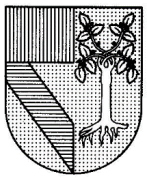


UNIVERSIDAD PANAMERICANA
FACULTAD DE INGENIERÍA
Con estudios incorporados a la
Secretaría de Educación Pública

“MACHINE LEARNING AND FEATURE REPRESENTATION FOR PREDICTING NCAA DIVISION I BASKETBALL TOURNAMENT OUTCOMES”

TESIS

QUE PARA OBTENER EL GRADO DE

MAESTRÍA EN CIENCIAS

P R E S E N T A

ANDRES GREGORI ALTAMIRANO

ASESOR:

DAVID ESCOBAR CASTILLEJOS

COASESOR:

ARI YAIR BARRERA ANIMAS

CDMX 2026

To my mother, Patricia Gregori, whose love made everything possible,
to my wife, Ana Isabel Castilla, whose support carried me through.

Acknowledgments

I would like to express my sincere gratitude to my thesis advisor, Dr. David Escobar Castillejos, for his guidance, patience, and generosity with his time throughout the development of this work. I am equally grateful to my co-advisor, Dr. Ari Yair Barrera Animas, for always lending a ear when in doubt on techical matters. I also want to thank Dr. José Mauricio Pardo Benito for awakening in me the art of research and Dr. Ernesto Moya Albor for his advices during the seminars.

I am thankful to Universidad Panamericana, and in particular to the Facultad de Ingeniería, for providing the academic environment and resources that made this research possible.

I want to thank my in-law family: Rafael Castilla and Mónica González, for always hearing out any new idea on my research.

Finally, I owe a deep personal debt to my two most important women in my life. To my mother, ma, for the constant encouragement and the example of perseverance she has given me throughout my life. To my wife, Anafabel, for her unwavering support, patience, and belief in this project at every stage, this thesis would not exist without them.

Table of Contents

Abstract

Machine learning has become a common tool in sports analytics because it can model nonlinear relations among performance indicators, contextual variables, and competitive outcomes. In basketball, this is especially relevant because modern datasets include box-score statistics, pace-adjusted measures, ratings, seeding information, and other derived indicators. These variables can improve prediction, but they also introduce redundancy and multicollinearity. This thesis examines whether Principal Component Analysis (PCA) can improve or preserve predictive performance in NCAA Division I men's basketball tournament forecasting.

The thesis combines a systematic view of recent machine learning applications in sports prediction with an empirical study focused on March Madness. The related work synthesizes evidence from recent sports forecasting studies and identifies dominant tasks, model families, data sources, and evaluation practices. The empirical component constructs a matchup-level dataset of NCAA tournament games from 2003 to 2025 using regular-season team statistics, tournament seeds, Elo ratings, and a latent team-quality metric. Three classifiers were evaluated: Support Vector Classification, Random Forest, and Extreme Gradient Boosting. Each model was trained on both the original feature space and a PCA-transformed representation, with the 2025 tournament held out for evaluation.

Results show that all six configurations achieved competitive performance, with F1-scores ranging from 81.95% to 84.33%. PCA-SVC obtained the highest F1-score, 84.33%, and outperformed the corresponding SVC model trained on raw features. PCA had only a marginal effect on XGBoost and reduced performance for Random Forest. These findings indicate that dimensionality reduction is not universally beneficial in tournament prediction. Its value depends on the interaction between feature representation and classifier structure. The thesis concludes that PCA can support compact and competitive basketball prediction pipelines, particularly for margin-based classifiers, but that model-specific evaluation remains necessary when interpretability and predictive performance must be balanced.

Chapter 1. Introduction

Background

Sports analytics has moved from descriptive performance reporting toward predictive systems that support scouting, preparation, injury monitoring, and commercial decision-making. Early work in the field focused on organizing match statistics and performance indicators for coaches and managers [1–4]. More recent work uses machine learning to estimate match outcomes, player performance, injury risk, and in-game events from structured statistics, sensor data, tracking systems, and video-derived features [5–7].

Basketball is a strong setting for this type of work because team behavior can be summarized through many measurable quantities: scoring, field-goal attempts, rebounding, turnovers, pace, opponent activity, ratings, and schedule-adjusted indicators. These variables provide useful information, but they are often correlated. A team that scores efficiently may also have a favorable point differential, rating measures, and tournament seed. When several predictors describe overlapping parts of the same underlying strength, model estimation can become less stable and harder to interpret [8–10].

The NCAA Division I Men's Basketball Tournament, commonly known as March Madness, adds another layer of difficulty. Unlike professional playoff series, the tournament is single elimination. On an unusual shooting night, a matchup effect, an injury, or a late-game decision can end a season. This structure makes tournament forecasting difficult even when historical data are available [11, 12]. For this reason, the tournament is a useful case for studying whether feature representation can help machine learning models generalize under uncertainty.

The problem is also relevant beyond the tournament itself. Sports organizations increasingly use predictive models to support time-sensitive, resource-constrained decisions. A model that performs well but is difficult to interpret may be less useful for analysts who need to explain its output. Conversely, a transparent model that ignores important nonlinear structure may not provide enough predictive value. This thesis, therefore, treats prediction as both a computational and methodological problem. The goal is not only to obtain a high score on a single holdout season, but also to understand how a common preprocessing decision changes model behavior.

Motivation

Machine learning research in sports often emphasizes the choice of algorithm. Random Forest, Gradient Boosting, Support Vector Machines, Logistic Regression, and Neural Networks are frequently compared across sports prediction tasks [5, 8]. However, model performance also depends on how the input space is represented. In tabular sports datasets, this issue is practical rather than cosmetic. If many predictors encode similar information, the model may learn redundant patterns, become overly sensitive to the training data, or lose interpretability.

Principal Component Analysis (PCA) offers one way to address this problem. PCA transforms correlated variables into orthogonal components ordered by explained variance [13]. In sports analytics, PCA has been used to summarize player and team profiles and to reveal latent performance dimensions [14]. Yet its value for NCAA tournament prediction remains uncertain because different classifiers respond differently to transformed feature spaces. A margin-based classifier may benefit from orthogonal standardized components, while tree-based ensembles may already handle correlated variables through recursive splitting.

This thesis is motivated by that methodological tension. The main question is not whether PCA is mathematically useful in general, but whether it improves tournament prediction when compared against the same models trained on raw basketball features. The answer matters for predictive sports analytics because preprocessing choices can affect both accuracy and interpretability.

The study also addresses a gap in how sports prediction studies are often reported. Many papers compare algorithms, but fewer isolate the role of feature representation while keeping the dataset, task, and validation design fixed. In a tournament setting with limited observations, this distinction matters. If two models produce similar scores, the more useful model may be the one whose feature representation is easier to maintain, explain, or transfer to a future season.

Problem Statement

NCAA tournament prediction uses data that are informative, limited, and noisy. The number of tournament games is small relative to regular-season data, and each matchup must be represented using information available before the game. At the same time, the available feature set contains correlated measures of team strength, scoring behavior, defensive activity, rebounding, seed, and rating quality. This creates a methodological problem: it is unclear whether models should be trained directly on the original variables or on a reduced representation that removes correlation and concentrates variance.

The problem addressed in this thesis is the evaluation of raw and PCA-transformed feature spaces for predicting NCAA Division I men's basketball tournament outcomes. The study examines whether PCA affects predictive performance across classifiers with different learning mechanisms.

The central research question is therefore stated as follows: how does replacing the original matchup-level basketball features with PCA-transformed components affect the predictive performance of SVC, Random Forest, and XGBoost for NCAA tournament outcomes?

General Objective

The general objective of this thesis is to evaluate the effect of PCA-based feature representation on machine learning models for predicting NCAA Division I men's basketball tournament outcomes.

Particular Objectives

1. To synthesize recent machine learning research in sports prediction and identify the main model families, prediction tasks, and evaluation practices.
2. To construct a matchup-level NCAA tournament dataset using pre-tournament team information from the 2003 to 2025 seasons.
3. To represent tournament matchups using regular-season team statistics, seed information, Elo ratings, and a latent team quality metric.
4. To train and evaluate Support Vector Classification, Random Forest, and XGBoost models using both raw and PCA-transformed feature spaces.
5. To compare model performance using precision, recall, and F1-score on a season-wise holdout design.
6. To discuss the tradeoff between compact feature representation, predictive performance, and interpretability in applied basketball analytics.

Hypothesis

The hypothesis of this thesis is that PCA-transformed features can improve or maintain the predictive performance of machine learning models for NCAA tournament outcome prediction by reducing feature redundancy and multicollinearity. The expected benefit is model-

dependent. A stronger improvement is expected for Support Vector Classification because it is sensitive to scale and correlation structure, whereas smaller or negative effects are expected for tree-based models, as they can handle redundant predictors through split selection.

This hypothesis does not assume that PCA is universally superior. It assumes that the usefulness of PCA depends on the classifier. If the classifier benefits from an orthogonal representation, PCA should improve performance or at least preserve it. If the classifier benefits from the original thresholds and meanings of the basketball variables, PCA may provide little benefit or even reduce performance.

Thesis Structure

Chapter 2 presents the related work. It synthesizes recent machine learning research in sports prediction and positions the empirical study within broader findings on model choice, feature representation, and validation. Chapter 3 describes the methodology, including dataset construction, feature engineering, PCA, model selection, and evaluation protocol. Chapter 4 reports the empirical results. Chapter 5 discusses the findings in relation to the literature and the methodological implications of using PCA in sports prediction. Chapter 6 concludes the thesis and outlines future research directions.

Chapter 2. Related Work

Machine Learning in Sports Prediction

Sports analytics has developed from post-game statistical description into a set of predictive tools used for preparation, recruitment, injury prevention, and tactical analysis. Earlier work in the field emphasized the systematic collection of match statistics and performance indicators, mainly to support coaching and managerial decisions [1–4]. As data collection improved, the analytical focus expanded. Researchers and practitioners began to ask not only what had happened in a match or season, but also what could be inferred about future performance, risk, and competitive outcomes.

Machine learning fits this shift because many sports problems involve nonlinear relationships, interactions among variables, and noisy observations. A team's scoring margin may depend on shooting efficiency, opponent strength, rebounding, turnovers, pace, and situational effects. An athlete's injury risk may depend on workload history, acceleration patterns, recovery, previous injury, and position. Linear models can be useful in these settings, but machine learning methods offer greater flexibility when the relationship between predictors and outcomes is not well captured by a simple parametric form [5, 7, 15].

The systematic review used as background for this thesis identified 118 recent studies, published from 2020 to 2025, that applied machine learning to sports prediction. The review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) statement, screening 978 indexed records from Scopus and Web of Science and applying explicit eligibility, screening, and quality-assessment phases before retaining the final 118 articles [16]. The literature can be organized into four main tasks: match outcome prediction, player performance prediction, injury prediction, and in-game event forecasting. These tasks share the broad goal of prediction, but they differ in their data structures and practical uses. Match outcome studies usually use team statistics, rating systems, match context, or betting-related variables. Player performance studies often use athlete-specific workloads, technical indicators, physiological data, or tracking features. Injury prediction studies rely more heavily on biomechanical, medical, and load variables. In-game event prediction is usually the most data-intensive category because it depends on temporal, positional, or video-derived information.

The review also documented a sustained expansion of the field over the period analyzed, with publication output rising from 12 articles in 2020 to a peak of 31 in 2024 and 27 in 2025. This trajectory is consistent with the broader observation that the proliferation of wearable technologies, digital performance platforms, and tracking systems has expanded the volume and

dimensionality of sports data [17–19]. At the same time, the review reports that the field continues to face interpretive challenges related to integrating multimodal data sources, ensuring data quality, and translating analytical outputs into actionable insights [18, 19]. These observations frame the present thesis as a study performed within a maturing but methodologically fragmented research domain, in which preprocessing and validation choices remain consequential for replicable inference.

Table 1: Main prediction tasks identified in recent sports machine learning literature.

Prediction task	**Reports**	**Proportion**
Match outcome prediction	50	42%
Player performance	34	29%
Injury prediction	25	21%
In-game events	9	8%

Table 1 shows that match outcome prediction represents the largest portion of the reviewed work. This matters for the present thesis because NCAA tournament forecasting belongs to this category. Outcome prediction has a long history in sports analytics because it is easy to define and directly connected to strategic and commercial decisions. However, its apparent simplicity can hide methodological difficulty. A win-loss label is clean, but the game result may be shaped by many interacting factors, including opponent strength, game location, team style, injuries, and randomness.

The same review also shows that the field remains uneven. Some sports and tasks have large, well-structured datasets, while others rely on small, private datasets or manually curated information. According to the disciplinary distribution reported in the review, soccer accounts for 24% of the included studies (n = 28) and basketball for 18% (n = 21), with cricket (12%) ranking third; together, these three sports cover more than half of the analyzed literature. Individual sports and lower-resource disciplines such as athletics, weightlifting, badminton, baseball, hockey, and rugby each contribute one to four studies. The review attributes this imbalance to differences in data accessibility and standardization: team sports benefit from advanced event-capture systems, public databases, and commercial analytics partnerships, whereas individual sports often depend on proprietary sensor systems or manually curated datasets. This imbalance affects what can be generalized from the literature, since a method that performs well in a data-rich team sport may not transfer cleanly to a sport with fewer observations or weaker data standardization.

The review further argues that predictive difficulty across sports cannot be attributed solely to the number of available games per season. Sports with very large schedules do not necessarily produce more accurate forecasts than sports with intermediate schedules, and basketball and soccer, with 60–80 matches per season, frequently exhibit higher and more stable predictive

performance. The reported pattern points to structural features of each discipline, including event randomness, contextual volatility, competitive parity, and the degree to which individual actions accumulate toward outcomes. This observation is directly relevant to NCAA tournament forecasting, where the forecasting challenge is not the absence of historical data but the structural variability that shapes how informative those data are.

Prediction Tasks and Data Sources

Match outcome prediction is the most relevant task for this thesis, but it is useful to place it beside the other predictive problems in sports analytics. Outcome prediction typically uses pre-game or historical information to estimate the probability of a win, loss, score margin, or ranking. These studies are often built from tabular features. Examples include prior performance, offensive and defensive efficiency, ratings, seed, injuries, venue, or schedule strength. The advantage of this structure is that the features are interpretable and often available before the event.

Player performance prediction has a different emphasis. Instead of predicting a team result, these studies estimate individual contribution, future workload, fatigue, or performance level. They may use physical load, movement profiles, tactical role, past performance, or physiological signals. This task is important for training design and talent development, but it can be harder to standardize because player roles vary across teams and sports.

Injury prediction is more closely connected to sports medicine. Models in this area aim to estimate the risk of an injury event based on workload, movement, previous injury, recovery, and related variables. The practical value is high because the prediction may support preventive decisions. The methodological difficulty is also high. Injury events are relatively rare, labels can be inconsistent, and data collection often involves sensitive medical or wearable-sensor information [20, 21].

In-game event forecasting is the newest and most technically demanding of the four categories. These models use time-dependent information, such as player positions, event streams, or video, to predict tactical actions, scoring events, possessions, or other match developments. Such studies can provide fine-grained insight, but they require large labeled datasets and careful temporal validation. For that reason, they are less common than studies based on team-level match statistics.

Across these four prediction families, the systematic review identifies a common pattern: model performance depends less on algorithmic novelty and more on data scale, signal complexity, and task design. Match outcome and player performance studies most frequently rely on tabular match statistics or aggregated event-level summaries, whereas injury and in-

game event prediction increasingly leverage tracking, physiological, or video-derived inputs. The review further notes that publications from 2023 onward more often adopt multi-metric evaluation and reproducible frameworks, indicating progress in benchmarking but persistent inconsistency in validation protocols and data partitioning. These observations support the orientation of the present study, which is positioned within the largest and most data-mature category, match outcome prediction, but uses a comparatively small tournament-level dataset that is therefore especially sensitive to design decisions.

This thesis focuses on outcome prediction with structured pre-game variables. That choice is narrower than the full sports machine learning landscape, but it is appropriate for NCAA tournament forecasting. Before a tournament game, the most reliable and reproducible inputs are regular-season summaries, ratings, seeds, and team-level indicators. The challenge is not the absence of data, but how to represent overlapping team-strength information in a way that supports prediction.

Models Commonly Used in Sports Analytics

Model selection in sports analytics is shaped by the type of data available. Tabular match and team statistics often favor methods that work well with mixed predictors, limited samples, and nonlinear interactions. Tracking and video data make deep learning more attractive, but those methods usually require more observations, more computation, and stronger validation procedures. As a result, traditional machine learning methods remain common even as deep learning becomes more visible in the literature.

Tree-based ensembles are widely used because they handle nonlinear effects and interactions without requiring a manually specified functional form. Random Forest reduces variance by combining many decision trees, each trained with bootstrap sampling and randomized feature selection [22]. Gradient Boosting and XGBoost use a different logic: they add trees sequentially, with later trees correcting errors made by earlier ones [23]. These methods are useful in sports prediction because performance indicators often interact. For example, a high scoring average may mean something different depending on pace, opponent strength, and defensive profile.

Support Vector Machines and Logistic Regression remain important baselines. Logistic Regression is transparent and useful when the goal is to estimate a probability from a modest number of predictors. Support Vector Classification is less directly interpretable, but it can perform well when classes are separable in a transformed or standardized feature space. SVC is especially relevant to this thesis because its margin-based objective can be affected by predictor scale, correlation, and feature geometry [24].

Neural Networks are increasingly common in studies that use sensor, tracking, or video data [5, 6]. Their flexibility allows them to model high-dimensional temporal patterns, but it also raises practical concerns. Without large datasets and careful validation, deep models can overfit. Their outputs may also be harder to explain to coaches, analysts, or medical staff. For structured team-level tournament prediction, the literature does not imply that a neural model is automatically preferable to well-tuned classical classifiers.

Table 2: Frequently used model families in recent sports prediction studies.

Model family	**Reports**
Random Forest	45
Gradient Boosting and XGBoost	36
Support Vector Machines	33
Neural Networks	32
Logistic Regression	31
Decision Trees	19
Naive Bayes	15
k-Nearest Neighbors	13

Table 2 summarizes the most frequent model families in the reviewed studies. The systematic review further reports that ensemble-based methods alone account for more than half of all implementations across the 118 included studies, and that the dominance of these techniques is also visible within the match outcome subcategory, where Random Forest appears in 20 studies, Gradient Boosting and XGBoost in 17, and Neural Networks in 16. The review describes this pattern as evidence that ensemble and boosting techniques perform consistently on tasks involving tabular or event-level datasets, owing to their resilience to overfitting and minimal preprocessing requirements, while neural and hybrid architectures are increasingly adopted in data-rich contexts that include video, positional, or sequence information. Logistic Regression and Support Vector Machines are described in the review as enduring baselines that offer interpretability and benchmarking stability across diverse data contexts.

The empirical study in this thesis uses SVC, Random Forest, and XGBoost because they represent different ways of learning from the same feature set. SVC emphasizes the separating boundary. Random Forest emphasizes aggregated tree decisions. XGBoost emphasizes sequential error correction and regularized boosting. This selection is also aligned with the conference-paper version of the present empirical study, which justifies the same three classifiers on the grounds that they correspond to distinct modeling assumptions: bagged tree ensembles, boosted tree ensembles, and margin-based classifiers [5]. Comparing these models across raw and PCA-transformed feature spaces enables the study of whether feature representation interacts with classifier design.

Basketball Outcome Prediction

Basketball is a natural domain for predictive modeling because much of team performance is captured in structured statistics. Traditional box-score variables capture scoring, shooting attempts, rebounds, assists, turnovers, blocks, steals, and fouls. Advanced metrics add pace adjustment, offensive and defensive efficiency, possession-based rates, rating systems, and shot-quality information. Tracking systems can add player movement, spacing, lineup behavior, and shot-location context [25, 26].

Outcome prediction studies in basketball often rely on aggregated team statistics because they are available across many seasons and competitions. Prior work has used shooting efficiency, rebounding, turnovers, schedule strength, tempo, and defensive indicators to estimate game results [8, 9, 27]. These features can describe meaningful differences between teams. A team that protects the ball, rebounds well, and limits opponent shots may have a higher probability of winning. However, many of these variables are not independent. Strong teams tend to perform well across multiple metrics simultaneously, creating redundancy in the predictor space.

The conference paper that anticipates the present empirical study explicitly formulates this issue. It argues that contemporary basketball datasets combine traditional box-score measures, advanced efficiency metrics, and contextual information on pace, lineups, and shot selection, and that the resulting feature spaces are both richer and more redundant than those used in earlier outcome studies [25, 26]. High correlation among predictors in this setting can increase redundancy, complicate model estimation, and reduce interpretability, with the additional risk of overfitting and unstable generalization when the number of observations is limited relative to the number of features [8–10]. This characterization is consistent with the systematic review, which reports that the growing adoption of advanced metrics and player-tracking data has increased the dimensionality of the input space without uniformly improving the conditions under which models are trained and validated.

The NCAA tournament adds specific forecasting challenges. The single-elimination format means that one game determines advancement. This differs from professional playoff series, where repeated games can reduce the effect of random variation. In March Madness, a poor shooting night, a favorable matchup, foul trouble, or late-game variance can overturn a prediction. Prior studies have therefore examined the predictive value of seed, rating systems, and historical tournament performance [11, 12]. These variables are useful because they summarize team quality, but they cannot remove the uncertainty created by the tournament structure.

Another issue is sample size. Regular-season college basketball features many games, but the tournament itself offers far fewer observations. If the model is trained only on tournament

games, the amount of direct target data is limited. If regular-season information is used, it must be transformed into pre-tournament team summaries. This creates a design problem: the model depends heavily on how team strength is encoded prior to the tournament. That is why this thesis places feature representation at the center of the analysis.

March Madness and Tournament Forecasting

March Madness prediction is not only a basketball problem. It is also a methodological test for supervised learning under uncertainty. A tournament model must be trained on past seasons, applied to future matchups, and evaluated on games that can be decided by small margins. The model cannot use post-game information, later tournament performance, or features that were not known before the matchup. This temporal ordering is essential for honest evaluation.

Seeds are among the most visible tournament features. They summarize the selection committee's judgment and regular-season performance into a compact ordinal variable. However, seed alone does not fully describe matchup quality. Two teams with the same seed in different regions or seasons may not be equally strong. Rating systems such as Elo provide another way to summarize strength by updating team ratings based on prior results. Team quality estimates based on scoring margins provide an additional measure of underlying strength. Each of these variables is useful, but they may measure the same concept only in part.

This overlap motivates the use of dimensionality reduction. If seed, Elo, point differential, and quality estimates all reflect team strength, then a model trained on raw variables may receive the same signal several times. Some classifiers can tolerate this redundancy, while others may become sensitive to it. The tournament setting, therefore, provides a practical case for comparing original variables against PCA components.

A complementary observation is offered in the conference paper introducing the empirical pipeline analyzed in this thesis. It argues that the combination of rich historical data and substantial outcome variability has made March Madness a widely studied setting for sports prediction research, and that even modest improvements in prediction quality can affect resource allocation in scouting, roster construction, opponent preparation, and risk evaluation [11, 12]. The same paper notes that despite the widespread use of advanced basketball metrics, the effect of replacing raw team statistics with PCA-based representations for NCAA tournament prediction has remained insufficiently examined in a comparative setting. The present thesis is therefore positioned to address an explicitly identified gap rather than an inferred one.

Feature Representation and Dimensionality Reduction

Feature representation determines what information is available to the model and in what form. In sports analytics, this choice is rarely neutral. A game can be represented with raw team averages, differences between teams, ratings, recent form, player-level summaries, or latent components. Each representation carries assumptions about what matters. For tournament prediction, difference variables such as seed difference or Elo difference are particularly useful because the outcome depends on the relative strength of the two teams.

High correlation among predictors is common in basketball. A strong team may have a good seed, a high Elo rating, a strong point differential, an efficient offense, and favorable defensive metrics. These variables are not identical, but they can contain overlapping information. In some models, this redundancy may increase variance or obscure interpretation. In others, it may not create a serious problem. The effect depends on the learning algorithm and the sample size.

Dimensionality reduction replaces a larger feature set with a smaller representation. The goal is not merely to reduce the number of columns. It is to preserve the main information in the data while reducing redundancy, noise, or instability. In a thesis focused on predictive modeling, this matters because a compact representation can improve generalization by removing irrelevant variation. At the same time, it can harm performance by removing useful structure or making relationships harder for the model to learn.

PCA is one of the most established unsupervised dimensionality-reduction techniques. It transforms standardized predictors into orthogonal components ordered by explained variance [13]. The first components capture the largest directions of variation in the feature matrix. Later components capture smaller directions. In sports analytics, PCA has been used to summarize player profiles, team styles, and latent performance dimensions [14]. This makes it a reasonable candidate for basketball data, in which many indicators capture related aspects of team strength.

PCA also has limitations. Because it is unsupervised, it does not select components based on their predictive value for winning. It preserves variance in the predictors, not necessarily information most relevant to the outcome. A component that explains a large share of feature variance may or may not help classify tournament winners. Another limitation is interpretability. A raw variable, such as seed difference, is easy to explain. A principal component may combine seed, Elo, scoring margin, and other statistics with positive and negative weights. This makes the model less transparent for basketball analysts.

The literature, therefore, leaves an applied question open. PCA may improve the feature space for models that are sensitive to correlation and scale, but it may not help models that

already handle redundant variables. This thesis addresses that question by holding the dataset, task, split, and classifiers constant while changing the feature representation.

The conference paper associated with this thesis formalizes this reasoning by treating feature representation as a central design choice rather than a purely preprocessing step. In its account, the relevance of PCA in basketball arises from two simultaneous properties of the available data: many predictors describe related aspects of team performance, including offensive and defensive efficiency, rebounding, shooting, turnover rates, and schedule-adjusted indicators, and the dimensionality of the feature space has grown alongside the adoption of advanced metrics and tracking-derived variables [14, 25, 26]. Within this framing, the comparison between raw and PCA-transformed inputs is interpreted as a test of whether models with different inductive biases respond differently to the same compressed signal. The systematic review supports the same general logic by reporting that algorithmic choice in sports analytics is shaped more by data scale and structure than by the sport, and that the dominance of ensemble and boosting methods on tabular data stems precisely from their tolerance for limited preprocessing.

Validation and Evaluation Practices

Evaluation practice remains one of the main weaknesses in sports prediction research. Studies often report accuracy, precision, recall, F1 Score, AUC, RMSE, MAE, or related metrics, depending on the task. These metrics are useful, but they do not solve the broader validation problem. If the train-test split does not match the intended deployment setting, the reported performance can be optimistic.

Random splits are common in machine learning, but they can be problematic in sports data. Games from the same season, league, or team context may share information. If observations are randomly mixed across train and test sets, the model may benefit from temporal or contextual leakage. Season-wise validation is often more appropriate for forecasting because it asks the model to learn from the past and predict the future. This is especially important in NCAA tournament prediction, where the real use case is to train before a tournament and then forecast upcoming games.

Metric selection also matters. Accuracy can be misleading when classes are imbalanced or when different error types have different consequences. Precision and recall separate false-positive and false-negative behavior. F1-score combines both and is useful when a single comparison measure is needed for classification. Probabilistic forecasting would require additional metrics, such as log loss or Brier score, but the present thesis focuses on binary classification performance.

Recent methodological critiques argue that sports data require validation practices adapted to their domain structure [28, 29]. A model trained on one league, season, or competition may not generalize to another. Features that are stable in one sport may be unstable in another. Reporting a single metric from a single split is therefore not enough to establish broad model reliability. This thesis does not solve all of these issues, but it follows a season-wise holdout design to keep the empirical comparison aligned with the forecasting task.

The systematic review reaches a comparable conclusion by synthesizing validation practices from 118 studies. It reports that current work often uses custom train-test splits, different performance metrics, and non-comparable experimental settings, which complicate cross-study synthesis, and it recommends that future studies report data-partitioning strategies in detail, employ repeated or nested cross-validation when sample sizes are small, and adopt evaluation schemes that mirror the intended deployment scenario, including rolling-origin validation for temporal data. The same review further argues that systematic use of multiple metrics, including calibration and probabilistic scoring rules, would improve transparency beyond a single accuracy value. These recommendations provide direct context for the design choices in this thesis. The 2025 holdout follows the rolling-origin logic at a single point; the F1-score is reported alongside precision and recall; and the limitations of inferring distributional behavior from a single tournament are explicitly acknowledged in the discussion of results. The review also flags concerns about external validity across seasons, competitions, and populations, and notes the methodological value of techniques such as transfer learning, domain adaptation, and hierarchical modeling. These directions are not implemented here, but they delineate the boundary of claims that can be drawn from a single-season holdout.

Position of the Present Study

The literature supports three points that guide this thesis. First, machine learning is now well established in sports prediction, but model performance depends heavily on data structure and validation design. Second, basketball outcome prediction is a suitable domain for studying feature representation because many available variables are useful but correlated. Third, PCA is a plausible preprocessing strategy, but its value depends on the classifier and on the tradeoff between compactness and interpretability.

The present study contributes by comparing the raw and PCA-transformed feature spaces for NCAA tournament outcome prediction under identical experimental conditions. Rather than asking whether a single classifier is universally optimal, the thesis asks how changes in representation affect models with different learning mechanisms. This framing connects the broader sports analytics literature to a specific methodological question in basketball forecasting.

The contribution can be situated more precisely with respect to two reference points. The systematic review identifies methodological fragmentation in validation practices, dataset transparency, and benchmarking, and recommends that preprocessing and evaluation be reported in sufficient detail to support cross-study synthesis. The conference paper, in turn, frames feature representation as a central design choice in sports prediction and identifies the comparative analysis between raw and PCA-transformed feature spaces for NCAA tournament forecasting as an open problem. The present thesis combines these two angles by holding the dataset, task, split, and classifiers constant while varying only the feature representation, thereby providing a controlled basis for evaluating whether dimensionality reduction interacts with each classifier's inductive bias in a tournament setting.

Chapter 3. Methodology

Research Design

This thesis follows a comparative experimental design. The same dataset, target variable, train-test split, and evaluation metrics were used across all model configurations. The factor being studied is the representation of the input features. In one condition, models received the original standardized basketball variables. In the second condition, models received a PCA-transformed version of the same information.

This design was selected to isolate the effect of dimensionality reduction. If the dataset or split changed between models, performance differences could be due to sampling variation rather than differences in feature representation. By holding these elements constant, the comparison focuses on whether PCA changes the behavior of SVC, Random Forest, and XGBoost in the NCAA tournament prediction task.

Dataset

The empirical component was built as a pre-game prediction task for NCAA Division I men's basketball tournament games. Each row in the final dataset represents one tournament matchup, not one team-season record. The dataset spans the 2003 to 2025 seasons and contains 136 matchups after the construction and filtering steps. Before dimensionality reduction, the conference paper describes the matchup-level table as containing 34 candidate predictors derived from regular-season team summaries, tournament seeds, Elo ratings, and an estimated team-quality parameter. The source data are publicly distributed through the Kaggle "March Machine Learning Mania" competition, which makes the construction reproducible and aligns the study with the systematic review's recommendation that dataset provenance be reported in sufficient detail to support cross-study synthesis. Every matchup was described with information available before tipoff. This restriction was used to avoid leakage from tournament outcomes into the predictors.

The tournament-game unit of analysis was chosen because the research question concerns predicting game outcomes rather than season rankings. This means that the model must compare two teams in a specific matchup. Team-level season summaries were therefore transformed into matchup-level rows, allowing each observation to include information for both teams and selected difference variables.

The choice of unit of analysis also reflects an intentional position within the four-task taxonomy described in the systematic review. The dataset belongs to the match outcome prediction

category, which the review identifies as the largest single research focus in recent sports machine learning, with 50 of 118 reviewed studies (42%) addressing this task. At the same time, the matchup-level NCAA dataset is comparatively small. With 136 observations and 34 candidate predictors prior to reduction, the ratio between observations and features is modest. This is consistent with the review's observation that sports analytics frequently operates with limited samples relative to feature dimensionality and that this asymmetry contributes to instability in model estimation when correlated predictors are not addressed by preprocessing or by the inductive bias of the classifier [8–10].

Outcome Definition

The prediction task is framed as a binary classification of matchup outcomes. Each row in the modeling table corresponds to one tournament game, and the target is a single indicator that takes the value one if the team designated as Team 1 in that row was the winner of the matchup and the value zero otherwise.[1]

The Team 1 / Team 2 assignment is fixed deterministically: within every matchup, the side with the lower internal team identifier is designated as Team 1 and the other side as Team 2. This rule is purely an ordering convention and carries no competitive meaning. It is needed because the classifier expects a label that is consistent across rows; without it, the same game could legitimately be encoded with a one or with a zero depending on which side was listed first, and the resulting model would learn a contradiction rather than a pattern. After applying the ordering, the dataset is filtered to remove the mirrored copy of each game, since the NCAA tournament is played at neutral venues and the two orderings of the same matchup describe the same competitive event rather than two independent observations.

A consequence of this construction is that the binary target should not be read as a marker of favored team, home-court advantage, seed, or any basketball-meaningful direction. It is, by design, only the outcome that is consistent with the fixed Team 1 / Team 2 assignment, and any predictive signal that the classifier discovers must come from the relative features of the two sides rather than from the labelling convention itself.

Feature Construction

Regular-season games are first aggregated into per-team seasonal profiles, with every variable computed strictly from games played before the start of the tournament. Each profile is then

[1] The same target encoding, expressed as a piecewise definition, appears in the related conference paper; the prose form used here is the one adopted for the thesis.

duplicated into a matchup-level row, one copy per side of the matchup, and the two halves are joined by a small set of difference variables that capture the relative gap between the two opponents on dimensions where the gap itself, rather than either side's level, drives the outcome.

The resulting feature set is best understood as four content blocks, organized by basketball concept rather than by storage convention.

Scoring volume. The first block describes how a team produces offense. It includes the season-average points scored and the season-average number of field-goal attempts. These two variables are kept separate because they decouple efficiency from pace: two teams may average the same point total while attempting different shot volumes, and that distinction is relevant when comparing styles in an elimination setting.

Rebounding and rim activity. The second block summarizes the way each team interacts with the rim and with possessions that follow a missed shot. It includes the average number of offensive and defensive rebounds per game and the average number of blocks recorded by the team. These three variables together describe possession recovery on offense, possession termination on defense, and rim protection.

Foul exposure and opponent pressure. The third block captures how each team behaves under contact and how opposing teams behave against it. It includes personal fouls committed by the team, fouls committed by opponents against the team, opponent field-goal attempts allowed, and blocks recorded by opponents against the team. Together these variables sketch the defensive activity of the team and the offensive load that opponents are typically able to generate against it.

Outcome-summary indicators. The fourth block contains variables that summarize team strength directly rather than describing a single facet of play. Tournament seed encodes the selection committee's pre-tournament judgement; Elo rating provides a continuous strength score updated game by game throughout the regular season; the average regular-season point differential captures how decisively the team won or lost; and a latent quality coefficient (described below) re-expresses point margins after adjusting for the opponents faced. These four variables are conceptually overlapping by construction, in that all of them are proxies for a single underlying competitive-strength attribute.

In addition to these absolute team descriptors, the matchup row contains three difference variables that compare the two teams directly. Writing $\Delta(\cdot)$ for the gap on a given dimension between Team 1 and Team 2:

$$\Delta\text{Seed}, \qquad \Delta\text{Elo}, \qquad \Delta\overline{\text{PointDiff}},$$

defined respectively as the difference of pre-tournament seeds, the difference of pre-tournament Elo ratings, and the difference of average regular-season point margins. The exact column-level naming convention used to store these descriptors in the modeling table is documented in the supplementary listing referenced at the end of this section, which is adapted from the dataset description in the related conference paper.

A latent quality score was estimated from regular-season point margins to capture team strength after accounting for opponent quality. Let $Y_{i,j}$ denote the point differential in a game between Team i and Team j. The model expressed the observed margin as the difference between team strength parameters:

$$Y_{i,j} = \alpha_i - \alpha_j + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2),$$

where α_i and α_j are team-specific quality parameters. The parameters were estimated with a generalized linear model using a Gaussian identity link. Team identities were included in the model as fixed effects, and the intercept was suppressed so that each coefficient could be interpreted relative to the field. Under this formulation, the gap between the two team quality scores approximates an expected point-margin advantage conditioned on the opponents faced during the regular season. The conference paper presents this construction as a continuous-margin adaptation of a Bradley-Terry-style framework, in which each team's coefficient reflects intrinsic quality relative to the league average, and the metric implicitly accounts for strength of schedule by conditioning on the opponents faced over the season.

The four content blocks together yield the modeling features used in this thesis. The thesis dataset deliberately preserves the same predictor structure as the related conference paper, so that any difference in results across feature representations can be attributed to PCA rather than to a redefinition of the feature set. The overlap among the outcome-summary indicators — seed, Elo, point differential, and the latent quality coefficient — is the empirical motivation for the comparison studied in Chapter 4: if the four variables are alternative proxies for one underlying attribute, a representation that compresses redundant directions may either help or harm a classifier depending on how the classifier exploits that redundancy.

Table 3 summarises the four content blocks at a glance. It is grouped by basketball concept rather than by the storage names used in the modeling table, since the latter follow the per-team `T1_*`/`T2_*` convention adopted from the public dataset and reproduced in the conference paper. The exact column names per block are listed in the supplementary feature dictionary (Appendix 8) for reference and reproducibility.

Table 3: Feature content blocks used in the matchup dataset, grouped by basketball concept rather than by storage name. Per-team column names following the public-dataset convention are reproduced in Appendix 8.

Content block	**Variables included**
Scoring volume	Average points scored per game; average field-goal attempts per game.
Rebounding and rim activity	Average offensive rebounds; average defensive rebounds; average blocks.
Foul exposure and opponent pressure	Personal fouls; opponent field-goal attempts allowed; opponent blocks against the team; fouls drawn from opponents.
Outcome-summary indicators	Tournament seed; pre-tournament Elo rating; average regular-season point differential; latent quality coefficient.
Relative matchup variables	ΔSeed, ΔElo, and $\Delta\overline{\text{PointDiff}}$, computed as the gap between Team 1 and Team 2 on the corresponding outcome-summary indicator.

Data Preprocessing

The numerical feature matrix was placed on a common scale before model fitting. Each continuous predictor was centered by subtracting its training-fold mean and divided by its training-fold standard deviation, so that every column entered the modeling pipeline with comparable magnitude and dispersion. Concretely, the means and standard deviations used for centering and rescaling were computed exclusively over the seasons assigned to model development; the 2025 holdout column was transformed using those same training-derived statistics rather than statistics computed on its own values. This ordering matters because two of the methods used downstream — Principal Component Analysis and Support Vector Classification — interact with the input geometry through Euclidean distances and inner products, so any rescaling applied with knowledge of the holdout season would silently leak target-period information into the training stage [13, 24].

The final modeling table contained no missing values and no categorical predictors. As a consequence, the preprocessing pipeline reduced to scale alignment followed, in the PCA condition, by a learned linear projection. This narrowness is a property of the dataset rather than a design choice, since matchup-level summaries derived from regular-season averages are by construction continuous and dense, and it aligns with the systematic review's observation that match outcome studies typically operate on tabular numerical data prepared from regular-season summaries. The principal preprocessing risks in this configuration are there-

fore concentrated in scaling and dimensionality reduction, both of which are managed within the training pipeline so that the 2025 holdout tournament does not influence preprocessing parameters.

Principal Component Analysis

PCA was evaluated as an alternative representation of the same standardized predictors [13]. The procedure operates on the centred training feature matrix and produces a new set of axes that are orthogonal in the standardized space and ranked by the share of feature variance they capture. Each principal axis is defined by a unit-norm loading vector that mixes the original basketball variables with positive and negative weights, and each observation is then re-expressed as the inner product between its standardized vector and these loadings. The resulting representation has the same number of rows as the input but a smaller number of columns once a retention rule has been applied.

In notation aligned with the rest of this thesis, write X_{train} for the standardized training feature matrix, $\mathbf{w}_1, \ldots, \mathbf{w}_k$ for the leading loading vectors stacked column-wise into a matrix W, and $Z_{\text{train}} = X_{\text{train}} W$ for the corresponding score matrix. The same loadings W — fitted exclusively on the training fold — are then applied to the held-out 2025 tournament rows, producing the test-time representation $Z_{\text{test}} = X_{\text{test}} W$. The number of retained components, k, is not fixed in advance: it is the smallest integer for which the cumulative explained variance on the training data reaches 95%. This makes k data-dependent and tied to the seasonal split.

The 95% retention threshold was selected because it balances two competing concerns. Setting the threshold lower would discard directions of variation that, although smaller, may still contribute information about competitive style. Setting it higher would defeat the purpose of reduction, since the trailing components in standardized basketball features tend to absorb noise and individually-correlated micro-variation rather than substantive structure. Critically, the threshold is defined on predictor variance rather than on classification performance, which keeps the PCA step unsupervised and prevents the dimensionality choice from being implicitly tuned on the target variable.

Machine Learning Models

Three supervised classifiers were compared. Random Forest was included as a bagged ensemble of decision trees that reduces variance through bootstrap sampling and randomized split selection [22]. XGBoost was included as a regularized boosted-tree method in which trees are added sequentially to reduce the current prediction error [23]. Support Vector Clas-

sification was included because its margin-based objective makes it sensitive to the geometry, scale, and correlation structure of the input space [24].

Each classifier was trained twice. One version used the standardized original predictors, while the other used the PCA component representation. Hyperparameters were selected through random search using only the training seasons. The preprocessing and modeling steps were treated as a single pipeline, keeping the holdout tournament isolated until final evaluation.

The three classifiers were selected to make the comparison methodologically useful. SVC represents a margin-based classifier, Random Forest represents a bagged tree ensemble, and XGBoost represents a boosted tree ensemble. If PCA affects these models differently, the results can be interpreted in relation to how each algorithm uses the feature space.

Training and Evaluation Protocol

Evaluation followed a season-wise holdout design. Games from 2003 to 2024 were used to develop the model, and the 2025 tournament was reserved for final testing. This split was chosen because it resembles the intended use case: a model is trained on past tournaments and then applied to a future tournament.

Precision, recall, and F1-score were used to evaluate classification behavior. F1-score served as the primary comparison metric because it combines false positives and false negatives into a single metric. Since every classifier was evaluated under both feature representations with the same split, the comparison focuses on the effect of the representation rather than on changes in the validation protocol.

The holdout design has one important implication. The reported results describe performance on the 2025 tournament, not an average across all possible tournament years. This makes the evaluation realistic but also limited. For this reason, the discussion treats small differences with caution and emphasizes model behavior rather than claiming broad superiority based on a single season.

Chapter 4. Results

Overview of the Evaluation

This chapter presents the empirical comparison between raw and PCA-transformed feature representations. The evaluation focuses on the 2025 holdout tournament, which was not used during model fitting or hyperparameter selection. For each classifier, the raw-feature and PCA-feature versions were compared using the same test set and performance metrics.

The results should be read as a comparison of model configurations under one realistic forecasting scenario. A higher score indicates better performance on the 2025 holdout set, but small differences should be interpreted cautiously because tournament outcomes are noisy and the holdout sample is limited. The goal is therefore to identify the direction and size of the PCA effect for each classifier, not to claim a universal ranking of all possible tournament models.

Model Performance

Table 4 reports the 2025 holdout performance for the six model configurations. The models produced a narrow performance band, with F1-scores ranging from 81.95% to 84.33%. The best result was obtained by SVC trained on PCA components. That configuration reached 84.33% precision, 84.33% recall, and 84.33% F1-score.

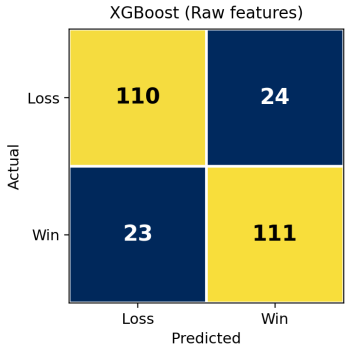
Among the raw-feature models, Random Forest obtained the strongest result, with an F1-score of 83.58%. The raw SVC model reached 82.84%, and raw XGBoost reached 81.95%. Among the PCA-based models, PCA-SVC ranked first, PCA-RF reached 82.83%, and PCA-XGBoost reached 82.09%. These values show that PCA did not produce a uniform improvement across the model set.

Effect of PCA by Model

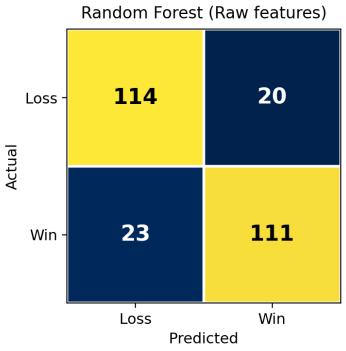
The three classifiers responded to the change in representation in three qualitatively different ways, and they are most clearly read in the order of the magnitude and direction of that response.

Random Forest is the clearest negative case. Its F1-score on the original feature space, 83.58%, dropped to 82.83% once the matchup was re-expressed as principal components — a decline of roughly three quarters of a percentage point. A plausible reading is that the

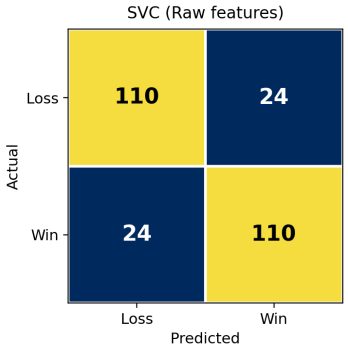
Raw feature representation



(a) XGBoost

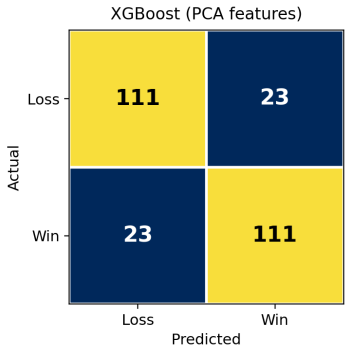


(b) Random Forest

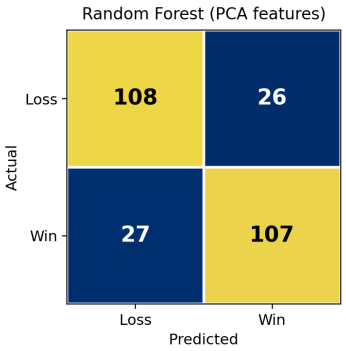


(c) SVC

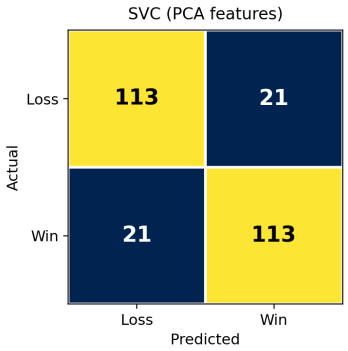
PCA-transformed feature representation



(d) PCA-XGBoost



(e) PCA-RF



(f) PCA-SVC

Figure 1: Confusion matrices on the 2025 holdout set, organised by feature representation. Each panel uses the cividis colormap, which is perceptually uniform and accessible to viewers with deuteranopia and protanopia; the colorbar is omitted because cell values are annotated directly. Reading column-wise, each pair (raw above and PCA below) corresponds to the same classifier under the two representations and isolates the effect of dimensionality reduction. Reading row-wise, the three panels in each row span the boosted-tree, bagged-tree, and margin-based classifier families.

Table 4: Performance on the 2025 holdout set, paired by classifier and feature representation. The ΔF1 column reports the change in F1-score induced by replacing the raw feature space with the PCA representation, holding the classifier fixed. The Rank column orders all six configurations from highest to lowest F1.

Classifier	Representation	Precision	Recall	F1-score	ΔF1 vs. Raw	Rank
XGBoost	Raw	82.58%	81.34%	81.95%	—	6
	PCA	82.10%	82.09%	82.09%	+0.14 pp	5
Random Forest	Raw	83.59%	83.58%	83.58%	—	2
	PCA	82.90%	82.84%	82.83%	−0.75 pp	3
SVC	Raw	82.84%	82.84%	82.84%	—	4
	PCA	84.33%	84.33%	84.33%	+1.49 pp	1

recursive-partitioning logic of a forest is well aligned with raw basketball variables that admit informative thresholds (e.g., a seed gap large enough to flip a node, an Elo difference past which the historical win rate shifts). When those variables are linearly recombined into orthogonal directions, the same information survives in the data but the threshold structure that the trees can exploit is rotated, and individual splits become less crisp.

XGBoost sits at the opposite extreme: representation barely matters. Going from 81.95% on the raw features to 82.09% on the PCA components is a 0.14-point change, well inside the noise floor implied by a single-tournament holdout. The boosted ensemble appears equally able to extract signal from either input geometry, which is consistent with its sequential error-correction objective and its built-in regularization on tree weights and leaf scores. In neither direction does the change in representation move the model meaningfully.

SVC shows the only clearly positive response. Its F1-score rose from 82.84% on the raw features to 84.33% on the principal components, a gain of about 1.5 percentage points and the largest movement of any model–representation pair in the experiment. This is the configuration most exposed to the geometry of the input space: a margin-based classifier optimizes a decision surface defined in Euclidean coordinates, and that optimization benefits when correlated directions in the basketball variables are merged into orthogonal axes of comparable scale. The corresponding confusion matrices in Figure 1 are consistent with this interpretation: PCA-SVC distributes its remaining errors more symmetrically across the two outcome classes than the raw-feature SVC.

Comparison Across Feature Representations

The raw feature space yielded the best result for Random Forest, whereas the PCA feature space yielded the best result for SVC. This contrast is the main result of the chapter. It indicates that the value of dimensionality reduction depends on how the classifier uses predictor information. PCA changes the geometry of the feature space. For SVC, this change appears helpful. For Random Forest, it appears to remove or blur information that was useful in the original variables.

The XGBoost comparison was nearly neutral. The PCA version was slightly better by F1-score, but the difference was only 0.14 percentage points. This suggests that XGBoost was relatively robust to changes in representation in the present dataset. The boosted model may have been able to learn from either the original variables or the component representation without a large change in classification behavior.

The narrow range of scores also shows that no configuration failed on the task. Even the lowest F1-score, 81.95%, was close to the strongest raw-feature and PCA-feature models. For applied tournament forecasting, this means that model selection should consider more than the largest point estimate. Interpretability, maintenance, computational simplicity, and year-to-year stability are also relevant.

Summary of Findings

The empirical results support a model-dependent interpretation of PCA. The highest score came from PCA-SVC, but the strongest raw-feature model was Random Forest. The narrow F1-score range indicates that all configurations were viable for this dataset. The practical conclusion is that feature representation matters, but it should be selected with the classifier in mind rather than applied as a default preprocessing step.

These findings partially support the thesis hypothesis. PCA improved the classifier expected to be most sensitive to feature geometry, namely SVC. It did not provide a meaningful advantage for XGBoost and reduced Random Forest performance. The hypothesis is therefore supported only in its model-dependent form.

Chapter 5. Discussion

Interpretation of Results

The model-level differences can be understood by how each classifier uses the input variables. SVC depends strongly on the geometry of the feature space. When predictors are correlated or unevenly scaled, the separating boundary can become sensitive to redundant directions in the data, especially with a small tournament sample. PCA reduced that redundancy and replaced the original predictors with orthogonal components, which likely made the margin optimization more stable.

Tree-based ensembles use the feature space differently. Random Forest and XGBoost make recursive splits and can select informative variables even when related predictors are present. For Random Forest, the raw feature set performed better. One likely explanation is that variables such as seed difference, Elo difference, and point differential contain threshold patterns that trees can exploit directly. PCA turns those variables into linear combinations, which can make such thresholds less clear to a tree ensemble.

The strength of the comparative conclusion is also bounded from below by an empirical observation about the floor of the experiment. The weakest of the six configurations still achieved 81.95% on F1, and every other configuration sat within roughly two and a half points of that value. Read in this direction, the comparison does not surface a dominant strategy so much as a band of broadly serviceable ones, and the interesting differences are in which classifier benefits from which representation rather than in absolute predictive power. That, in turn, shifts the practical decision toward the tradeoff among accuracy, feature compactness, and interpretability that the next section develops.

Relation to the Literature

The findings are consistent with the broader sports analytics literature, which shows that ensemble methods, SVMs, and regression-style baselines remain competitive on structured sports datasets [5, 8]. The results also match the systematic review's observation that model choice must be evaluated in relation to data type and validation design. For tabular team statistics, traditional machine learning methods remain strong. More complex models are not automatically superior unless the data structure supports them.

The thesis also contributes to the literature on feature representation. Prior studies have used PCA to summarize sports performance profiles [14], but fewer studies directly compare raw and PCA-transformed feature spaces in NCAA tournament prediction. The present findings

show that dimensionality reduction can be useful, but the advantage is conditional rather than general.

The result also fits the broader validation concerns discussed in Chapter 2. Sports prediction studies often report a single best model without showing how design decisions affect that model. Here, the same classifier can behave differently after a change in representation. This supports the idea that preprocessing is part of the modeling strategy, not a neutral technical step.

Interpretability

Interpretability is a central tradeoff. Raw variables allow analysts to connect predictions to basketball concepts. Seed difference, Elo difference, scoring margin, and rebounding statistics have direct meaning. PCA components are more compact and reduce correlation, but they are less transparent. In applied sports analytics, this loss matters because coaches, analysts, and decision-makers often need to understand why a prediction was made.

Therefore, PCA is most attractive when the objective is compact prediction or when the classifier benefits from orthogonal inputs. When explanation in domain terms is the priority, the raw feature space may be preferable, especially for tree-based models that already handle correlated variables reasonably well.

This distinction matters in basketball analytics. A coach or analyst can reason about seed differences, Elo differences, or point differentials. These variables connect the prediction to basketball knowledge. A principal component may improve model behavior, but it requires an additional explanation layer. The appropriate choice, therefore, depends on whether the model is used mainly for forecasting accuracy, exploratory analysis, or decision support.

Limitations

Several limitations should be considered. First, the evaluation used a single 2025 holdout tournament. Although this design reflects a realistic forecasting scenario, it does not provide a distribution of performance across many future seasons. Second, the dataset contained 136 tournament matchups, which limits the statistical power of model comparisons. Third, the study used team-level regular-season summaries and excluded player injuries, lineup combinations, travel effects, and tracking-derived information. Fourth, PCA was evaluated using a 95% explained variance threshold, but alternative thresholds or supervised dimensionality reduction methods may yield different results.

An additional limitation is that the comparison uses only classification metrics. Tournament prediction is often used in probabilistic settings, such as bracket simulation or upset-risk estimation. A model with a similar F1-score may still differ in calibration. Future work should therefore evaluate not only whether the predicted winner is correct but also whether the predicted probabilities match the observed frequencies.

Future Work

Future research should evaluate the same modeling pipeline across multiple rolling season-wise splits. This would make it possible to estimate confidence intervals and test whether the observed PCA-SVC advantage persists across years. Additional work should incorporate player-level statistics, injury reports, lineup information, and possession-level features. These data may improve model sensitivity to matchup-specific conditions.

Future studies should also compare PCA with alternative representation methods, including feature selection, partial least squares, autoencoders, and supervised embedding approaches. Finally, probabilistic forecasting and calibration metrics should be added because tournament prediction is often used to estimate probabilities, not only binary winners.

Chapter 6. Conclusion

This thesis evaluated whether PCA-based feature representation improves machine learning models' ability to predict NCAA Division I men's basketball tournament outcomes. The work was motivated by a common problem in sports analytics: modern basketball datasets contain many useful variables, but several of them describe overlapping aspects of team quality. In that setting, preprocessing decisions can change model behavior as much as the choice of classifier.

A matchup-level dataset was built from the 2003 to 2025 NCAA tournament seasons using information available before each tournament game. The predictors included regular-season team statistics, seed information, Elo ratings, and latent team quality estimates. SVC, Random Forest, and XGBoost were each trained under two input conditions: original standardized features and PCA-transformed features. Performance was evaluated for the 2025 tournament using precision, recall, and F1-score.

The results showed that all evaluated models performed competitively. PCA-SVC achieved the best F1-score at 84.33%, while Random Forest trained on raw features was the strongest non-PCA model at 83.58%. PCA improved SVC, had a negligible effect on XGBoost, and reduced Random Forest performance. These findings support the thesis hypothesis in a qualified way: PCA can improve or preserve predictive performance, but its benefit depends on the classifier.

The main conclusion is that dimensionality reduction should be treated as a model-specific design decision in sports analytics. PCA can provide a compact and effective representation when the classifier benefits from orthogonal standardized inputs. However, raw features remain valuable when interpretability and threshold-based relationships matter. For NCAA tournament prediction, preprocessing choices should therefore be evaluated empirically rather than assumed to be beneficial.

The thesis makes three main contributions. First, it connects the empirical study to a broader synthesis of machine learning research in sports prediction, showing that outcome forecasting remains a central but methodologically varied task. Second, it provides a direct comparison between raw and PCA-based feature spaces for NCAA tournament prediction. Third, it shows that the same preprocessing step can help one classifier while weakening another, which reinforces the need to evaluate feature representation as part of the modeling design.

Future research should extend the analysis across multiple rolling holdout seasons to determine whether the PCA-SVC advantage persists beyond the 2025 tournament. Additional work should include player-level variables, injury reports, lineup information, and richer possession-level or tracking-derived features. Probabilistic evaluation should also be added,

since tournament forecasting often requires calibrated probabilities rather than binary winner predictions alone. Finally, alternative representation strategies, such as feature selection, partial least squares, supervised embeddings, or nonlinear dimensionality reduction, could be compared with PCA to determine whether other compact representations preserve interpretability while improving generalization.

Appendix A. Supplementary Feature Dictionary

This appendix documents the storage-level column names of the variables described, by basketball concept, in Section 3.4. The naming convention follows the public Kaggle "March Machine Learning Mania" dataset and is reproduced for reference and reproducibility; it is adapted from the dataset description provided in the related conference paper. Throughout, `T1_*` denotes a column corresponding to Team 1 in the matchup row and `T2_*` denotes the analogous column for Team 2.

Scoring volume. `T1_avg_Score`, `T2_avg_Score`: average points scored per regular-season game by each team. `T1_avg_FGA`, `T2_avg_FGA`: average field-goal attempts per regular-season game.

Rebounding and rim activity. `T1_avg_OR`, `T2_avg_OR`: average offensive rebounds per game. `T1_avg_DR`, `T2_avg_DR`: average defensive rebounds per game. `T1_avg_Blk`, `T2_avg_Blk`: average blocks per game.

Foul exposure and opponent pressure. `T1_avg_PF`, `T2_avg_PF`: average personal fouls committed per game. `T1_avg_opponent_FGA`, `T2_avg_opponent_FGA`: average field-goal attempts allowed per game. `T1_avg_opponent_Blk`, `T2_avg_opponent_Blk`: average blocks recorded by opponents against the team. `T1_avg_opponent_PF`, `T2_avg_opponent_PF`: average fouls drawn from opponents.

Outcome-summary indicators. `T1_seed`, `T2_seed`: NCAA tournament seed assigned to each team. `T1_elo`, `T2_elo`: pre-tournament Elo rating of each team. `T1_avg_PointDiff`, `T2_avg_PointDiff`: average point differential per regular-season game. `T1_quality`, `T2_quality`: latent quality coefficient estimated from the regular-season GLM described in Section 3.4.

Relative matchup variables. `Seed_diff`, `elo_diff`, and `PointDiff_diff` denote the corresponding paired differences between Team 1 and Team 2 on seed, Elo, and average point differential.

Identifiers and metadata. `Season` (tournament year), `T1_TeamID` and `T2_TeamID` (team identifiers), and `men_women` (a constant indicator retained for compatibility with the source schema) are kept for record-linkage purposes and are not used as model inputs.

References

[1] Michael Mondello and Christopher Kamke. The introduction and application of sports analytics in professional sport organizations. *Journal of Applied Sport Management*, 6

(2):11, 2014. doi: 10.7290/jasm06ldmv.

[2] Louis Passfield and James G. Hopker. A mine of information: Can sports analytics provide wisdom from your data? *International Journal of Sports Physiology and Performance*, 12(7):851–855, 2017. doi: 10.1123/ijspp.2016-0644.

[3] Bryan L. Boulier and Herman O. Stekler. Are sports seedings good predictors? an evaluation. *International Journal of Forecasting*, 15(1):83–91, 1999. doi: 10.1016/S0169-2070(98)00067-3.

[4] Bryan L. Boulier, Herman O. Stekler, Jason Coburn, and Timothy Rankins. Evaluating national football league draft choices: The passing game. *International Journal of Forecasting*, 26(3):589–605, 2010. doi: 10.1016/j.ijforecast.2009.10.009.

[5] Nader Chmait and Hans Westerbeek. Artificial intelligence and machine learning in sport research: An introduction for non-data scientists. *Frontiers in Sports and Active Living*, 3, 2021. ISSN 2624-9367. doi: 10.3389/fspor.2021.682287.

[6] Yunke Jia, Norli Anida Abdullah, Hafiz Eliza, Qingbo Lu, Deyou Si, Hengwei Guo, and Wenliang Wang. A narrative review of deep learning applications in sports performance analysis: Current practices, challenges, and future directions. *BMC Sports Science, Medicine and Rehabilitation*, 17(1):249, 2025. doi: 10.1186/s13102-025-01294-0.

[7] Christopher Leckey, Nicol van Dyk, Cailbhe Doherty, Aonghus Lawlor, and Eamonn Delahunt. Machine learning approaches to injury risk prediction in sport: A scoping review with evidence synthesis. *British Journal of Sports Medicine*, 59(7):491–500, 2025. doi: 10.1136/bjsports-2024-108576.

[8] Tomislav Horvat and Josip Job. The use of machine learning in sport outcome prediction: A review. *WIREs Data Mining and Knowledge Discovery*, 10(5):e1380, 2020. doi: 10.1002/widm.1380.

[9] Elia Morgulev, Ofer H Azar, and Ronnie Lidor. Predicting outcomes in sport: A meta-analysis of the predictive yielding of experts, statistical models and the crowd. *International Review of Sport and Exercise Psychology*, 11(1):89–115, 2018.

[10] José Miguel Contreras, Elena Molina Portillo, and Juan Manuel Fernández Luna. Evaluation of hierarchical clustering methodologies for identifying patterns in timeout requests in euroleague basketball. *Mathematics*, 13(15), 2025. ISSN 2227-7390. doi: 10.3390/math13152414. URL `https://www.mdpi.com/2227-7390/13/15/2414`.

[11] Gang Shen, Di Gao, Qian Wen, and Rhonda Magel. Predicting results of march madness using three different methods. *Journal of Sports Research*, 3(1):10–17, 2016. doi: 10.18488/journal.90/2016.3.1/90.1.10.17.

[12] Sheldon West. March madness: How to pick a winner. *Nature*, 531(7594):S10–S12, 2016.

[13] Ian T. Jolliffe. Principal component analysis. *Springer Series in Statistics*, 2002.

[14] Scott Bruce. A scalable framework for nba player and team comparisons using player tracking data. *arXiv preprint arXiv:1511.04351*, 2015.

[15] Konstantinos Apostolou and Christos Tjortjis. Sports analytics algorithms for performance prediction. In *Proceedings of the 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–4, 2019. doi: 10.1109/IISA.2019.8900754.

[16] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, et al. The prisma 2020 statement: An updated guideline for reporting systematic reviews. *Systematic Reviews*, 10(1):89, 2021. ISSN 2046-4053. doi: 10.1186/s13643-021-01626-4.

[17] Elia Morgulev, Ofer H. Azar, and Ronnie Lidor. Sports analytics and the big-data era. *International Journal of Data Science and Analytics*, 5(4):213–222, 2018. doi: 10.1007/s41060-017-0093-7.

[18] Dan Cristian Mănescu. Big data analytics framework for decision-making in sports performance optimization. *Data*, 10(7):116, 2025. ISSN 2306-5729. doi: 10.3390/data10070116.

[19] Vanessa Meyer, Lena Wiese, and Ahmed Al-Ghezi. An information system for training assessment in sports analytics. In *Proceedings of the 26th International Conference on Enterprise Information Systems (ICEIS)*, pages 149–160. SciTePress, 2024. ISBN 978-989-758-692-7. doi: 10.5220/0012724200003690.

[20] Ahmet Cagdas Seckin, Bahar Ates, and Mine Seckin. Review on wearable technology in sports: Concepts, challenges and opportunities. *Applied Sciences*, 13(18):10399, 2023. ISSN 2076-3417. doi: 10.3390/app131810399.

[21] Nitin Singh. Sport analytics: A review. *The International Technology Management Review*, 9(1):64–69, 2020. doi: 10.2991/itmr.k.200831.001.

[22] Leo Breiman. Random forests. *Machine Learning*, 45, 2001. doi: 10.1023/A:1010933404324.

[23] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016. doi: 10.1145/2939672.2939785.

[24] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2002.

[25] Stephanie A. Kovalchik. Player tracking data in sports. *Annual Review of Statistics and Its Application*, 10:677–697, 2023. doi: 10.1146/annurev-statistics-033021-110117.

[26] Joachim Gudmundsson and Michael Horton. Spatio-temporal analysis of team sports. *ACM Computing Surveys*, 50(2):1–34, 2017. doi: 10.1145/3054132.

[27] Tomislav Horvat, Josip Job, Robert Logozar, and Caslav Livada. A data-driven machine learning algorithm for predicting the outcomes of nba games. *Symmetry*, 15(4):798, 2023. doi: 10.3390/sym15040798.

[28] Jesse Davis, Lotte Bransen, Laurens Devos, Arne Jaspers, Wannes Meert, Pieter Robberechts, Jan Van Haaren, and Maaike Van Roy. Methodology and evaluation in sports analytics: Challenges, approaches, and lessons learned. *Machine Learning*, 113(9): 6977–7010, 2024. doi: 10.1007/s10994-024-06585-0.

[29] Elia Morgulev and Felix Lebed. Beyond key performance indicators. *German Journal of Exercise and Sport Research*, 54(3):335–340, 2024. doi: 10.1007/s12662-024-00944-8.