with the selection outcome encoded as a distinct shape on the PCA scatterplot.

Further analyses were conducted to deepen the interpretation of the identified clusters and understand the distinct performance characteristics within each group. Heatmaps were generated to visualise aggregated statistics, enabling the detection of high-level trends and outliers across clusters. These heatmaps facilitated the comparison of central tendencies, such as batting averages, strike rates, bowling economy rates, and wicket-taking ability, providing an intuitive overview of how player performance metrics varied between clusters. Radar charts were then employed to create multidimensional profiles of representative players from each cluster. This approach enabled the

comparison of performance "archetypes" by displaying strengths and weaknesses across multiple metrics simultaneously, thereby making the contrasts between groups more interpretable.

Representative player profiling involved selecting one or two players per cluster whose individual statistics were nearest to the cluster centroids, effectively serving as exemplars of the overall group profile. These players were utilised as case studies to demonstrate typical performance patterns within each cluster, thereby assisting in contextualising the broader findings in a manner directly relevant to cricket performance analysis.

To explore potential relationships between clustering results and player selection outcomes, selection rates were compared across clusters using both descriptive and inferential statistical techniques. First, cross-tabulations were used to quantify the proportion of players selected versus not selected within each cluster. Subsequently, a chi-square independence test was performed to evaluate whether the observed distribution of selection outcomes was significantly associated with cluster membership. The chi-square statistic was calculated using the formula:

$$\chi^2 = \sum \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

Where O_{ij} represents the observed frequency of players in cluster i with selection outcome j, and E_{ij} denotes the expected frequency, assuming independence. A statistically significant result would indicate that selection outcomes are not randomly distributed across clusters, suggesting that specific performance profiles are more strongly associated with being selected.

This multi-layered analytical approach — combining visualisation, player profiling, and statistical hypothesis testing — provided a comprehensive understanding of how player performance archetypes relate to selection decisions. By integrating these analyses, it became possible to identify clusters containing potentially undervalued players, uncover biases in the selection process, and highlight the cricketing attributes (and, by extension, the skills) most strongly associated with advancement to higher levels of the sport.

Thus, the EDA revealed distinct performance patterns among players, uncovering meaningful relationships between batting, bowling, and fielding metrics through both visual and statistical techniques. Dimensionality reduction using PCA clarified the underlying structure of the dataset, while K-means clustering highlighted three well-defined player archetypes representing batting specialists, bowling-dominant players, and balanced all-rounders. Comparing selection outcomes across these clusters suggested that while performance metrics played a significant role, certain selection decisions may have been influenced by additional subjective factors. These insights establish a robust analytical foundation, providing both the rationale and direction for developing predictive models. This was a crucial step in setting up the next part of this analytics pipeline – the XGBoost model. These outcomes will help enhance the accuracy and interpretability of the XGBoost model outputs.

3.4 XGBoost Model

The core of this entire pipeline is the XGBoost model. This powerful supervised learning algorithm works as a classifier, with the goal of looking at selected vs non-selected players (based on their selection status from trials in 2024) and uncovering hidden trends and patterns to "guess" which combination of performance metrics leads to a successful selection. It is then tasked with using these learned patterns to extract a list of the top 20 players based on their performance during the 2025 season, so that the list may be compared to the top 20 from last year. Having cleaned and inspected

the dataset, this step is essentially the last and most important phase of the cricket analysis pipeline designed here.

Ordinarily, most statistical analysis in sports, particularly cricket, is conducted using basic statistical methods or unsupervised learning methods (chiefly, PCA and Clustering). In many cases, as is traditionally seen, this analysis would be enough. However, given the recent expansion of supervised learning techniques in sabermetric analysis, this pipeline was designed to complement traditional analytical techniques and explore how much more ML techniques can contribute by combining them with basic insights from simple EDA (as seen earlier). There are many powerful ML-driven solutions available, such as LightGBM, Convolutional Neural Networks (CNNs), Long Short-Term Memory Networks (LSTMs), and Transformers (such as BERT, ViT/Swin and Encoder-Decoders). However, these methods rely on very large and highly detailed datasets that also often incorporate longitudinal datapoints such as videographic footage of actual players, weather data and other conditional/contextual information. In the case of the dataset we have at hand, due to its small size, limited scope, simple structure, and lack of any relevant contextual data, transformers and neural networks would result in highly biased and overfitted outputs. As such, the options for supervised learning methods are limited to basic ML techniques – namely, Random Forests (RFs), Support Vector Machines (SVMs), and XGBoost. Of these, XGBoost was selected as the best candidate of the three. This type of model was selected due to its ability to stably handle tabular datasets with multi-column structure, its ability to handle non-linear relationships, and its robustness against overfitting (a huge risk, considering the small size of the dataset) (Zedda, 2024).

At its core, XGBoost (short for Extreme Gradient Boosting) is an ensemble learning method that builds upon the principles of decision trees. A decision tree in isolation is a weak learner – it recursively partitions the dataset according to certain conditional constraints to arrive at predictions, but it is often prone to overfitting and may struggle with generalisation. To overcome this, ensemble methods combine the outputs of many weak learners to construct a more powerful predictive model. There are two main ways to do this – bagging, where models are trained in parallel and their predictions aggregated (as in Random Forests), and boosting, where models are trained sequentially, with each new model attempting to correct the errors of its predecessors. XGBoost belongs to the latter category, refining its predictions step by step through gradient boosting (Chen & Guestrin, 2016).

Gradient boosting works by fitting an initial simple model (usually a single split or "stump") and then iteratively adding new trees that predict the residual errors of the previous iteration. Each added tree is essentially a corrective mechanism, guiding the model closer to the true outcome by minimising a differentiable loss function (Knoll and Natekin, 2016). As this is a classification task, the loss function used here was log loss. What makes XGBoost particularly effective is the way it formalises this process. It uses a second-order Taylor approximation of the loss function, incorporating both first-order gradients and second-order derivatives. This enables it to make more precise updates at each boosting step, compared to traditional gradient boosting which only uses the gradient. Mathematically, the optimisation objective at each stage can be written as a combination of the training loss and a regularisation term:

$$Obj^{(t)} = \sum_{i=1} l\left(y_i, \widehat{y_i}^{(t-1)} + f_t\left(x_i \right) \right) + \Omega\left(f_t \right)$$

Where l is the loss function, f_t is the new tree being added at stage t, and $\Omega(f_t)$ is a regularisation penalty on tree complexity. This penalty term is one of XGBoost's key innovations, as it helps to prevent overfitting by discouraging overly complex trees (Wiens et al, 2025).

Another technical advantage of XGBoost is its use of shrinkage (learning rate), column subsampling, and efficient handling of sparsity. Shrinkage scales the contribution of each tree so that the model converges more cautiously and is less prone to overfitting. Column subsampling, akin to what is seen in Random Forests, introduces diversity in the weak learners and reduces correlation between trees. Furthermore, XGBoost can natively handle missing values in a dataset by learning default directions in the tree structure, a property that is particularly advantageous given the real-world messiness of sports data (Chen and Guestrin, 2016).

In the context of this dissertation, the XGBoost pipeline begins with cleaning and standardising the batting, bowling, and fielding datasets. Then it applies PCA to reduce dimensionality and highlight the underlying structure. Clustering provides initial groupings of players, offering insight into latent categories of performance. However, while such unsupervised methods are descriptive and exploratory, they do not directly address the question of prediction: namely, which players are most likely to be selected based on their performance metrics. This is where XGBoost comes in, bridging the gap between exploratory analysis and predictive modelling.

The model is trained using this year's data of the 50 players who attended the U15 trials last year, 20 of whom were selected into the U15 squad. By encoding the selection outcome as the target variable (1 = selected, 0 = not selected), the XGBoost classifier learns which combinations of batting, bowling, and fielding attributes best differentiate between those chosen and those overlooked. This has already been in the initial data preparation step, when the column "SELECTION" was added to the dataset with the appropriate 1 or 0 value. According to the model design, the model does not merely identify single performance indicators in isolation, but rather the interactions between features. For instance, a player with average batting but exceptional fielding may still emerge as a strong candidate, a nuance that simple linear models would struggle to capture, further validating the need to use XGBoost.

The target variable, "SELECTION", indicates whether a player was selected for the U15 squad. The dataset is split into training and testing subsets using an 80/20 split, with stratification applied to ensure that the proportion of selected and non-selected players is consistent across both sets. This helps maintain balance in the classification task and avoids bias towards the majority class, as much as possible. An ensemble of five XGBoost models is used, each initialised with a different random seed (by varying the *random_state* hyperparameter) to introduce variation in the learning process. These models are trained independently on the same training data and then combined by averaging their predicted probabilities. This ensemble approach reduces variance and improves the overall robustness and stability of the predictions compared to relying on a single model. Before arriving on the solution to use 5 models with varying seeds, two approaches were tested out:

1. An untuned model was run with default hyperparameters. This achieved an accuracy of 80% and an F1 score of 0.65.
2. A model was manually tuned and then optimised using the RandomSearchCV method to find the optimal values for the hyperparameters. This led to a sudden dip in accuracy to 70%, along with a very low F1 score of 0.58. These values remained very similar in the case of manual hyperparameter tuning alongside automated tuning via RandomSearchCV.

The best approach was then decided by combining the merits of both these methods – the most optimal estimates for specific hyperparameters were tuned manually (with default values for the remaining hyperparameters), and an ensemble of 5 models was run with varying seeds.

Hyperparameter	Value	Explanation
max_depth	4	Complexity of the tree – no of layers to which each tree goes. A value that is too low leads to a risk of overfitting.
learning_rate	0.05	Controls the step size during gradient descent, determining how much each tree's contribution is shrunk to prevent overfitting and improve model generalisation. Lower values are better but are slower to train and more computationally expensive.
n_estimators	150	Number of rounds the boosting takes place for
subsample	0.8	Fraction of rows (of data) sampled per tree
colsample_bytree	0.8	Fraction of features sampled per tree
eval_metric	"logloss"	Evaluation Metric used to assess model performance at each stage; "logloss" penalises incorrect classifications more and encourages loss minimisation, leading to more correct classifications.

Fig 3.5 – Table explaining hyperparameter tuning for the XGBoost model

After training, the ensemble produces averaged prediction probabilities for every player in the dataset. These probabilities represent the model's confidence in each player being selected for the U15 squad, based on the patterns it has learned. The average probabilities of each model m, $\widehat{p_m}$ are averaged over the number of models M as

$$\widehat{p_{ens}(x)} = \frac{1}{M}\sum_{m=1}^{M} \widehat{p_m}(x)$$

Where $\widehat{p_{ens}(x)}$ is the average probabilities of ensembles. By adding these probabilities to the dataset and then sorting players based on these probabilities, the top 20 predicted players are extracted and compared against the actual top 20 selected players from the previous year.

This comparison helps validate last year's selection process, seeking to answer two key questions:

1. Were the methods and criteria used to rank last year's triallists accurate, and do they hold true?
2. How much does an ML technique (albeit a simple one) align with a coach's instincts and how much does it enhance/complement traditional statistical methodologies?

This seed ensemble serves as a straightforward method to reduce variability in the model's predictions – even though each part is already a boosted ensemble, changing the seeds alters how features are sampled and how splits are chosen in the trees, resulting in slightly different error patterns. By averaging the results, the quirks and errors of any one model are smoothed over, a steady decision boundary is revealed (when working with limited data, as in this case), and performance metrics like F-scores on the test set are boosted. For the held-out test data, a 0.5 threshold (50% probability of being selected) is applied to the predictions to work out overall accuracy and produce the classification report; yet, for the follow-on task of ranking players, the unadjusted probabilities are the key result, since they offer a more nuanced view than just yes-or-no labels in a selection scenario.

Model interpretability is enhanced using two methods. Firstly, a global feature importance is computed for each of the five models and averaged across the ensemble. This counters the unreliability of importance scores from a single model on small datasets. While measures of tree-based importance (such as gain or weight) are tied to the specific model, this averaged view highlights reliable factors – including batting average, strike rate, bowling economy, and wicket-taking measures – that consistently appear despite random variations in model fits, indicating feature

importance. Secondly, the probability outputs serve as a straightforward way to rank players – assigning a selection likelihood to each one gives a sense of graded certainty, rather than rigid binary calls, making it easier to cross-check against past selection records. Essentially, this selection probability is indeed the main output of the model, serving the model's "opinion" on how much chance a player has of being selected based on learned patterns and hidden trends.

On the practical side, probabilities are generated for the entire group using the same averaged ensemble across all rows of the feature matrix (stored as *preds_full*), followed by the extraction of the top 20 players via sorting by Prediction_Prob. This new data matrix is saved as a CSV file, passing the output dataset cleanly to an easy-to-display tabular file and allowing for its use in dashboards (the final step of the entire analytics pipeline). It is essential to reiterate that PCA and clustering were not part of the XGBoost workflow; they were utilised solely in the EDA phase to understand data patterns and differences, rather than to shape features for supervised modelling. However, their results will aid in the *interpretation* of the XGBoost model's results.

The feature importance values from all five models are also averaged to identify which player performance metrics contribute most to the selection process. Visualising these feature importances provides valuable insights into the underlying factors influencing the model's decisions, allowing coaches and analysts to interpret the model's outputs more effectively. The final step of the XGBoost pipeline, after this feature importance visualisation, is yet another feature importance calculation and visualisation – the SHapley Additive exPlanations, or SHAP – a mathematical framework that assigns values to each feature in the dataset, indicating its importance in influencing model outcomes.

3.5 SHAP (SHapley Additive exPlanations)

SHAP, short for SHapley Additive exPlanations, is a framework designed to enhance the explainability of AI algorithms. By analysing the outputs of a given ML algorithm, the SHAP calculates contribution scores for each attribute in a dataset, thereby revealing which attributes contribute the most and the least to an algorithmic output. SHAP is derived from the concept of Shapley values in cooperative game theory, introduced by Lloyd Shapley in 1951.

Shapley values are a way to fairly divide the total payoff (such as gains or costs) among participants in a cooperative game[2], based on each player's actual contribution. The idea is to calculate what each player brings to every possible combination of factors, average these contributions, and assign the fair share accordingly. This makes the solution robust in situations where players contribute unevenly, and it satisfies fairness axioms: efficiency, symmetry, additivity, and the dummy player property (Narahari, 2012). Mathematically speaking, suppose N is the set of all players, and $\upsilon(S)$ is the value (payoff) the coalition can acquire. The Shapley value for a given "player" i (a participant in the scenario) is given by the formula

$$\varphi_i(\nu) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} \left[\nu(S \cup \{i\}) - \nu(S)\right]$$

Essentially, this formula sums the marginal contributions of a player i over all possible coalitions S where that player i does not have a contribution, thus finding the proportion of that player's contribution (Shapley, 1951). In the context of the assignment at hand, the game is the dataset of

[2] In cooperative game theory, a "game" is simply the term used to describe a scenario affected by multiple quantifiable variables. Anything – from a observation of the stock market, the financial health of a restaurant, to a sports match – can be classified as a "game" and then studied through the lens of game theory.

cricket metrics for 50 U14 and U15 players, while the players are the individual performance metrics contributing to their selection status, and thus, selection probability as per the XGBoost algorithm.

The application of SHAP in this study is primarily motivated by the need for transparency and interpretability within the player selection framework. While XGBoost models are highly effective at capturing complex, non-linear relationships within performance data, their ensemble structure makes them inherently difficult to interpret. In the context of sports analytics, particularly player selection, understanding the reasoning behind predictions is essential. SHAP enables this by identifying the extent to which individual performance metrics influence the model's output. This is especially relevant when comparing algorithmic decision-making with human judgment, as it provides insights into whether the model prioritises similar factors to coaches and selectors or identifies patterns that may otherwise be overlooked. By making the model's reasoning open, SHAP ensures that the outputs are not only accurate but also interpretable and actionable for decision-makers. These outputs can be used by coaches at the various academies in the county to understand player performance behaviours and target specific skillsets to improve players and enhance their chances of selection.

To implement SHAP, a *TreeExplainer* was initialised using the trained XGBoost model, taking advantage of the method's optimised compatibility with tree-based algorithms. SHAP values were then computed for all players in the dataset, quantifying the contribution of each feature to the predicted likelihood of selection. A summary plot was generated to visualise global feature importance, allowing an assessment of which performance metrics most strongly influenced the model's decisions across all players. In addition, individual force plots were produced to provide local explanations for specific predictions, illustrating the relative impact of each feature on a player's selection probability. Together, these visualisations provided a comprehensive understanding of the decision-making process at both the aggregate and player-specific levels, thereby supporting a more transparent, interpretable, and robust evaluation of the model's outputs within the overall player selection framework. A single force plot was generated as an example – for more player specific plots, an interactive Streamlit dashboard was used where radar plots showed individual player performance.

3.6 Streamlit Dashboard

To tie the entire analytics pipeline together a Streamlit dashboard was developed to provide an interactive interface for visualising, analysing, and interpreting player performance data within the Durham County Junior Cricket framework. Streamlit is an open-source Python library designed for building interactive web applications directly from scripts, allowing data scientists to deploy analytical tools without requiring extensive front-end development. By integrating data processing, statistical modelling, and visualisation in a single platform, the dashboard served as a centralised tool for exploring the findings of this study and supporting evidence-based decision-making.

The use of a dashboard was justified by the need to make complex analyses accessible to selectors, coaches, and other stakeholders involved in player identification. Given the variety of statistical methods applied in this dissertation—including principal component analysis (PCA), clustering, and machine learning—it was essential to create a medium where results could be visualised interactively rather than relying solely on static tables and charts. The dashboard enables users to dynamically select players, compare statistics, explore model-driven insights, and understand selection recommendations. By combining interpretability with interactivity, the tool bridges the gap between technical analysis and practical decision-making, ensuring that findings can be applied in a real-world talent identification context. Furthermore, the dashboard was designed to be used as a tool and not specifically tuned to this dataset alone. With the ability to add any data file, the dashboard utilises the core mechanics of the analytics pipeline in a way that can be applied on any merged cricket dataset,

thereby overcoming one of the biggest limitations of the project – the difficulty in obtaining fair, complete and consistent data.

Implementation involved several stages. The dashboard was built using Streamlit, with visualisations created using *plotly* for interactive charts and tables. The sidebar navigation allowed users to switch between multiple analytical sections, including an overview of player performance, PCA-driven dimensionality reduction, K-means clustering for grouping similar players, and XGBoost-based selection modelling. SHAP explainability plots were integrated to provide local and global insights into model predictions, helping stakeholders understand the key drivers behind selection outcomes. The entire code from the previous steps in the pipeline was incorporated into the dashboard to perform these functions. The *Player Comparison* module enabled selectors to evaluate two players side by side across batting, bowling, and fielding metrics. Finally, an *Insights and Recommendations* section summarised the model-driven findings, offering clear guidance for future selection decisions. Overall, the dashboard facilitated transparent, interpretable, and user-friendly exploration of the data and modelling outputs. By consolidating advanced analytical methods into a single accessible platform, it enhances both the usability and practical value of the study's findings.

4. Results

4.1 Introduction

This chapter presents the findings of the analysis as per the methodologies detailed in the previous chapter and discusses their implications. Statistical interpretation is combined with cricketing knowledge to understand what the dataset tells us about selection patterns in the Durham Cricket Junior League. The results are derived from the custom analytics pipeline discussed in the previous chapter, which begins with Exploratory Data Analysis (including a PCA and Clustering component), followed by data modelling and result extraction using an XGBoost framework, culminating in a presentation of results and insights via an interactive Streamlit dashboard. The chapter presents results and a discussion of their interpretations and implications together.

The analysis begins by examining overall performance trends within the dataset, using a myriad of plots and statistical outputs to understand the basic make-up of the dataset. This is followed by PCA to reduce dimensionality and highlight underlying relationships between batting, bowling, and fielding metrics. Clustering is then used to identify potential player groupings based on performance profiles, providing initial insights into patterns that may influence selection outcomes. This is followed by an XGBoost ensemble comprising multiple models with varying seeds. The model ranks the players based on calculated selection probabilities; this is used to create and extract a list of top 20 players for comparison with last year's top 20. The model is evaluated via an accuracy score and a classification report. Feature importance is visualised, followed by a deeper insight into the contributions of key performance indicators via SHAP. All the outcomes of these unsupervised and supervised learning methodologies are presented to coaches and analysts via a Streamlit dashboard with interactive HTML/CSS elements.

Please note that all non-essential outputs (including console output snippets, tables and data files) are available in the appendix

4.2 Exploratory Data Analysis Part 1: Standard Statistical Measures

Having loaded the merged dataset into the system as "cdata.csv", a basic analysis of the structure of the dataset was conducted using the *.shape, .info( ),* and *.describe( )* commands. Additionally, the "SELECTION" column was examined via the *.value_counts( )* command. The outputs revealed that the dataset consisted of 50 players, each with 26 performance metrics. With 20 players selected from the 50, the dataset exhibited a 40% selection rate. Batting, bowling and fielding metrics showed significant variation, indicating diverse skill levels. However, beyond an initial understanding of skill variation, it is also important to note that these outputs also highlight one of the key limitations of the analysis – that of a somewhat unfair dataset for comparison. Due to several factors, not all players have equivalent amounts of data. Some players have more than 50 matches under their belt, while a small number of others have played less than 15 games (the ideal low-end cut-off to consider the performance metrics to be a fair representation of the players' skill levels). This means that raw cumulative metrics, such as Runs and Wickets, see massive variations due to the difference in the sheer volume of games played, while other calculated metrics, such as Strike Rate, Bowl Strike Rate and Economy, see positive/negative inflations. The outputs for these initial analyses can be seen in the Appendix.

Further analysis was done by generating a feature distribution matrix. The matrix provides several key insights regarding the highly skewed nature of the dataset, along with several bimodal/multimodal patterns. These outcomes helped shape the next steps in the analytics pipeline.

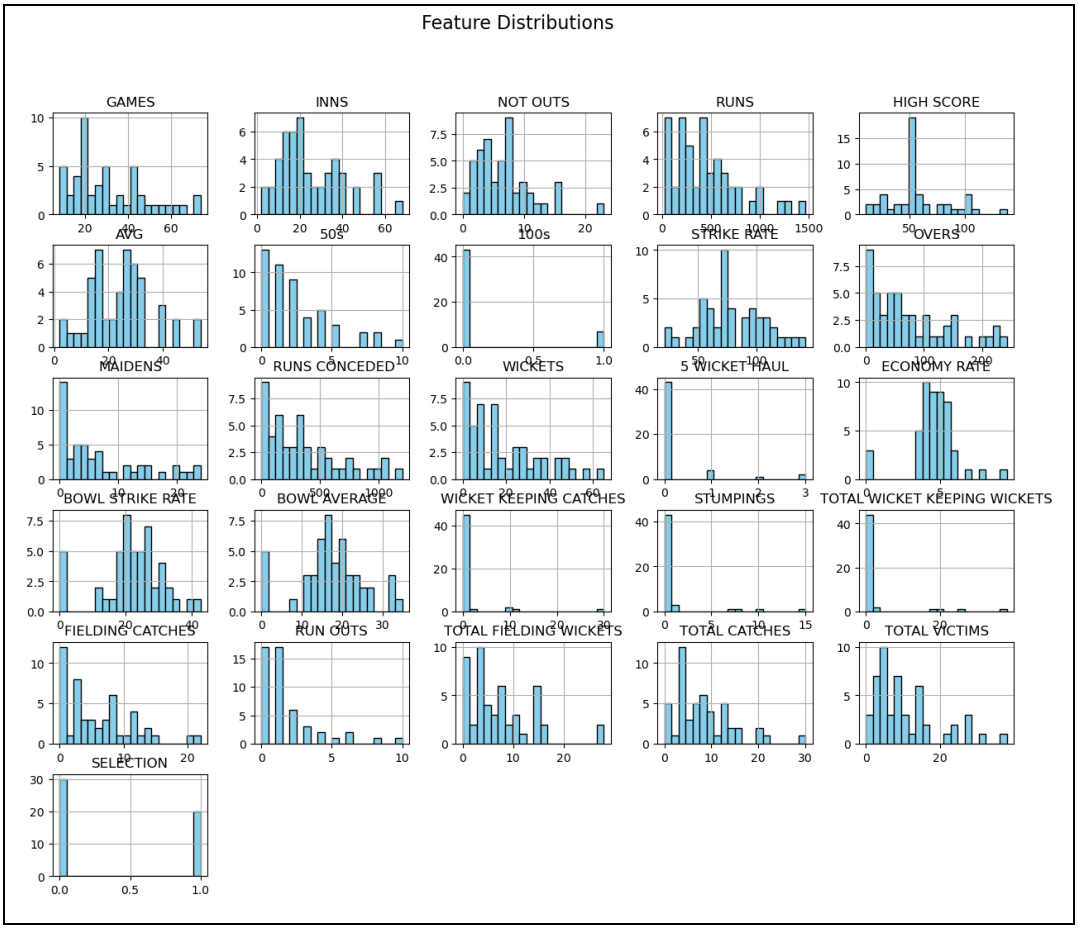

Fig 4.1 – Feature Distribution Matrix for the Merged Dataset "cdata.csv"

The Games, Innings, and Runs metrics exhibit an extreme level of right-skewness, with most players clustered around lower values. This indicates that the majority of players have limited exposure and opportunities, while some players (likely the older and more talented few) have played more matches. These insights track in the context of this study, since it makes perfect logical sense that a small group of players would have more opportunities than the others – with only about 40% players being selected for the higher levels of the game, a small subset of players being given more game time at their local clubs is a natural occurrence. While not every one of the players comprising the top 40% will have significantly more game time than the rest of their cohort, a greater number of players from that small subset will naturally get greater opportunities due to their talent, match-winning ability, and in the interest of their development as cricket players who could potentially go on to represent Durham at the senior levels of county cricket. Indeed, some of these 40% may end up as England Internationals and top picks at the IPL and BBL[3] in the coming years! The wicketkeeping metrics show an extreme right skew, as there are just a few specialist wicketkeepers/players who have taken on keeper duties. 5-wicket hauls (henceforth referred to as 5WI) and 100s are rare, with the overwhelmingly vast majority of players having 0 values in these columns. As such, players who indeed have 5WIs and 100s are standout players – it is likely that the model will be biased in their favour. A specific

[3] The Indian Premier League (IPL) and the Big Bash League (BBL) are seasonal franchise-based cricket tournaments from India and Australia respectively. These leagues are aflush with monetary value and are the pinnacle of the sport at the franchise level, meaning that is a great honour for a player to be picked to play in these leagues.

wicketkeeping stat, Stumpings, is nearly zero for all players except 6 players out of the 50. Of those 7, 4 players have between 8-15 stumpings while the other 3 have 1 stumping each. Batting Averages have a reasonable bell-curve distribution, with a mean of approximately 24.7. This suggests that Batting Averages differentiate players well, irrespective of the difference in the number of matches played. Similarly, Strike Rate shows a normal distribution with a clustering of values around 79.6, indicating consistent measurement across skill levels. Economy Rate is another metric that shows normal distribution and can be a useful tool for measuring bowling efficiency and rating performance. Finally, fielding metrics show bimodal patterns, showing clusters around 0-5 catches and 8-15 catches taken, possibly reflecting a distinction between fielding specialists versus generalists in the field.

There are several key takeaways from this feature distribution matrix. It is seen that calculated metrics are better suited for a comprehensive analysis of player skill over cumulative metrics, as they inherently cancel out imbalances brought forth by uneven match exposure across the cohort. The extreme skewness in several variables suggests that the XGBoost model in the next stage likely benefited from its robustness to non-normal distributions. Traditional linear models would have struggled with these patterns, further justifying its application in this pipeline. The high number of zero-inflated variables (wicket-keeping, bowling specialties) and the bimodal distributions across fielding metrics will likely create distinct principal components and clusters separating specialists from all-rounders. As such, these distributions validate the need for robust data preprocessing via standardisation. The skew in distributions indicates that traditional averages might miss players with limited opportunities/datapoints but strong per-game performance.

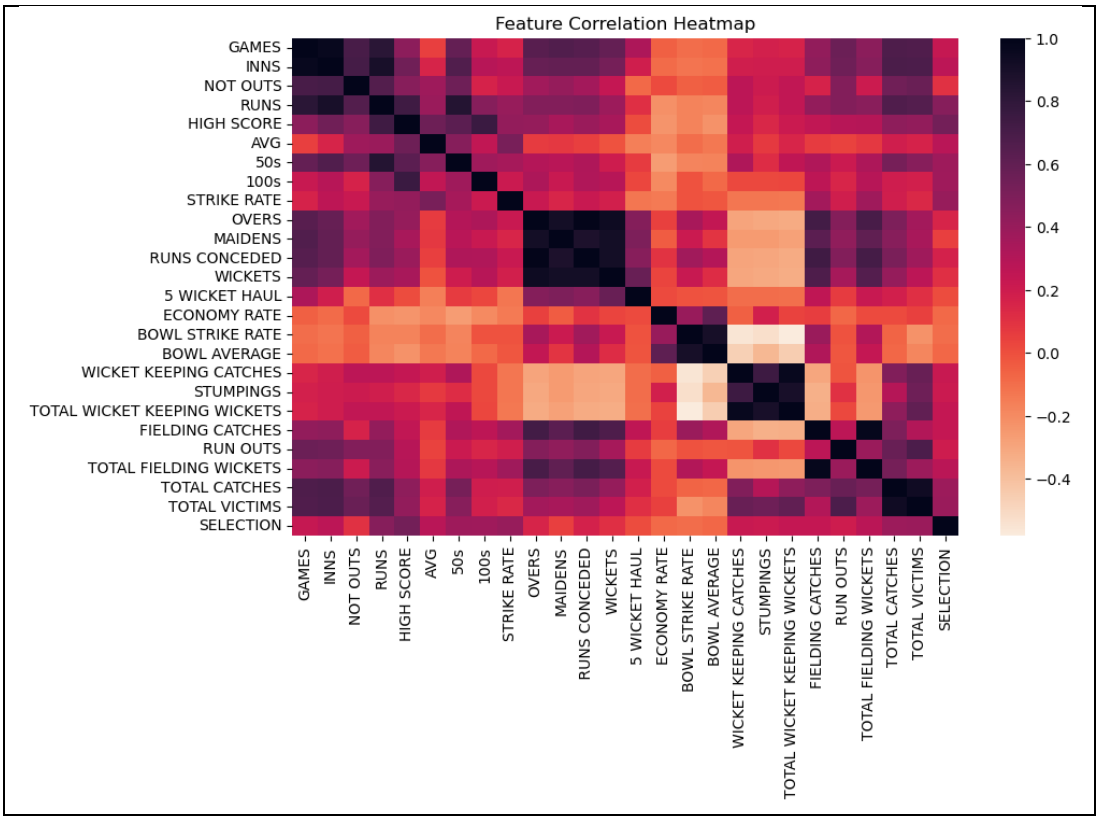

Fig 4.2 – Correlation Heatmap

The correlation heatmap above highlights several notable patterns. Strong positive correlations are observed among batting-related variables, innings, and total runs, forming a distinct cluster indicating batting specialisation. This relationship is expected, as players with more opportunities to bat naturally accumulate higher run totals. Similarly, wicket-keeping metrics such as catches, stumpings, and total

wicket-keeping dismissals cluster tightly, indicating internal consistency as seen in the feature distribution matrix. Fielding-related metrics also demonstrate strong internal alignment, and variables capturing total catches or victims effectively summarise combined contributions from both fielding and wicket-keeping activities.

Conversely, negative correlations are identified in key bowling metrics. Bowling averages display an inverse relationship with total wickets taken, indicating that stronger bowlers typically achieve lower averages through effective wicket-taking. Similarly, economy rates reveal a trade-off between run prevention and wicket-taking ability; bowlers who adopt attacking strategies may concede more runs while simultaneously creating greater wicket opportunities. The true standouts here are the players who understand this trade-off and strike a balance between conservative bowling and a wicket-taking mindset. Importantly, the selection variable exhibits moderate to strong associations with several batting and bowling measures. Higher runs, better strike rates, and increased overs bowled demonstrate positive relationships with selection probability, offering early indications of performance attributes that Durham coaches may prioritise when making decisions.

Pairplot analysis (visualised below) was used to compare selected and non-selected players across the top variance-driving features. For total runs, selected players generally achieve higher values, although some overlap remains, indicating that strong batting performance is advantageous but not the sole determinant. For runs conceded, the relationship is more complex. Some selected players concede more runs, which likely reflects that they bowl more frequently and are entrusted with higher workloads. Similarly, selected players display a wider distribution of overs bowled, reinforcing the idea that selection favours those trusted with greater on-field responsibility. In contrast, strike rate shows clearer differentiation, with selected players clustering within higher ranges (60–120 and beyond), suggesting that efficient scoring is a significant selection driver.

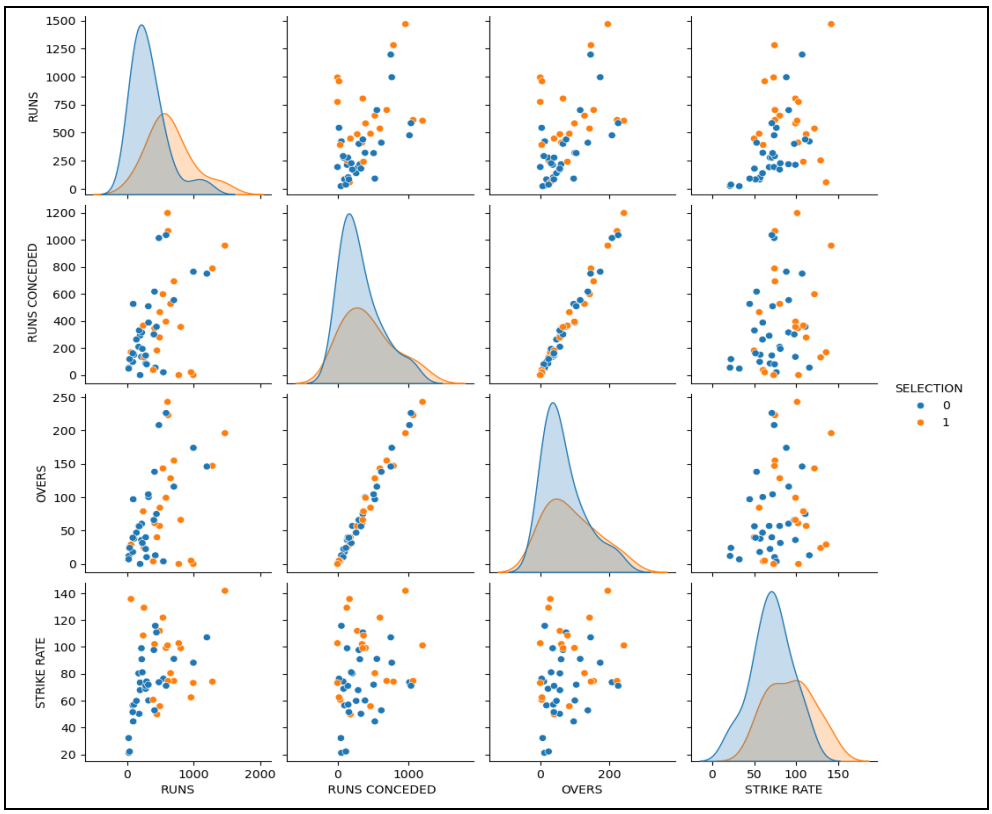

Fig 4.3 – Pairplot analysis visualised

Finally, Variance Decomposition identified the features that most contribute to differences in performance. Runs and Runs Conceded are the main sources of variation, followed by Overs (bowled), Strike Rate (batting), and high score. These results support the multi-dimensional approach to player evaluation.

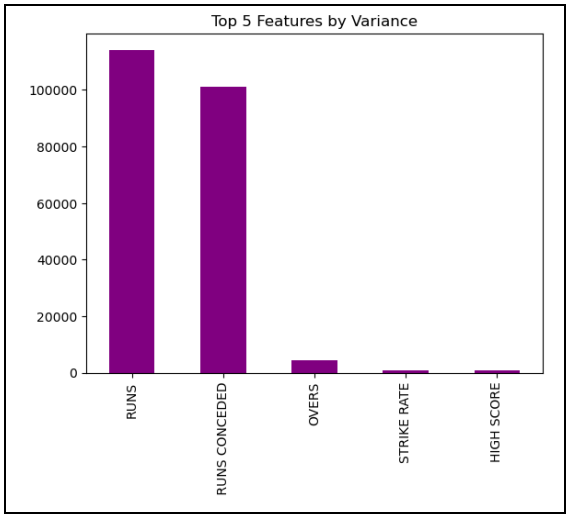

Fig 4.4 – Variance Decomposition

From this initial phase of the EDA, we see that selection decisions are not determined by a single attribute but are made by integrating batting, bowling and fielding performance metrics. There is clear evidence of the trade-off between opportunities and performance, as players with greater workloads (more overs bowled/more innings batted) may still be selected despite conceding more runs or having lower batting averages than their colleagues who did not play as much. The pairplots reveal potential performance thresholds, implying that players must meet minimum competency levels across several metrics rather than relying on exceptional strength in one domain. Finally, the patterns suggest an "all-rounder premium", as selected players tend to contribute across multiple facets rather than excelling in only one. However, this must not be an overhyped insight, as it is natural to see players at the youth stages of cricket performing a range of disciplines. Taking forward the discussion on player development from earlier, it is natural to see selected players, whose development will be a top focus for the county clubs, are made to play across these disciplines in line with youth development paradigms in modern cricket. Further performance structure, such as player specialisations, will be captured via the second phase of EDA – unsupervised learning using PCA and K-Means Clustering.

4.3 Exploratory Data Analysis Part 2: PCA and Clustering

To better understand patterns within player performance and selection dynamics, dimensionality reduction through Principal Component Analysis (PCA) was conducted, followed by clustering analysis to identify potential player archetypes. These complementary techniques provide deeper insights into how underlying performance profiles influence Durham's U15 selection decisions.

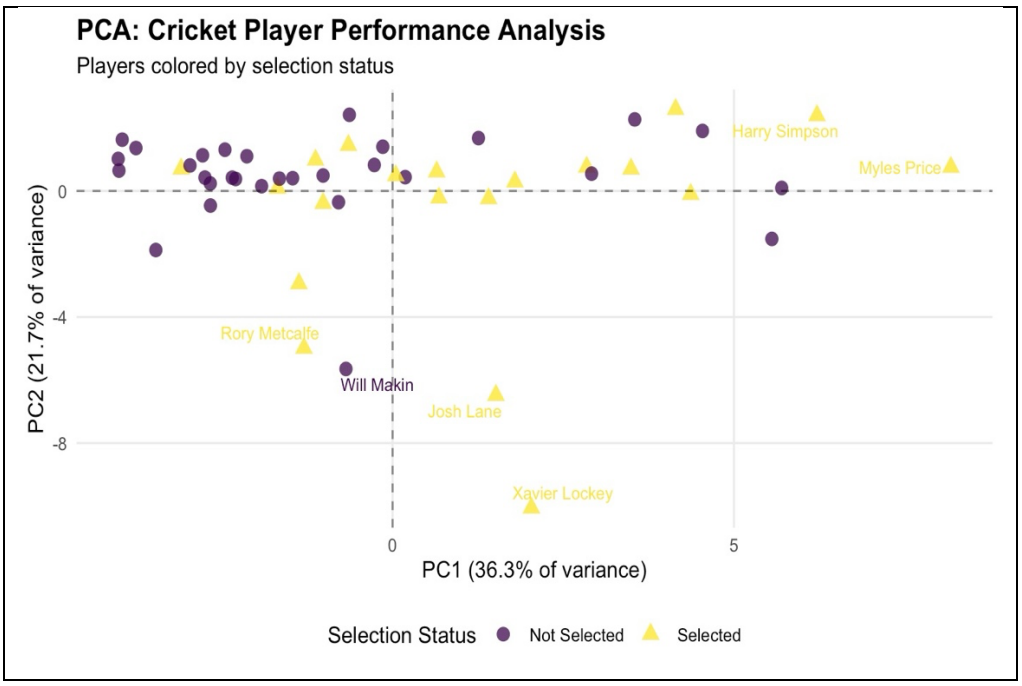

Fig 4.5 – Scatterplot of players in two dimensions as per extracted PCs

The discussion begins with an analysis of the PCA outcomes, as visualised in the form of a scatterplot above, distinguishing between selected and non-selected players across two dimensions. Players at the extreme ends of performance metric analysis are marked out specifically. The PCA results reveal two dominant components capturing 58% of the total variance, reflecting distinct dimensions of cricketing performance. The first component (PC1), explaining 36.3% of variance, represents overall cricket involvement and opportunity. High positive loadings are observed for Innings, Games, Runs, Overs, Maidens, and Runs Conceded, suggesting this dimension reflects exposure to match situations.

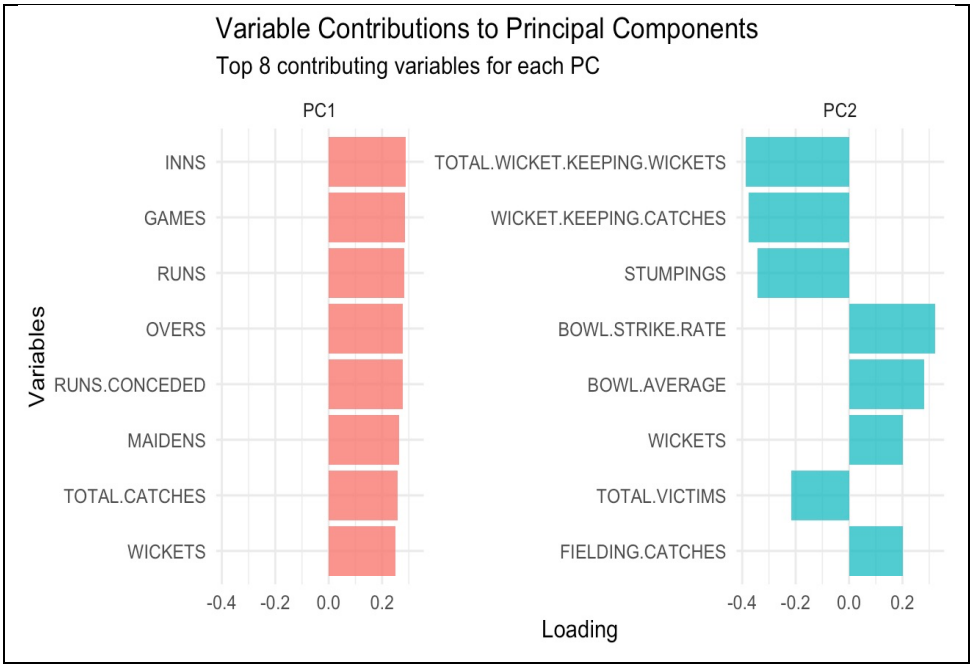

Fig 4.6 – Variable contributions to PC1 and PC2

Selected players exhibit substantially higher PC1 scores (with a mean of 1.36) compared with unselected players (with a mean of -0.9), indicating that players with greater opportunities tend to be favoured for selection – observed during initial EDA as well.

The second component (PC2), explaining 21.7% of variance, differentiates wicket-keeping specialists from other roles. Strong negative loadings are associated with wicket-keeping dismissals, stumpings, and total wicket-keeping contributions, while positive loadings correspond to bowling averages, strike rates, and general fielding metrics. Selected players, on average, have negative PC2 scores (-0.68), indicating that wicketkeepers are disproportionately represented in the selected group, reflecting their specialist value within teams. As such, it is evident that wicketkeepers may not match up to the best batters in the dataset, but stand out and bias selections simply due to their specialist skill set.

These findings imply a two-factor model of selection where decisions are primarily influenced by overall involvement (*i.e.* raw contributions) and player role specialisation. The steep decline in explained variance after the second component supports the idea of focusing analysis on these two dimensions. Notably, the PCA biplot reveals a clear spatial separation between selected and unselected players, validating the idea that the extracted PCs indeed capture performance differences.

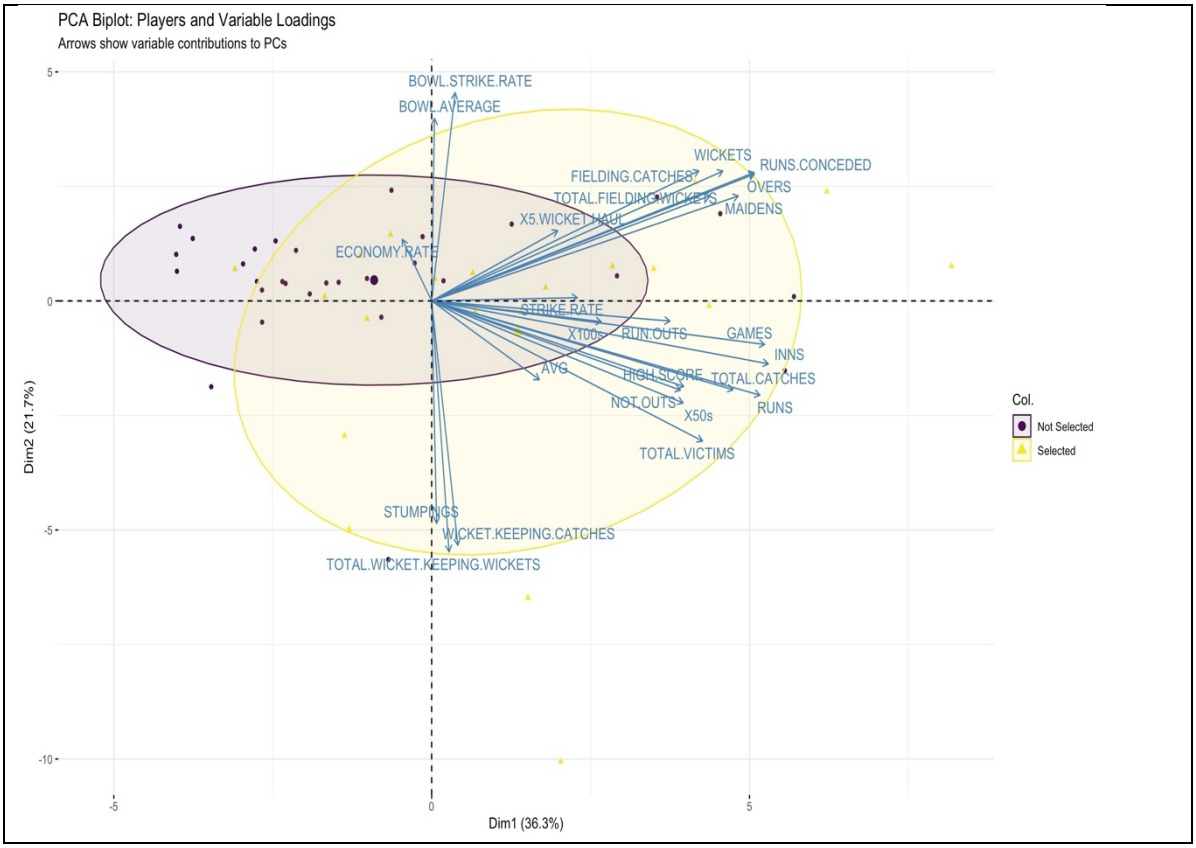

Fig 4.7 – PCA Biplot with feature loadings

Using the principal component scores as input, a k-means clustering approach was applied to group players based on their performance characteristics. Four distinct clusters emerged, representing meaningful differences in player types, though some overlaps remain due to the inherent nature of cricket as a sport in general, and youth cricket in particular, with most players (especially the selected ones) expected to fulfil multiple roles. These clusters are visualised and characterised as described below.

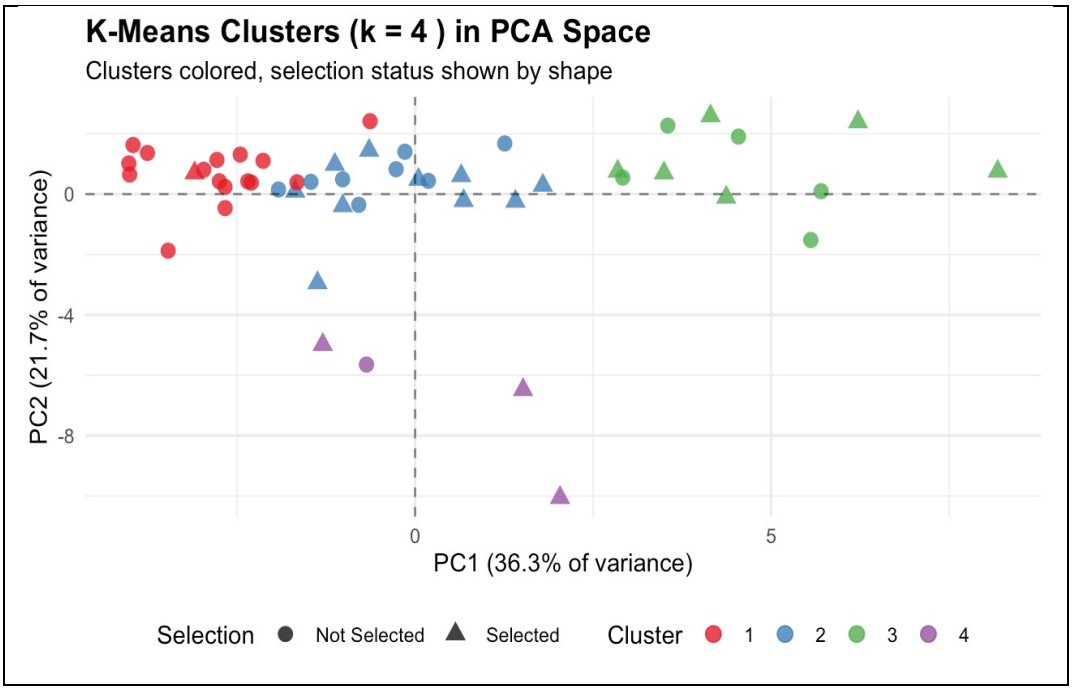

Fig 4.8 – Scatterplot of players in two dimensions separated into clusters (by colour)

Cluster 1: Comprising 17 players, this cluster exhibits the lowest selection rate (5.9%) and consistently underperforms across most metrics, including batting averages, strike rates, and wickets taken. Based on these findings it seems that these players appear to have limited on-field impact, making their low selection rate unsurprising.

Cluster 2: Containing 18 players, this cluster represents balanced all-round performers. With moderate batting averages (31.5), strong strike rates (93.3), and approximately 18 wickets on average, this cluster achieves a 55.6% selection rate. Despite solid contributions across batting and bowling, several players in this group remain unselected, suggesting possible oversights in talent identification. Scouts and coaches should look at the players from this cluster to point out untapped potential.

Cluster 3: Consists of 11 players specialising in bowling. These players average 42.5 wickets – significantly higher than other g roups – but their batting performance remains very average. Their 54.5% selection rate demonstrates that selectors recognise their bowling strengths despite limited contributions with the bat.

Cluster 4: Though the smallest with four players, this cluster has the highest selection rate at 75%. These individuals appear to hold specialist roles or exceptional skills that selectors consider critical, despite contributing minimally to traditional bowling statistics. These can be wicketkeepers, sloggers (high strike rates but low runs and fewer not outs), mystery spinners, etc. Some players in this cluster may even be amongst the youngest of the group, indicating exciting new talent where even a limited number of matches (due to their age and lack of relative cricketing maturity) has yielded promising performances.

The cross-tabulation of clusters and selection outcomes highlights significant differences in selection probabilities. While Cluster 1 contributes the least to selections, Clusters 2 and 3 demonstrate balanced distributions, suggesting selectors value both all-round consistency and specialist bowling roles. This, yet again goes back to the concept of nearly everyone being an all-rounder at the youth level, as

discussed, several times in the prior portions of this analysis Cluster 4, despite its small size, shows a strong positive bias towards selection, implying that some individual player attributes override statistical averages when forming final squads.

Statistical testing further supports these observations. A chi-square test confirms a significant association between cluster membership and selection status, suggesting that the identified groupings reflect meaningful performance archetypes recognised by selectors. However, a silhouette score of 0.17 indicates moderate clustering quality, suggesting some players do not fit cleanly into a single archetype. This is consistent with cricket's performance landscape, where player contributions can be seen as a continuum across disciplines rather than being demarcated as specific and discrete roles. Yet, there remains a degree of appreciation for specialised skills, as discussed earlier, regardless of this preference.

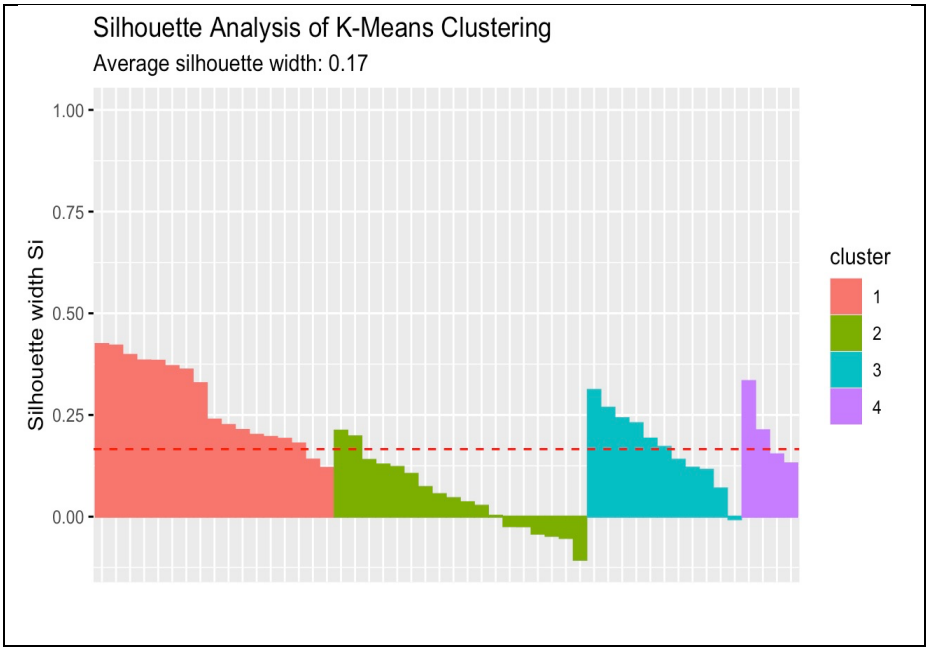

Fig 4.9 – Silhouette analysis of clusters

Visualisation of clusters in PCA space reveals some overlap (see Figs 4.5 and 4.8), particularly between balanced performers (Cluster 2) and specialist bowlers (Cluster 3). This overlap is expected, as some bowlers maintain competitive batting capabilities, blurring the boundary between roles. The diagnostics for optimal cluster number also indicate potential ambiguity, with different methods suggesting between three and six natural groupings. While the choice of four clusters is supported by interpretability and alignment with cricketing context, it does not represent perfectly separated categories.

A particularly valuable outcome of this analysis is the identification of potentially overlooked talent. Players such as Matthew Stokoe, Noah Workman, and Harry Keegan in Cluster 2 demonstrate performance metrics comparable to or exceeding those of selected peers, including high batting averages, strong strike rates, and substantial wicket-taking ability. Their non-selection suggests potential biases or other unquantified factors, such as training attendance, fitness, or tactical preferences, influencing final decisions. More such undervalued players can be seen in the output snippet below.

	Player	Cluster	cluster_selection_rate	AVG	STRIKE.RATE	WICKETS
	<chr>	<fct>	<dbl>	<dbl>	<dbl>	<dbl>
1	Matthew Stokoe	2	0.556	53.5	99.1	8
2	Alfie Armstrong	2	0.556	27.4	90.9	16
3	Noah Workman	2	0.556	25.1	97.8	20
4	Kanishk Sathishkumar	2	0.556	32.5	116.	4
5	Jack Harker	2	0.556	44	111.	17
6	Harry Keegan	2	0.556	16.9	60.2	36
7	Parker Lowe	2	0.556	53.9	72.0	24
8	Will Makin	4	0.75	29.7	76.4	0
9	Muhammad Hayyan	3	0.545	15.9	73.8	50
10	Matty Neal	2	0.556	16.8	52.8	45
11	Huw Morgan	3	0.545	16.2	71.1	65
12	Christopher Bennett	3	0.545	25.5	107.	37
13	Cialan McCarthy	3	0.545	16.3	91.1	33
14	William Shields	3	0.545	31.1	88.3	29

Fig 4.10 – Potentially undervalued players, as per Clustering analysis.

The combination of PCA and clustering provides evidence that selection decisions are influenced by both overall involvement and role specialisation, with selectors favouring players who demonstrate either high exposure or specialist skills, striking a balance between experience and unique talent/roles. The moderate quality of clustering reflects the nuanced nature of cricket, where individual roles and contributions often defy rigid categorisation. These insights will be critical when evaluating the predictive performance of machine learning models, as they highlight the importance of capturing multi-dimensional interactions between variables rather than relying on single metrics, justifying the use of such supervised learning techniques.

4.4 XGBoost Model and SHAP: Model Performance, Results and Influential Metrics

The ensemble XGBoost models developed for predicting player selection status demonstrated a consistent ability to identify patterns within the available performance data and translate them into probabilistic selection outcomes. Across multiple runs with varying *random_state*, the model achieved an overall accuracy of approximately 80% on the test set, thereby correctly predicting eight out of ten cases. While this level of accuracy suggests a strong alignment between the model's learned patterns and the actual selection criteria, the distribution of errors reveals underlying intricacies within the classification task. Specifically, the model displayed perfect precision when predicting player selection. However, this came at the cost of reduced recall for selected players, correctly identifying only half of the players who were ultimately chosen by selectors given the recall score of 0.50. In contrast, for players not selected, the recall was 1.0 (100%), meaning the model was highly confident and accurate in rejecting players who lacked the underlying patterns consistent with selections. These outcomes imply that the model learned to adopt a cautious stance towards recommending players for selection, prioritising precision over sensitivity in its decision-making framework. These results also indicate some degree of overfitting, especially for players labelled 0 under "SELECTION" (*i.e.* non-selected players).

The implications of this become clearer when evaluating the ensemble's behaviour across multiple seeds. While aggregate accuracy remained consistent, the models consistently maintained high probability thresholds for positive predictions. This meant that players were flagged as "selected" only when several performance metrics collectively supported the prediction, resulting in fewer false positives but a greater number of missed opportunities. This behaviour mirrors real-world selection practices to some extent, as selectors often prioritise conservative decision-making to ensure lower errors of inclusion, placing greater weight on consistently demonstrated performance rather than occasional spikes of brilliance. However, the trade-off between precision and recall suggests that

adjusting decision thresholds or using cost-sensitive learning techniques could prove useful, especially if the goal is to identify promising players who may have been previously overlooked rather than simply matching past selection decisions. The decision-making criteria were summarised by a feature importance graph, showing which metrics had the heaviest impact on selection decisions made by the model.

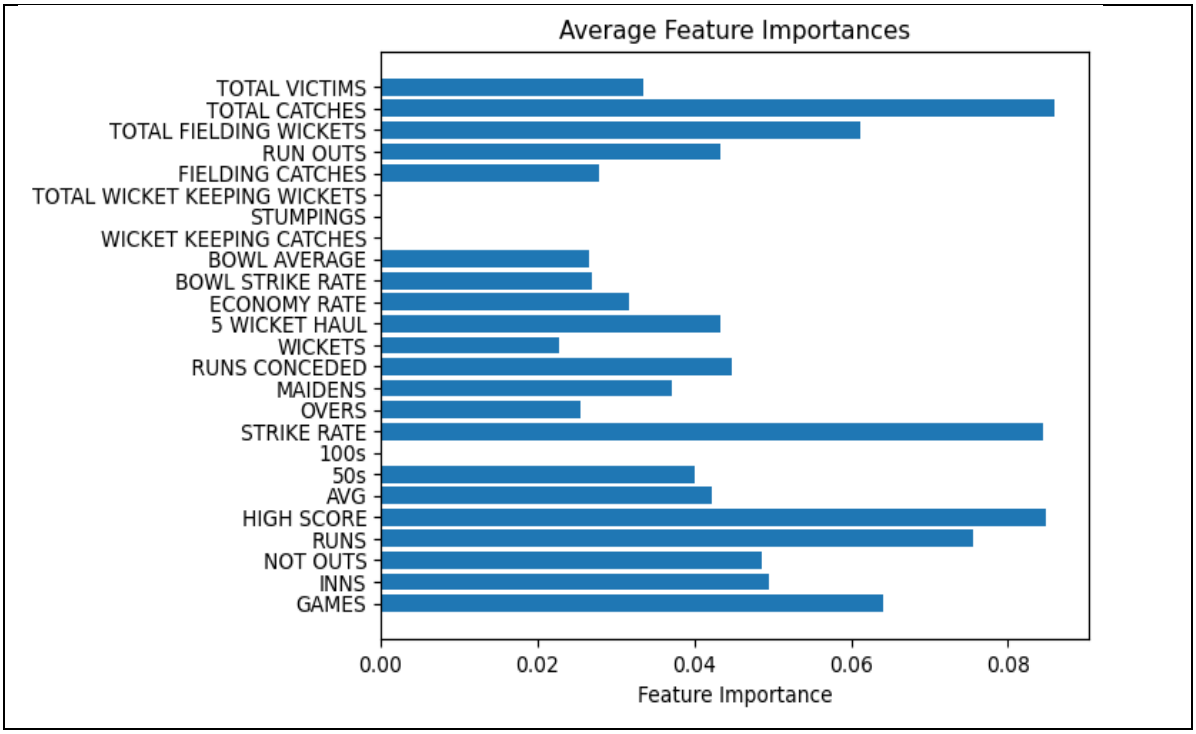

Fig 4.11 – Feature Importance Graph via XGBoost

To better understand the inner mechanics of the model and the drivers of its predictions, SHAP (SHapley Additive exPlanations) analysis was performed across the ensemble outputs. The interpretability offered by SHAP enabled a detailed assessment of feature importance and individual feature effects, providing insights into how specific performance variables affected selection probabilities. Among all predictors, total catches emerged as the most influential feature, with an average SHAP value of 0.085. This suggests that fielding contributions, especially in terms of reliable catching ability, carry significant weight in determining a player's likelihood of selection. An alternative (or perhaps, concurrent) interpretation of this could be that fielding skill is a rare trait – as such, fielding skills add greater value to a player's overall profile than other attributes because of their scarcity. Interestingly, strike rate ranked as the second most important variable (0.082), yet its influence showed an inverse relationship with selection probability. Contrary to what might be expected in the modern game, lower strike rates were linked with higher chances of selection, suggesting a preference among selectors for batsmen who demonstrate consistency and control rather than aggressive shot-making. This pattern hints at an embedded bias within the selection framework favouring calm and composed batsmen capable of building innings methodically, rather than those pursuing flashy, high-risk, high-reward strategies. This is quite surprising, especially within the context of English cricket. Given the current philosophy surrounding the so-called "bazball," one might expect the youth system to promote its core principles of risky, exciting, high-scoring, and aggressive cricket from an early age. Yet, this is not observed. This reaffirms a long-held belief among coaches, analysts and cricket enthusiasts worldwide – ultimately, traditional technique and elegant innings-building still surpass desperate slogging and relentless shot-making. Cricket is, after all, the gentleman's game!

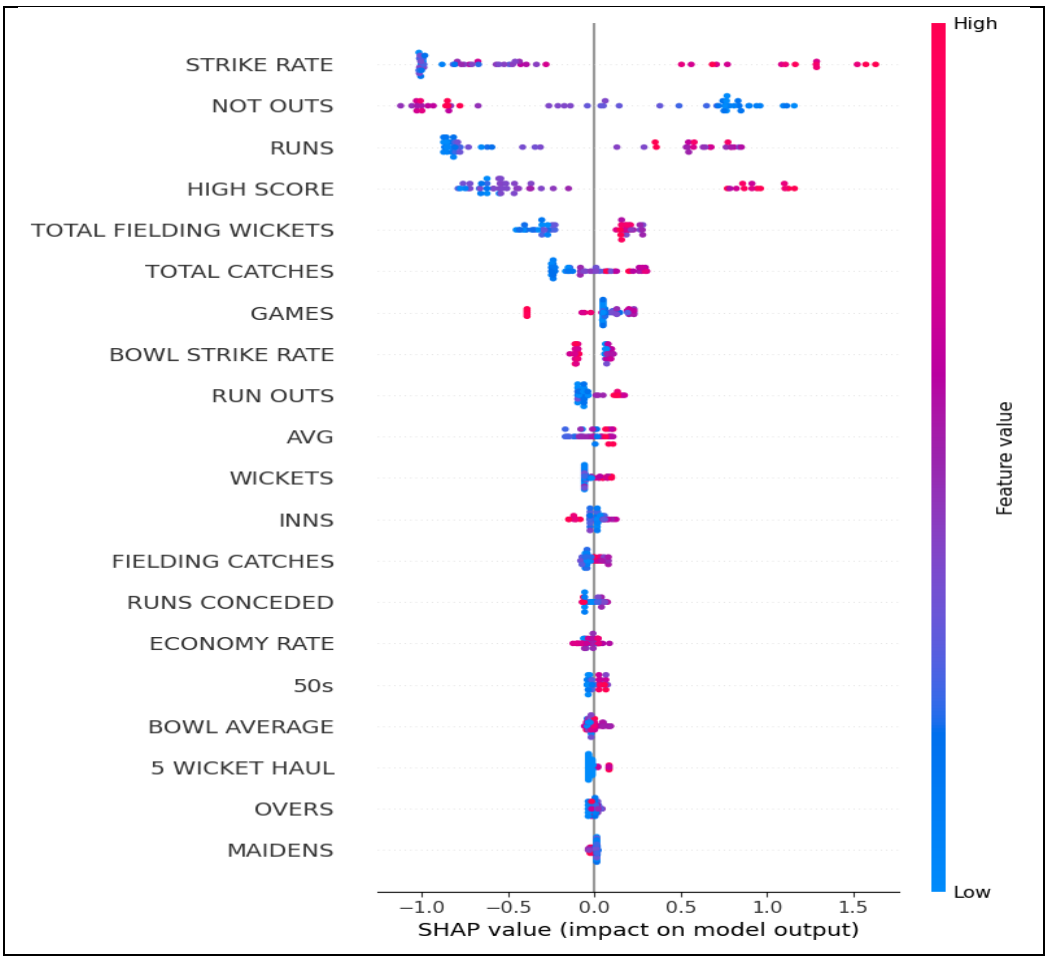

Fig 4.12 – SHAP output visualisation, showing key attributes contributing to the XGBoost model

Alongside these, high score (0.078) and overall runs accumulated (0.074) were also found to play critical roles. Their positive SHAP values reveal that higher batting productivity improves selection probabilities, but their influence remains secondary to fielding strength and batting temperament. Total fielding wickets (0.060) as the fifth-most important feature further reinforces the growing emphasis on all-round capability in the selection process. Players who consistently demonstrate value in multiple facets of the game — especially through fielding contributions — appear to receive a significant advantage in the model's predictions. Taken together, the feature importance rankings indicate that the XGBoost ensemble has effectively captured a multi-dimensional selection philosophy, one where competence across batting and fielding dimensions is weighted more heavily than extreme specialisation. These findings largely reinforce the specialisations and unique skills captured during the PCA and clustering phase of the EDA at the start of the analytics pipeline. The SHAP force plots further reinforce all these findings. For strike rate, the analysis shows a clear pattern where lower strike rates correspond to positive SHAP contributions, while higher strike rates produce increasingly negative impacts on predicted selection probabilities. This aligns closely with the hypothesis that selectors value accumulation and reliability over explosive but inconsistent performances. Conversely, the distribution of SHAP values for total catches and fielding wickets highlights a more linear relationship – greater fielding productivity directly translates into improved selection chances, suggesting that fielding remains a non-negotiable differentiator among otherwise comparable players. Importantly, these SHAP-derived insights were consistent across all ensemble runs, indicating the robustness of these patterns rather than random variance due to specific seeds or in certain subsets of the data. A singular such force plot is presented below.

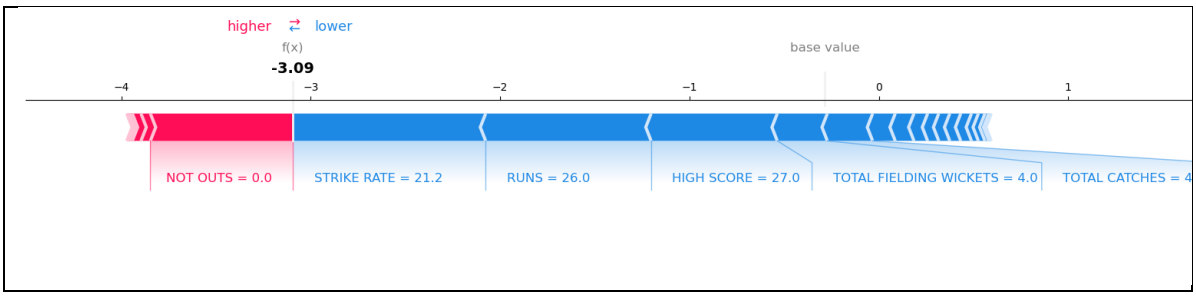

Fig 4.13 – Singular force plot showing variable influence as per SHAP. The output is originally supposed to be interactive.

The integration of these interpretability findings with the dataset's broader context provides deeper insights into the underlying patterns guiding the XGBoost model's outputs. The profiles of players who were consistently predicted as likely selections reveal several shared characteristics: they tend to record solid, reliable batting numbers, maintain strike rates in reasonable ranges (by cricketing standards), and make significant contributions in the field, particularly through catches and direct involvement in dismissals ("TOTAL VICTIMS"). On the other hand, players overlooked by the model often exhibit imbalances in their skillsets, such as strong but inconsistent batting performances, limited fielding impact, or overly aggressive scoring patterns that imply an increased risk of dismissal. This mirrors patterns observed within the actual selection outcomes, suggesting that the model has not simply overfitted to the data but has learned from statistical patterns seen in historical decision-making. It is, nonetheless, important to stress that despite these findings, it is entirely impossible to rule out at least some degree of overfitting given the size of the dataset, the classification based on the target variable and the resulting train/test split; however, these findings do indeed reinforce the performance of the model and increase confidence in its outcomes.

When comparing the model's top twenty predicted players against the actual twenty selected for U15 progression, there is a large overlap that underscores both the strengths and limitations of the approach. The model largely picks the same 20 that were picked in reality by the selectors last year, except for one player. Rory Metcalfe was selected for the team in last year's trials; the model, based on performances from the ongoing season, picks Christopher Bennett in his place. This can be explained in several ways. Firstly, by looking at the raw data and considering the metrics that the model prioritised, Christopher Bennett has performed much better than Rory Metcalfe this season. Bennett has more runs, more wickets (although, as the data shows, Metcalfe was likely a wicketkeeper) and better batting and bowling averages and strike rates. From a human vs AI perspective, selectors likely incorporated qualitative judgments unavailable to the model, such as perceived potential, physical development, leadership qualities, or contextual factors like performance under pressure, technique and raw skill. Also, as mentioned earlier, he was likely selected as a specialist wicketkeeper rather than as a batter/bowler/all-rounder, which can also be seen in the raw due to his exceptional wicketkeeping stats. Even though it was previously observed that the model recognises these specialisations, the significant gap in the other metrics between Bennett and Metcalfe means that the model overrides his specialist role. These qualitative elements remain outside the scope of the current dataset but could be integrated in future iterations to bridge the gap between algorithmic predictions and human judgment – it may be argued that this is a potential weakness of the model. Another explanation for this gap is purely data-related. In the raw data, Bennett has played 61 matches (as of the time of data collection), whereas Rory Metcalfe has only played 36 games. As Metcalfe's role as a keeper has already been established, it is only expected that his batting and bowling figures will not be amongst the best of the best, and his selection relies largely on his wicketkeeping skill, which makes him unique and hence grants him an edge. However, because of the massive gap in game exposure between the two, Bennett, who is already an extraordinary performer when it comes to batting and bowling metrics, significantly overtakes Metcalfe to the point that Metcalfe's keeping stats are

rendered inconsequential compared to the batting and bowling stats. Other than this, the rest of the 19 selections made by the model are entirely the same as the ones made by the coaches and selectors last year, thereby validating both their selections and indeed the model's ability to learn how players are ranked in the real world. However, it must again be noted that the model was trained by targeting the "SELECTION" variable, so there is indeed yet another reason to suspect some degree of overfitting, as established earlier. To mitigate these fears, it is important to look at the results of the PCA and Clustering performed earlier, thereby resulting in a more thorough and holistic view of the data. It is this approach of combining supervised and unsupervised techniques that makes this model unique and more reliable than using either technique in isolation.

From a practical standpoint, the XGBoost ensemble's emphasis on precision over recall means it is particularly effective when the strategic goal is to avoid false positives — ensuring that any player identified as a likely selection has a high probability of meeting selectors' expectations. This is valuable when recommendations must be defensible and trustworthy. However, in purely scouting contexts, where the discovery of untapped talent is more critical, threshold tuning or probabilistic ranking would enable better capture of borderline candidates who may not meet the model's conservative selection standards but otherwise exhibit promising underlying patterns. Indeed, using predicted probabilities rather than hard binary classifications would allow selectors to create a "fairer" shortlist, ranking players based on predicted success rather than hard and fast inclusion/exclusion criteria. However, as the purpose of this exercise was to validate selection decisions from last year, it was essential to take these criteria into account when training the model.

Ultimately, the XGBoost ensemble shows that machine learning can provide a data-driven perspective for assessing player performance and potential, revealing selection criteria that might otherwise remain hidden. By emphasising fielding contributions, balanced and seemingly level-headed batting, and overall reliability, the model reflects the increasing focus in modern cricket on adaptable, multi-skilled players who can finish games and win matches without resorting to flashy and unnecessary tactics. Although its conservative classification tendencies limit recall, its ability to mirror observable selection preferences underscores its validity as a tool rather than a replacement for human judgment. This is supported by the selection of Christopher Bennett over Rory Metcalfe – a coach might have chosen Metcalfe for his general skill and unique role as a keeper; the model, heavily influenced by the available data, failed to make that decision. In an ideal setting, such tools can complement on-field selection decisions to generate a list of players whose skills and talents are demonstrated both by the data and the instincts of a good coach. Larger, more comprehensive datasets with additional contextual metrics will enable the model to consider even more factors and learn more complex patterns, thereby improving its reliability. In fact, these models can serve as stepping stones towards more advanced modelling paradigms such as Neural Nets and Transformers, which can fully utilise the additional data.

4.5 Streamlit Dashboard

The results of all stages in the cricket analytics pipeline were presented via a Streamlit dashboard. These outputs (and indeed their resulting interpretations) were the same as what has been discussed in the prior sections of this results chapter. The dashboard is simply a means of displaying these results in an accessible, dynamic, and interactive environment. A few snapshots of the dashboard are presented below.

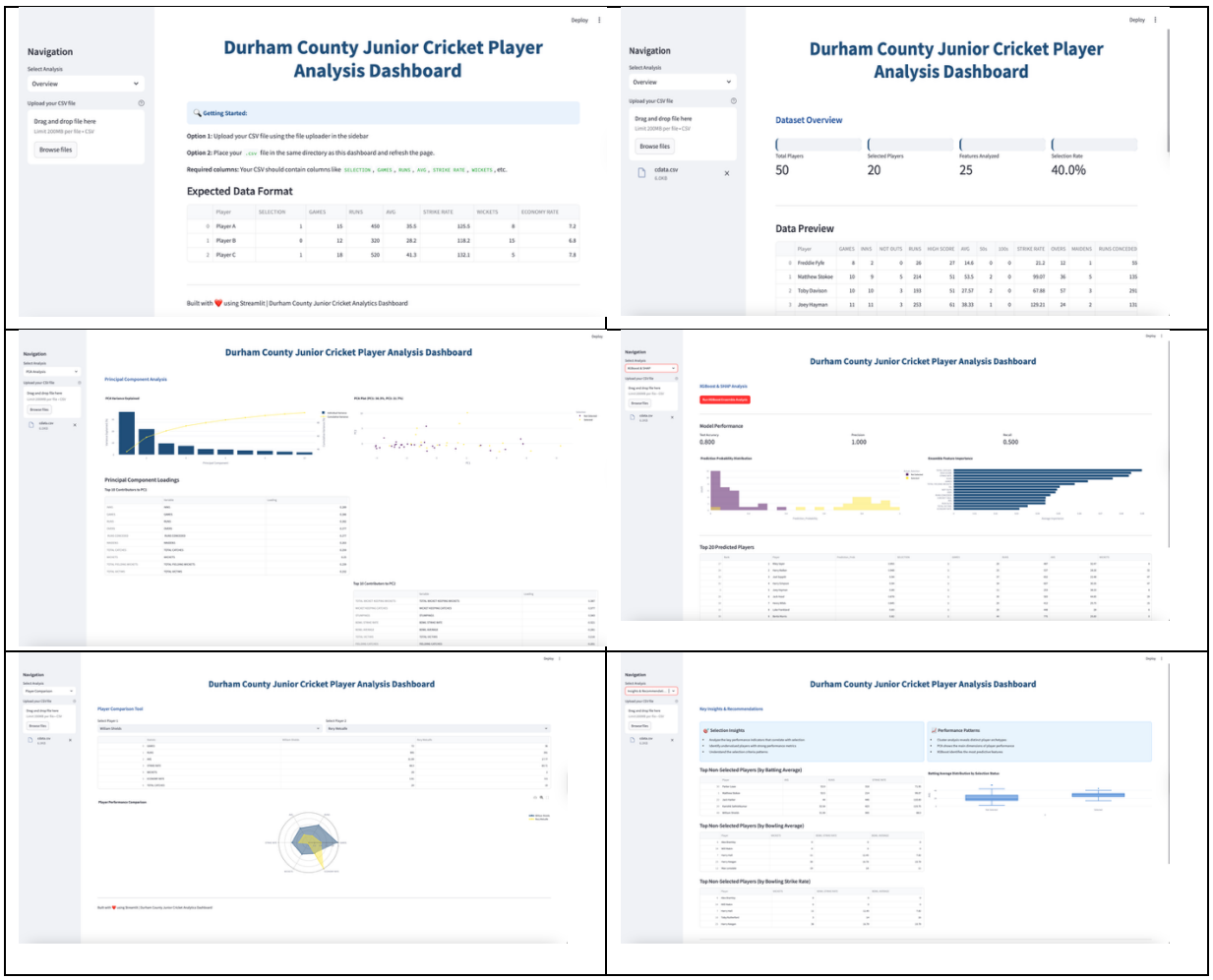

Fig 4.15 – Screenshots of Streamlit Dashboard

5. Conclusion

Talent identification is perhaps the most vital skill in the development of young athletes – cricket is no exception. When done correctly, the best talent identification programmes and the best scouts and coaches have built generational teams and legends of the game, often creating history and redefining the sport as we know it. This skill, however, is highly dependent on the knowledge and experience of coaches and scouts, which can be very hit-or-miss due to inherent biases and inconsistencies. With the growing use of data and machine learning in sports analytics, it is critical to develop analytics pipelines that take advantage of a variety of data processing and supervised/unsupervised learning techniques to augment the judgment made by human coaches and scouts on the field. With this goal in mind, this study aimed to combine traditional data analytics techniques, PCA, Clustering and XGBoost modelling to validate human selections. By running these analyses and models on a dataset of 50 Durham County U15 cricket triallists from 2024, this analytics pipeline tries to draw on their performance data from 2025 to see if the 20 selected players from those 50 are still the best 20 of the lot - confirming whether judgements made by coaches on the ground match up to decisions made by algorithms and statistical analyses.

The analysis begun by utilising standard exploratory data analytics techniques to understand the overall composition and structure of the dataset, including any variable interactions. Having cleaned and prepared the data, these methodologies revealed key aspects of the dataset, including positively and negatively correlated values in the dataset and the skew and distributions of different metrics. Exploratory analysis was then taken forward using PCA and Clustering. These methods revealed four main player archetypes and separated players into groups representing specific talents and skills. These methods also revealed some potentially overlooked players – these players may not have caught the coaches' eyes on-field but shone through based on what the algorithms could glean from the data. The final (and indeed core) stage of the analysis saw an XGBoost model train on this dataset, with the "SELECTION" variable (set at 1 or 0 to demarcate selected and non-selected players) as the target variable. This model achieved an accuracy of 80%, with a passable F1-score of 0.67 and very high precision but relatively mediocre recall, indicating conservative selection criteria. The model largely validated the selections made by the selectors in 2024, with all but one of the same players shining through due to their exceptional performances and unique skillsets. This indicated that the model had successfully matched its selection trends and inclusion criteria with methods employed by the coaches. However, the model was held back by a certain degree of overfitting and bias due to the small size of the dataset and the unfair distribution of metrics, with players selected in last year's trials having more matches than others and hence better raw numbers. The SHAP analysis of the model revealed that key metrics influencing selections included batting strike rates, runs accumulated, bowling economy and total wicket contributions from fielding. Rare achievements, such as 5-wicket hauls, hundreds, and wicketkeeping contributions, also improved players' selection probabilities. All these findings were present in an easy-to-use and highly generalisable Streamlit dashboard to help coaches take full advantage of the pipeline without having to write a single line of code.

The analysis of these results leads to several conclusions. Firstly, it was clearly seen that selections were heavily biased towards players with greater opportunities and exposure, meaning that players with lesser game time and late developers risk being overlooked. It also indicated that the quality of the analysis is only as good as the fairness and size of the dataset, with a larger and fairer dataset leading to potentially better conclusions. Finally, it highlighted the need for a holistic approach – while the pipeline ultimately validated the coaches' selections and summarised key attributes desirable in a player, it also revealed some weaknesses and biases that can only be overcome by on-field decisions fuelled by a coach's instinct. As such, these methods must be used *in conjunction* with real-life selections and human judgment and must never replace human scouts.

The study can be expanded in the future by making several improvements. First and foremost, larger datasets that more fairly represent player performances can be used to improve the trustworthiness of the output. By including more longitudinal data points such as player fitness, match conditions, video footage and other contextual attributes, more complex Deep Learning models can be used to get even better results and more detailed insights. This framework can be scaled and used to build a variety of hybrid pipelines that blend traditional statistical analytics with machine learning paradigms that can potentially be applied at the national level, used by other counties that come under the ECB to inform and validate their youth selections.

As evidenced by the results and conclusions of this study, the pipeline demonstrates the potential of introducing supervised and unsupervised learning into junior cricket talent identification programmes. As data collection improves and methodologies evolve, a balanced approach involving collaboration between selectors and algorithmic tools can significantly enhance fairness, transparency, and long-term player development.

References

(Please note that references are presented in alphabetical order and NOT in order of in-text citation)

1. *(42) The Role of Big Data in Talent Scouting and Athlete Recruitment | LinkedIn*. (n.d.). Retrieved 2 August 2025, from https://www.linkedin.com/pulse/role-big-data-talent-scouting-athlete-recruitment-zeavf/
2. Ajay, L. (2021, June 6). Machine Learning to Cluster Cricket Players. *Towards Data Science*. https://towardsdatascience.com/machine-learning-to-cluster-cricket-players-1d53beeb69b4/
3. Brown, T., McAuley, A. B. T., Khawaja, I., Gough, L. A., & Kelly, A. L. (2023a). *Talent Identification and Development in Male Cricket: A Systematic Review*. *6*(2).
4. Brown, T., McAuley, A. B. T., Khawaja, I., Gough, L. A., & Kelly, A. L. (2023b). *Talent Identification and Development in Male Cricket: A Systematic Review*. *6*(2).
5. Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics*, *15*(1), 27–33. https://doi.org/10.1016/j.aci.2017.09.005
6. Chakraborty, S., Mondal, A., Bhattacharjee, A., Mallick, A., Santra, R., Maity, S., & Dey, L. (2024). Cricket data analytics: Forecasting T20 match winners through machine learning. *International Journal of Knowledge-Based and Intelligent Engineering Systems*, *28*(1), 73–92. https://doi.org/10.3233/KES-230060
7. Chen, T., & Guestrin, C. (2016, March 9). *XGBoost: A Scalable Tree Boosting System*. arXiv.Org. https://doi.org/10.1145/2939672.2939785
8. Chicco, D., Oneto, L., & Tavazzi, E. (2022). Eleven quick tips for data cleaning and feature engineering. *PLOS Computational Biology*, *18*(12), e1010718. https://doi.org/10.1371/journal.pcbi.1010718
9. *CSS: Cascading Style Sheets | MDN*. (2025, July 14). MDN Web Docs. https://developer.mozilla.org/en-US/docs/Web/CSS
10. *Develop—Streamlit Docs*. (n.d.). Retrieved 22 August 2025, from https://docs.streamlit.io/
11. Gudmundsson, J., & Horton, M. (2017). Spatio-Temporal Analysis of Team Sports. *ACM Comput. Surv.*, *50*(2), 22:1-22:34. https://doi.org/10.1145/3054132
12. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition*. (n.d.). O'Reilly Online Learning. Retrieved 5 August 2025, from https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/
13. Jianjun, Q., Isleem, H. F., Almoghayer, W. J. K., & Khishe, M. (2025a). Predictive athlete performance modeling with machine learning and biometric data integration. *Scientific Reports*, *15*(1), 16365. https://doi.org/10.1038/s41598-025-01438-9
14. Jianjun, Q., Isleem, H. F., Almoghayer, W. J. K., & Khishe, M. (2025b). Predictive athlete performance modeling with machine learning and biometric data integration. *Scientific Reports*, *15*(1), 16365. https://doi.org/10.1038/s41598-025-01438-9
15. Jones, B. D., Hardy, L., Lawrence, G., Kuncheva, L. I., Preez, T. D., Brandon, R., Such, P., & Bobat, M. (2019). *The Identification of "Game Changers" in England Cricket's Developmental Pathway for Elite Spin Bowling: A Machine Learning Approach*. *2*(2).
16. K Kiran Babu, Srikanth Banoth, Vijaya Lakshmi Muvvala, Mohammad Shafee, & Shravan Kumar Ainala. (2025). Cricket player performance prediction: A machine learning. *World Journal of Advanced*

Research and Reviews, *25*(2), 953–961. https://doi.org/10.30574/wjarr.2025.25.2.0379
17. Kim, J.-H., Kim, J., Kang, H., & Youn, B.-Y. (2025). Ethical implications of artificial intelligence in sport: A systematic scoping review. *Journal of Sport and Health Science*, *14*, 101047. https://doi.org/10.1016/j.jshs.2025.101047
18. Kuhn, M., & Johnson, K. (2016). *Applied Predictive Modelling* (5th edn). Springer.
19. Larkin, P., & O'Connor, D. (2017a). Talent identification and recruitment in youth soccer: Recruiter's perceptions of the key attributes for player recruitment. *PLOS ONE*, *12*(4), e0175716. https://doi.org/10.1371/journal.pone.0175716
20. Larkin, P., & O'Connor, D. (2017b). Talent identification and recruitment in youth soccer: Recruiter's perceptions of the key attributes for player recruitment. *PLOS ONE*, *12*(4), e0175716. https://doi.org/10.1371/journal.pone.0175716
21. Monsees, L. M. (2025). "There is a lot more potential"—Practitioner perspectives on technology and data-driven talent identification, selection, and development in a German Bundesliga academy. *International Journal of Sports Science & Coaching*, *20*(2), 628–638. https://doi.org/10.1177/17479541241308519
22. *Microsoft/Streamlit_UI_Template*. (2025). [Python]. Microsoft. https://github.com/microsoft/Streamlit_UI_Template (Original work published 2024)
23. name, Y. (n.d.). *Colors & Fonts—Home*. Colors & Fonts. Retrieved 4 September 2025, from https://www.colorsanfonts.com
24. Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*, 21. https://doi.org/10.3389/fnbot.2013.00021
25. Nijenhuis, S. B., Koopmann, T., Mulder, J., Elferink-Gemser, M. T., & Faber, I. R. (2024). Multidimensional and Longitudinal Approaches in Talent Identification and Development in Racket Sports: A Systematic Review. *Sports Medicine - Open*, *10*, 4. https://doi.org/10.1186/s40798-023-00669-2
26. Noorbhai, H. (2022). Cricket coaching and batting in the 21st century through a 4IR lens: A narrative review. *BMJ Open Sport — Exercise Medicine*, *8*(3), e001435. https://doi.org/10.1136/bmjsem-2022-001435
27. Phillips, E., Davids, K., Renshaw, I., & Portus, M. (2010). Expert performance in sport and the dynamics of talent development. *Sports Medicine (Auckland, N.Z.)*, *40*(4), 271–283. https://doi.org/10.2165/11319430-000000000-00000
28. *PlayCricket*. (n.d.). Retrieved 19 May 2025, from https://www.play-cricket.com/
29. *Reshaping Baseball: The Impact of Analytics and Technology | ORMS Today*. (2019, May 20). https://pubsonline.informs.org/do/10.1287/orms.2025.02.07/full/
30. Reynoso-Sanchez, L. F. (2023). Tech-Driven Talent Identification in Sports: Advancements and Implications. *Health Nexus*, *1*(3), Article 3. https://doi.org/10.61838/kman.hn.1.3.11
31. Rossi, A., Pappalardo, L., Cintia, P., Iaia, F. M., Fernàndez, J., & Medina, D. (2018). Effective injury forecasting in soccer with GPS training data and machine learning. *PLOS ONE*, *13*(7), e0201264. https://doi.org/10.1371/journal.pone.0201264
32. *Scikit-learn: Machine learning in Python—Scikit-learn 1.7.1 documentation*. (n.d.). Retrieved 5 August 2025, from https://scikit-learn.org/stable/
33. Shapley, L. S. (1951). *Notes on the N-Person Game — II: The Value of an N-Person Game*. https://policycommons.net/artifacts/4837582/notes-on-the-n-person-game-ii/5674259/

34. *Streamlit*. (n.d.). Retrieved 4 September 2025, from https://cheat-sheet.streamlit.app/?ref=blog.streamlit.io
35. *Style Your Streamlit App with Custom CSS: A Simple Guide*. (n.d.). Retrieved 4 September 2025, from https://pythonandvba.com/blog/style_your_streamlit_app_with_custom_css/
36. Tibshirani, R., Hastie, T., Witten, D., & James, G. (n.d.). *An Introduction to Statistical Learning*. An Introduction to Statistical Learning. Retrieved 25 August 2025, from https://www.statlearning.com
37. Till, K., Cobley, S., Morley, D., O'hara, J., Chapman, C., & Cooke, C. (2016). The influence of age, playing position, anthropometry and fitness on career attainment outcomes in rugby league. *Journal of Sports Sciences*, *34*(13), 1240–1245. https://doi.org/10.1080/02640414.2015.1105380
38. Vaeyens, R., Lenoir, M., Williams, A. M., & Philippaerts, R. M. (2008). Talent identification and development programmes in sport: Current models and future directions. *Sports Medicine (Auckland, N.Z.)*, *38*(9), 703–714. https://doi.org/10.2165/00007256-200838090-00001
39. Wiens, M., Verone-Boyle, A., Henscheid, N., Podichetty, J. T., & Burton, J. (2025). A Tutorial and Use Case Example of the eXtreme Gradient Boosting (XGBoost) Artificial Intelligence Algorithm for Drug Development Applications. *Clinical and Translational Science*, *18*(3), e70172. https://doi.org/10.1111/cts.70172
40. Zedda, S. (2024). Credit scoring: Does XGboost outperform logistic regression?A test on Italian SMEs. *Research in International Business and Finance*, *70*, 102397. https://doi.org/10.1016/j.ribaf.2024.102397
41. Zhou, D., Keogh, J. W. L., Ma, Y., Tong, R. K. Y., Khan, A. R., & Jennings, N. R. (n.d.). Artificial intelligence in sport: A narrative review of applications, challenges and future trends. *Journal of Sports Sciences*, *0*(0), 1–16. https://doi.org/10.1080/02640414.2025.2518694

APPENDIX

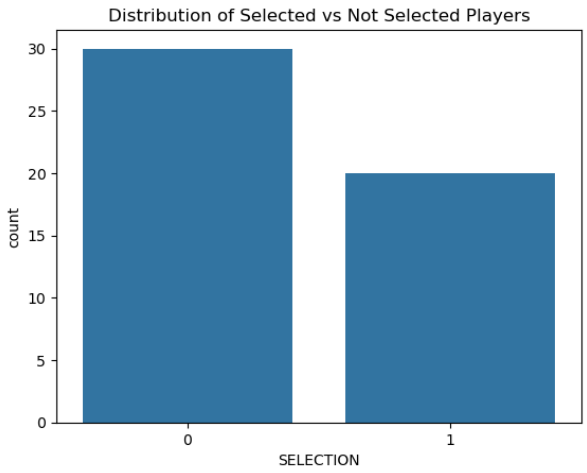

Figure 1 - Distribution of Selected vs Non-Selected Players

```
Top 10 variables contributing to PC1:
> print(pc1_contrib[c("Variable", "PC1")])
                                    Variable       PC1
INNS                                    INNS 0.2889883
GAMES                                  GAMES 0.2857936
RUNS                                    RUNS 0.2817234
OVERS                                  OVERS 0.2769702
RUNS.CONCEDED                  RUNS.CONCEDED 0.2767433
MAIDENS                              MAIDENS 0.2631960
TOTAL.CATCHES                  TOTAL.CATCHES 0.2586438
WICKETS                              WICKETS 0.2499801
TOTAL.FIELDING.WICKETS TOTAL.FIELDING.WICKETS 0.2389660
TOTAL.VICTIMS                  TOTAL.VICTIMS 0.2324790
```

Figure 2 - Variable Contributions for PC1

```
Top 10 variables contributing to PC2:
> print(pc2_contrib[c("Variable", "PC2")])
                                    Variable       PC2
TOTAL.WICKET.KEEPING.WICKETS TOTAL.WICKET.KEEPING.WICKETS -0.3867839
WICKET.KEEPING.CATCHES             WICKET.KEEPING.CATCHES -0.3766009
STUMPINGS                                       STUMPINGS -0.3429279
BOWL.STRIKE.RATE                         BOWL.STRIKE.RATE  0.3209902
BOWL.AVERAGE                                 BOWL.AVERAGE  0.2814082
TOTAL.VICTIMS                               TOTAL.VICTIMS -0.2158381
FIELDING.CATCHES                       FIELDING.CATCHES  0.2005119
WICKETS                                         WICKETS  0.2001754
RUNS.CONCEDED                             RUNS.CONCEDED  0.1978041
OVERS                                           OVERS  0.1954660
```

Figure 3 - Variable Contributions for PC2

	Not Selected	Selected
1	16	1
2	8	10
3	5	6
4	1	3

Figure 4 - Selected vs Non-Selected Players by Cluster

	Cluster	n_players	selected	selection_rate
	<fct>	<int>	<int>	<dbl>
1	1	17	1	5.9
2	2	18	10	55.6
3	3	11	6	54.5
4	4	4	3	75

Figure 5 - Selection Rate by Cluster

	Cluster	Players	Avg_Games	Avg_Batting_Avg	Avg_Strike_Rate	Avg_Wickets	Avg_Economy	Selection_Rate
	<fct>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
1	1	17	17.8	15.9	61.7	8.8	4.8	5.9%
2	2	18	26.7	31.5	93.3	18.4	4.4	55.6%
3	3	11	54.5	25.2	89	42.5	4.6	54.5%
4	4	4	40.8	29.6	68.2	0.5	4.7	75%

Figure 6 - Summary Stats by Cluster

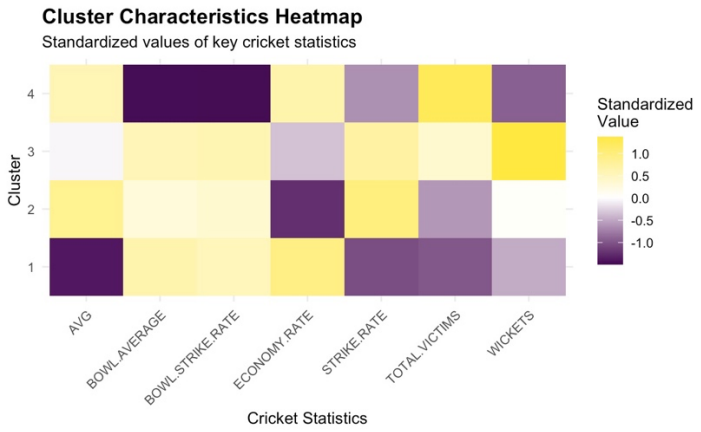

Figure 7 - Heatmap of Cluster Characteristics

All code and relevant data files can be found in the following GitHub Repository - https://github.com/kodiakthebear/durham-junior-cricket-analysis

NOTE: Generative AI was used to enhance the quality of the writing and as a paraphrasing tool during research.