IDS 705 Final Report

Forecast the 2024 College Basketball Tournaments

# **March Machine Learning Mania 2024**

****

**Team Members**:

*Shaila Guereca*

*Ayush Gupta*

*Cassie Kang*

**Team Number**: 6

# Abstract

In this study, we aim to forecast the outcomes of both the men's and women's 2024 collegiate basketball tournaments by employing a portfolio based approach rooted in historical data analysis. Our objective extends beyond mere prediction; we seek to investigate whether game strategy significantly changes with different coaches, focusing specifically on the Indiana team. Indiana was chosen due to its notable history of NCAA Basketball Sweet 16 Appearances, particularly amidst coaching transitions, including six different coaches from 1985 to 2023. We expect our clustering algorithm to partition the data into distinct coaching eras, each characterized by unique game strategies and team dynamics. Through the application of unsupervised learning techniques, such as t-SNE for dimensionality reduction and k-means clustering for data classification, we aim to uncover patterns that elucidate the impact of coaching changes on the team's playing style and overall performance throughout its history.

Additionally, we will explore predictive models for March Madness outcomes, including Logistic Regression and Recurrent Neural Networks (RNNs), to discern the most effective approach for forecasting game outcomes. Our study offers valuable insights into the dynamic relationship between coaching transitions and team performance in collegiate basketball.

# Introduction

In the world of college basketball, where excitement and accuracy come together, there's an exciting challenge: predicting who will win in the men's and women's 2024 tournaments.

What we find exciting about this project is the chance to explore March Madness, where every move in the game can change its course. We're eager to figure out what factors affect game results using past data.

We aim to predict the winners and losers of both tournaments accurately by creating a set of brackets based on evidence and careful analysis. Using historical data as our roadmap, we want to predict outcomes as accurately as possible to get the highest score.

# Background

Several studies have explored the potential of machine learning for predicting basketball game outcomes.

For example, Bae and Kim (2017) and Fernandez et al. (2013) directly investigate the use of machine learning techniques like neural networks for NBA game prediction. Their work highlights the potential of this approach in analyzing sports data.

Beyond just machine learning, researchers have also explored statistical models that incorporate factors beyond win-loss records. Beaver et al. (2019) focus on NCAA tournaments, developing a hierarchical random effects model that considers additional variables to improve prediction accuracy. Similarly, Byrnes and Johnson (2010) delve into the importance of team interactions, suggesting that solely focusing on individual player performance might not be sufficient for accurate predictions.

Extracting valuable data points from existing statistics is crucial for building strong prediction models. Sanchez et al. (2022) emphasize the importance of feature engineering, which involves creating relevant data points from raw statistics. This aligns with the work of Goulias and Stylianou (2016) and Johnson and Raab (2011) who focus on identifying key performance indicators that can be used for prediction.

With more sophisticated techniques Liu and Yang (2018) introduced a deep learning method for prediction based on offensive and defensive ratings and Xu and Cohen (2014) explored predicting college basketball performance using team statistics, demonstrating the applicability of these methods beyond the NBA.

All this research confirms that this topic has already been explored and that various machine learning techniques have been used for predicting basketball games.

# Data

The datasets for NCAA basketball tournaments that we will be working with provide a detailed and structured insight into the intricacies of college basketball, covering a wide range of aspects from team information to game outcomes and rankings. We have information about men's and women's teams, including their unique IDs, names, and their tenure in Division I. It also outlines the specifics of each season, such as start dates and regional identifiers, along with seeding information for NCAA tournaments which sheds light on team placements and structures within the tournaments.

We can also observe detailed game statistics for both regular season and tournament games are presented. These datasets offer a granular look at team performances across a variety of metrics including scoring, rebounds, assists, and defensive actions. This allows for a deep dive into the game-by-game performance of teams.

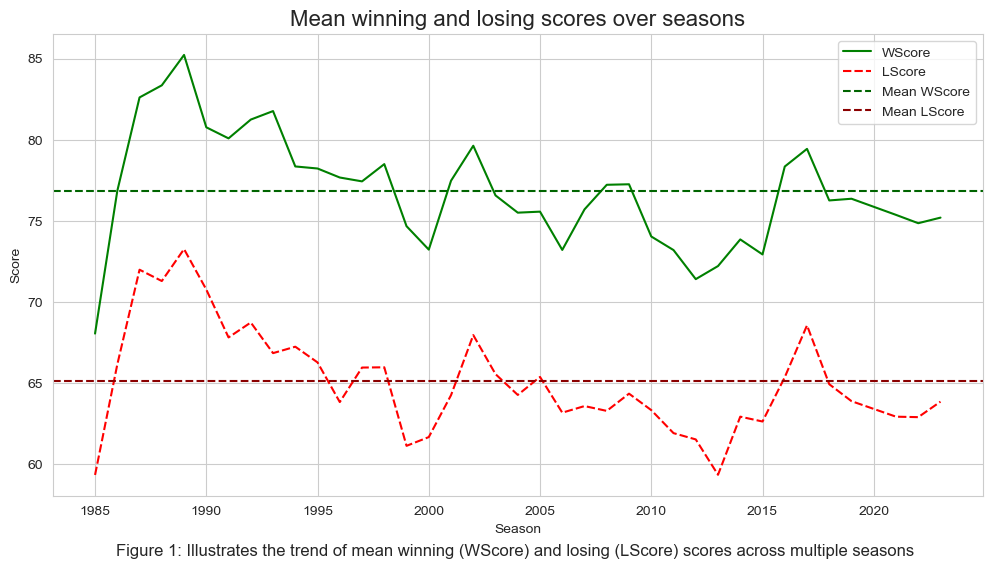
Geographically, the datasets pinpoint the locations where games have been played, listing cities and associating them with specific games in both regular seasons and tournaments. This geographical aspect adds another layer of analysis, potentially exploring home advantage or travel impacts on team performance.

The public rankings section provides a historical view of team performances according to various rating systems. These weekly rankings since the 2002-2003 season present a diverse perspective on team strengths and shifts in performance over time.

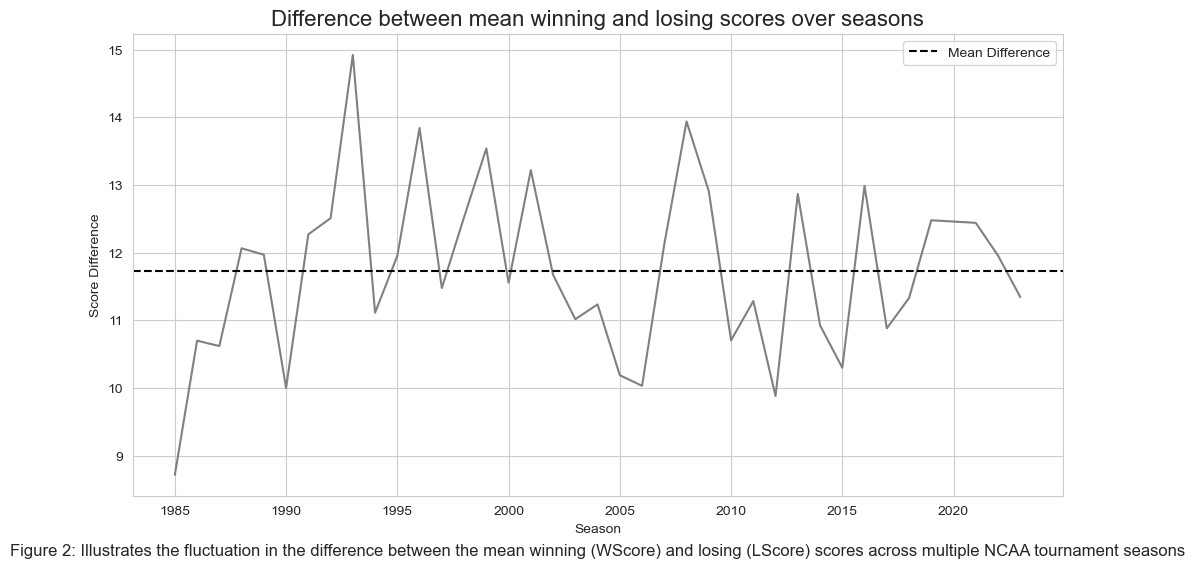
We also have supplemental information on coaching records, detailing head coaches' tenures for each team. Conference affiliations and tournament games offer insights into the competitive landscape of college basketball. Additionally, details on postseason tournaments outside of the NCAA provide a broader view of college basketball's postseason activities. Alternative team spellings ensure that external data can be accurately linked, enhancing the datasets' utility for comprehensive analysis.

The period we will be analyzing spans from 1985 to 2023, with a total of 378 male teams and 376 female teams throughout the entire period.

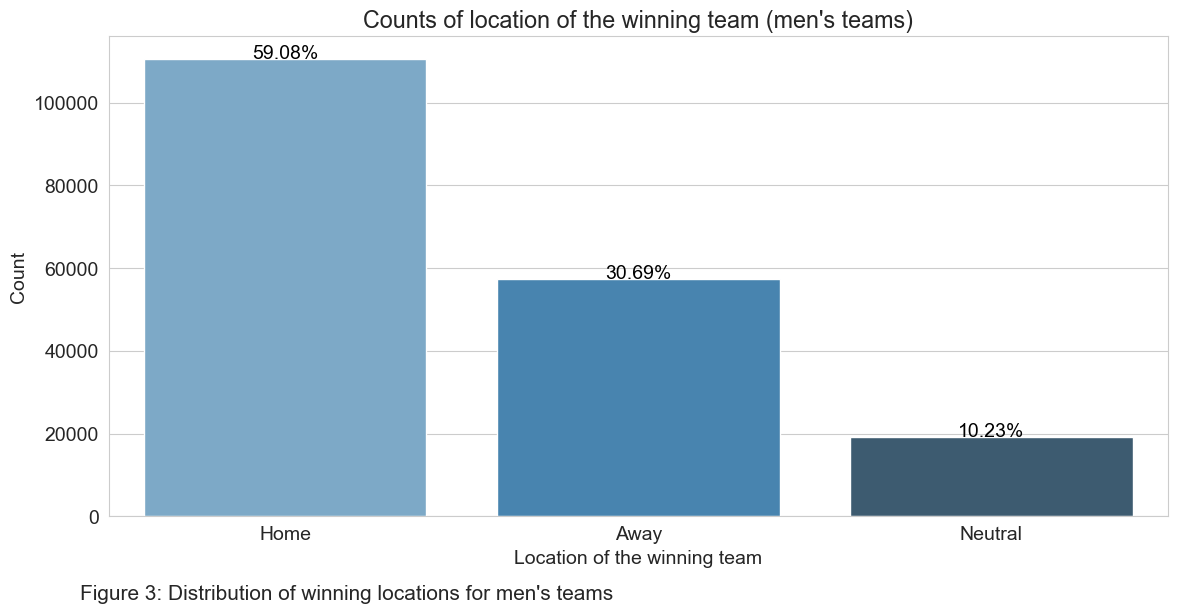
We want to investigate the behavior of the number of points scored when a team wins and loses a game. To do this, we can observe Figure 1, which shows that over time, the average points scored by the losing team peaked at 73 points during 1987, with the winning teams also reaching their peak at 86 points. It is interesting to note that the average points scored have decreased over time, with the current average being 75 points for winning teams and 64 points for losing teams.

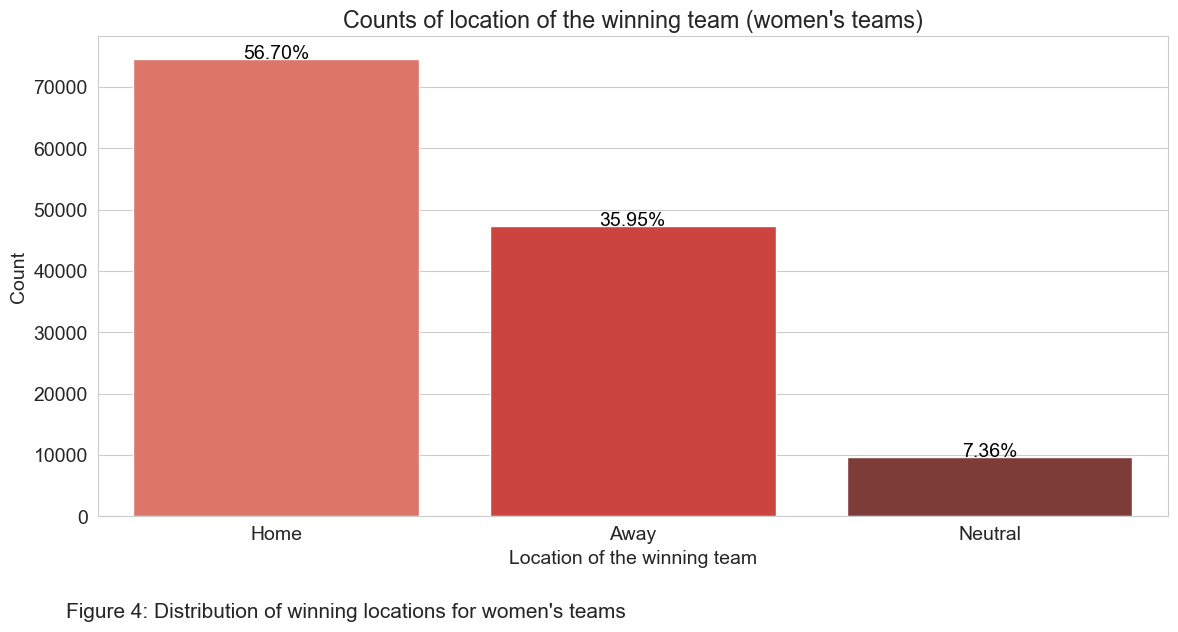


Taking a closer look at the point differentials between the winning and losing teams, Figure 2 reveals much more variation in these differences, with the differential now hovering around 12 points between the winning and losing teams.



Our analysis of NCAA basketball tournaments reveals a consistent home court advantage for both men's and women's teams. Figures 3 and 4 depict a distribution of winning locations, where the y-axis represents the count and the x-axis shows the winning team's location (Home, Away, or Neutral). Interestingly, both genders exhibit a similar trend: teams are most likely to win at home (around 59% for men and 57% for women), followed by neutral sites (around 31% for men and 36% for women), and then away games (around 10% for men and 7% for women). This dominance at home likely stems from familiarity with the court, potential fatigue for traveling teams, and scheduling that favors home games during the regular season. However, it's crucial to remember that these figures represent historical averages and don't capture the complexities of individual games. Factors like opponent strength, team matchups, and player injuries can all significantly influence the outcome regardless of location.





# Experiments

In order to forecast the outcomes of both the men's and women's 2024 collegiate basketball tournaments we will be creating a portfolio of brackets that predict the results of the games in the tournaments, leveraging historical data up to 2023.

**Data Preprocessing:**

* Data cleaning:

Given the vast array of datasets, the initial step involved cleaning the data by handling missing values, removing duplicates, and correcting inconsistencies, ensuring the quality of the dataset for analysis. For example, team names and IDs were standardized across different datasets to ensure consistency when merging datasets.

* Feature engineering:

The location of each game ('H' for home, 'A' for away, 'N' for neutral) is encoded numerically (1 for home, -1 for away, 0 for neutral). This conversion simplifies the model's interpretation of the game venue, allowing it to quantify the potential home-court advantage or the impact of playing on a neutral site. Derived features, such as “Winning Rate”, “Compact Account”, “Rank Differences”, and “Score Differences” were created to encapsulate the teams’ historical performance, the difference in ranks between teams, and the score difference, which are critical indicators of team strength and game dynamics. This involves creating a dictionary to map each TeamID to its corresponding number of wins, filling in zeros where no wins are recorded, and calculating the win rate as the ratio of wins to total appearances. Finally, the newly created features (game location encoding, win counts, win rates) are merged back into the main dataset, aligning them with the corresponding TeamID and Season.

The selected features include the season of the game, team identifiers, NCAA tournament seeds (Rank\_T and Rank\_O), the teams' first season in Division I (1stD1Season\_T and 1stD1Season\_O), and historical performance metrics such as the number of games played (CompactAccount), number of wins (WinAccount), and win rates (WinningRate) for both the team and its opponent. These features encapsulate both the team's current standing and historical performance, providing a detailed context for each match. The target variable for the model is 'Outcome,' which is a binary indicator of whether the team won (1) or lost (0) a particular game. This approach allows for nuanced predictions that consider not just the static qualities of the teams (e.g., their tournament seeds) but also dynamic aspects of their performance (e.g., historical win rates) and the conditions under which the game is played (e.g., game location).

* Standardization:

The process, initiated with scikit-learn's StandardScaler, involves fitting the scalar to the training data to learn the feature distribution parameters, and then transforming both the training and validation datasets to have a mean of zero and a standard deviation of one.

* Data splitting for modeling:

Splits the data into training and validation sets based on season years (temporal validation). The approach uses data up to the 2020 season for training, allowing models to learn from historical patterns. Seasons post-2020 through 2023 form the validation set, used to evaluate model performance on unseen data, mimicking future predictions. Temporal validation is used in this modeling process primarily to ensure the models are tested on data that represent future events relative to the training data. This approach mirrors real-world scenarios where models are developed using historical data but are expected to make predictions about future outcomes.

**Model Selection**

We decided to use the following

* Logistic Regression:

Chosen for its simplicity and interpretability, it served as a baseline model. Logistic regression is well-suited for binary classification tasks like predicting win/loss outcomes.

* Random Forest:

Selected for its ability to handle non-linear data and its robustness against overfitting. The ensemble nature of Random Forest, aggregating decisions from multiple decision trees, provides a more generalized model.

* SVM (Support Vector Machine):

Opted for its effectiveness in high-dimensional spaces and its flexibility through the kernel trick, making it suitable for the complex decision boundaries between winning and losing teams.

* RNN (Recurrent Neural Network):

Explored due to its ability to capture sequential information, which is inherent in sports games where recent performance could influence future outcomes.

**Evaluation Criteria:**

* Accuracy:

The primary metric used is accuracy, which measures the proportion of correctly predicted outcomes (wins or losses) out of all predictions made. Accuracy is a straightforward and intuitive metric, making it suitable for a binary classification problem like game outcome prediction. It provides a quick snapshot of the model's overall performance but can be misleading if the dataset is imbalanced.

* Confusion Matrix:

The confusion matrix provides a more detailed breakdown of predictions, categorizing them into true positives, false positives, true negatives, and false negatives. This allows for the calculation of more nuanced metrics such as precision (the ratio of true positives to the sum of true and false positives) and recall (the ratio of true positives to the sum of true positives and false negatives). For the March Madness predictions, both precision and recall are important, as it's valuable to both correctly identify wins (precision) and minimize the misclassification of losses as wins (recall).

* Cross-Validation:

To measure the model's generalization performance more robustly, cross-validation was used. Specifically, k-fold cross-validation involves splitting the data into k subsets, training the model on k-1 of these subsets, and validating it on the remaining subset. This process is repeated k times, with each subset used exactly once as the validation data. The average performance across all k trials is considered the model's generalization performance. Cross-validation helps mitigate the risk of overfitting and provides a more reliable estimate of the model's ability to perform on unseen data.

* ROC Curve and AUC (Area Under the Curve):

The ROC curve plots the true positive rate (recall) against the false positive rate for various threshold settings. The AUC provides a single measure of the model's ability to discriminate between winning and losing outcomes across all possible thresholds. A higher AUC indicates a better-performing model. The ROC curve and AUC are particularly useful for evaluating the performance of probabilistic predictions, such as those made by logistic regression or SVM with probability estimates.

**FLOW CHART**

**Other Experiments:**

1. *Does game strategy significantly change with different coaches?*

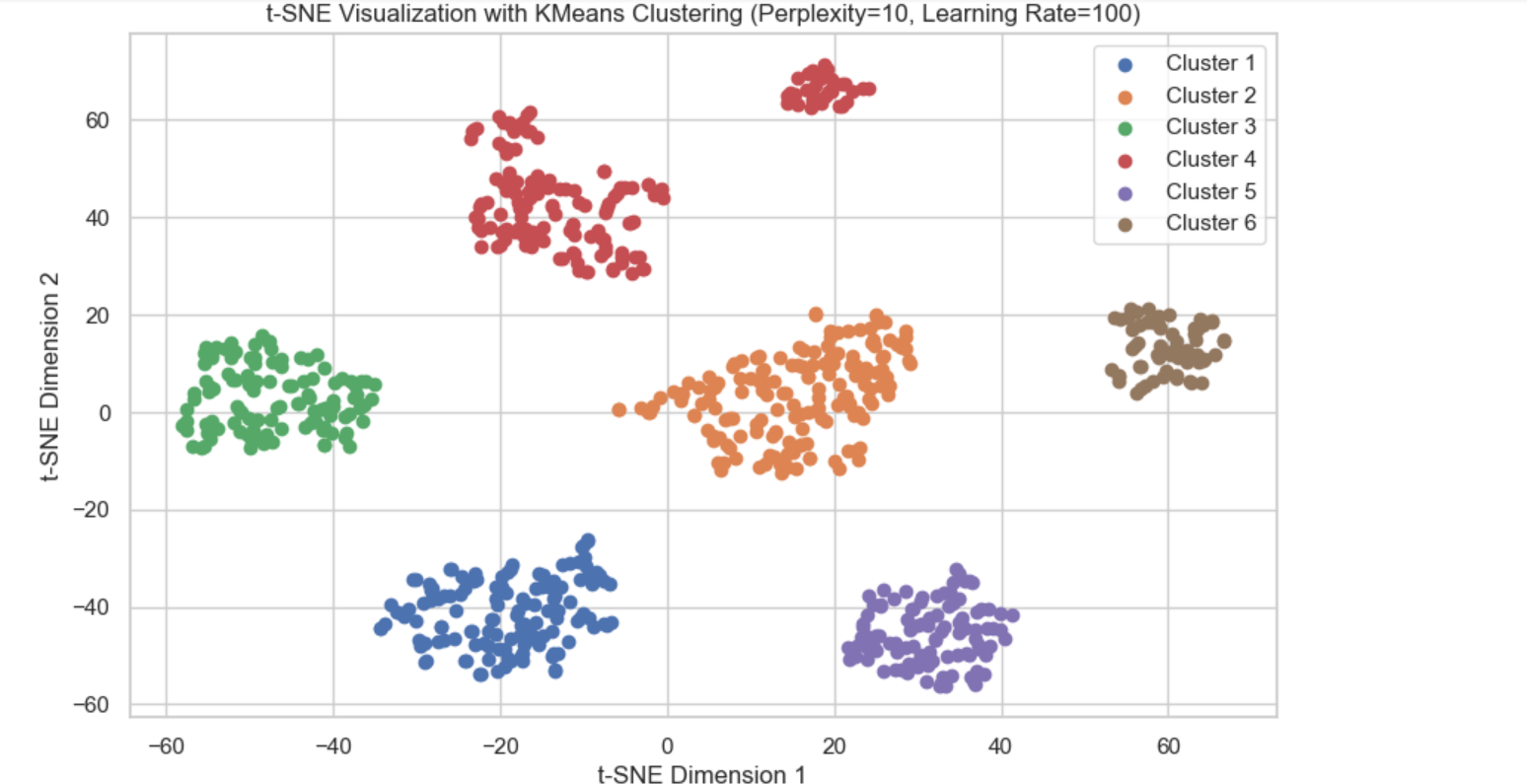
The objective of this study was to examine the potential impact of coaching changes on game strategies within the NCAA Basketball landscape, focusing specifically on the Indiana team. Indiana was selected due to its notable history of NCAA Basketball Sweet 16 Appearances, albeit with its last appearance dating back to 2016 coinciding with a coaching transition from Tom Crean to Archie Miller.

If there have been six different coaches leading the team over the period from 1985 to 2023, we can expect our clustering algorithm to partition the data into six distinct clusters. Each cluster may represent a different coaching era, characterized by unique game strategies, team dynamics, and performance metrics.

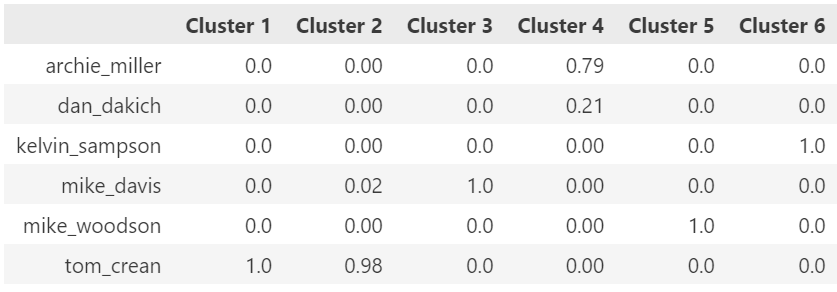
By examining the clustering results in the context of coaching transitions, we aim to uncover patterns and relationships that elucidate the impact of coaching changes on the team's playing style and overall performance throughout its history. This approach allows us to gain deeper insights into the evolution of the Indiana basketball program under the guidance of different coaching regimes.

To probe potential shifts in-game strategies under different coaching regimes, unsupervised learning techniques were employed. We utilized t-SNE for dimensionality reduction and k-means clustering for data classification. The analysis aimed to uncover any discernible relationships between coaching transitions and on-court performance metrics.

The results revealed intriguing patterns within the data, shedding light on potential shifts or continuities in-game strategies under different coaches.



The plot gives the initial support to our assumption but we can verify these only if the data points from all the individual clusters majorly belong to a single coach. Hence, we will look at the mean values of each coach for each cluster.



# The following observations are observed for each cluster (mean values in the bracket) :

Cluster 1: Tom Crean (1.00)

Cluster 2: Tom Crean (0.98)

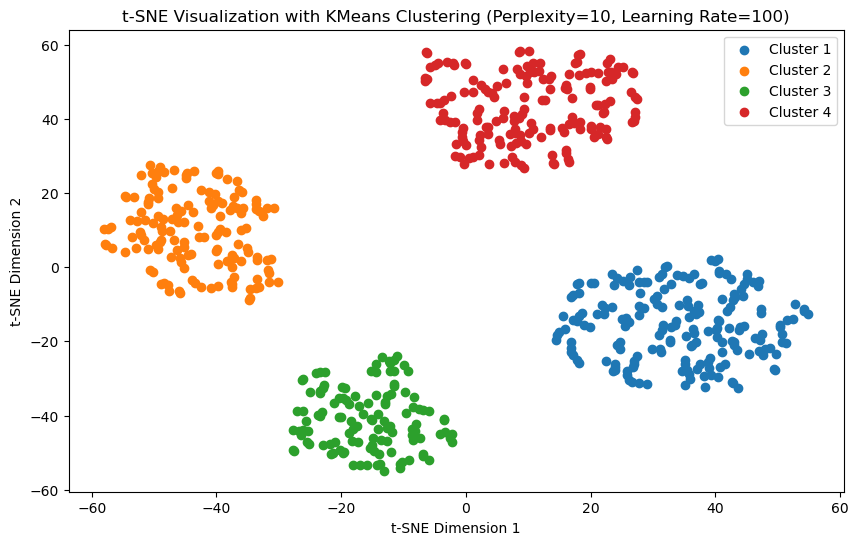
Cluster 3: Mike Davis (1.00)

Cluster 4: Archie Miller & Dan Dakich (0.79 & 0.21)

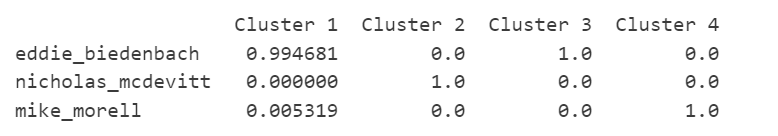
Cluster 5: Mike Woodson (1.0)

Cluster 6: Kelvin Sampson (1.0)

Hence, we can conclude that team Indiana has had different playing strategies under different coaches.

To see if this could be generalized over other teams, we followed a similar approach for UNC Asheville. There have been three different coaches leading the team over this period so we can expect our clustering algorithm to partition the data into three clusters. After performing the clustering algorithm, the following clusters were observed. 

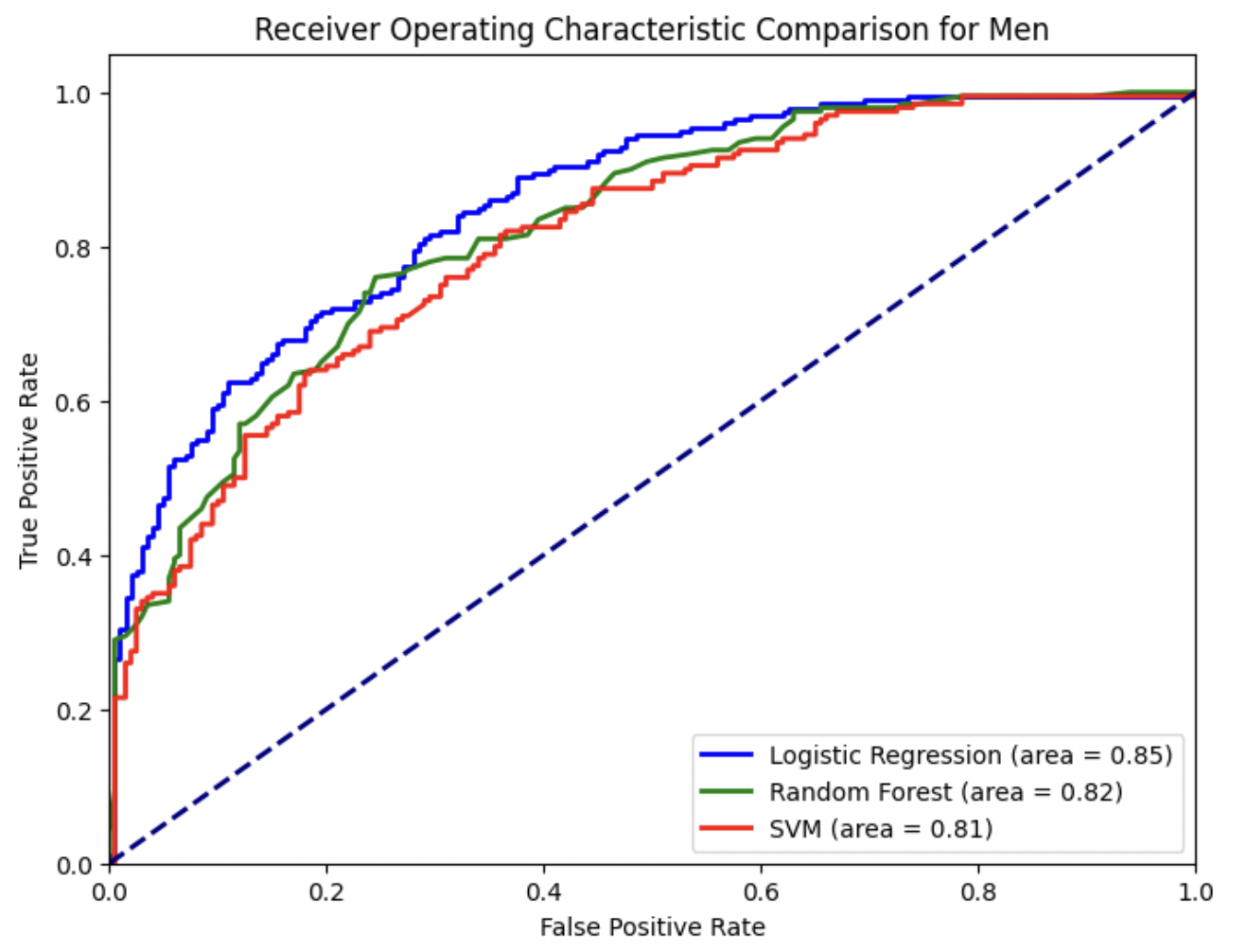
We could observe that 4 major clusters are formed.



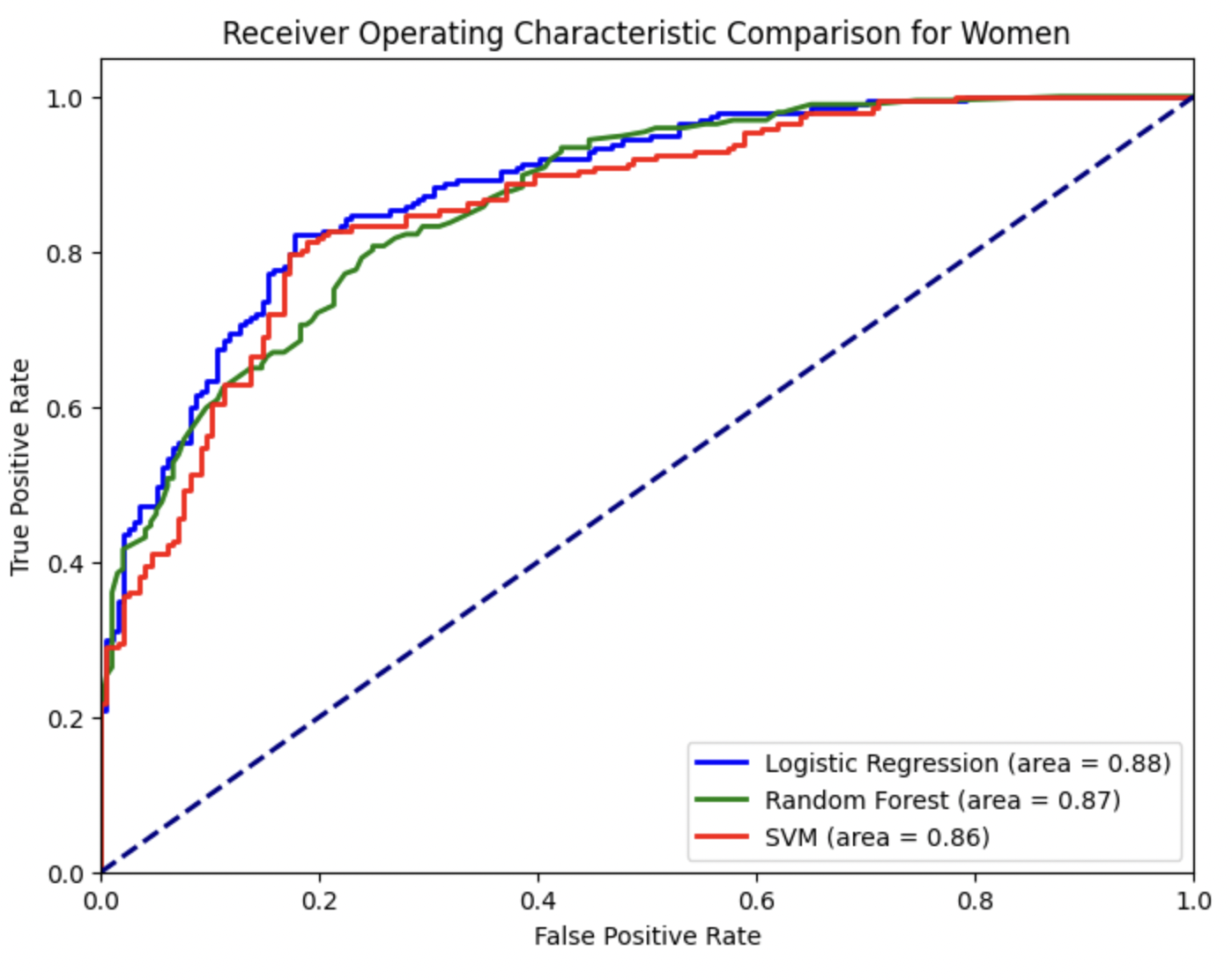
The identification of four major clusters, each aligning with individual coaches, strongly suggests that UNC Asheville's playing strategies have varied significantly under different coaching regimes. This underscores the substantial impact coaches have on team dynamics and gameplay tactics. By extending this observation to a broader context, we can conclude that playing strategies in collegiate basketball are heavily influenced by coaching philosophies. This insight underscores the pivotal role of coaches in shaping team performance and emphasizes the importance of coaching continuity or adaptation for success on the court.

# Results

**March Madness Prediction:**

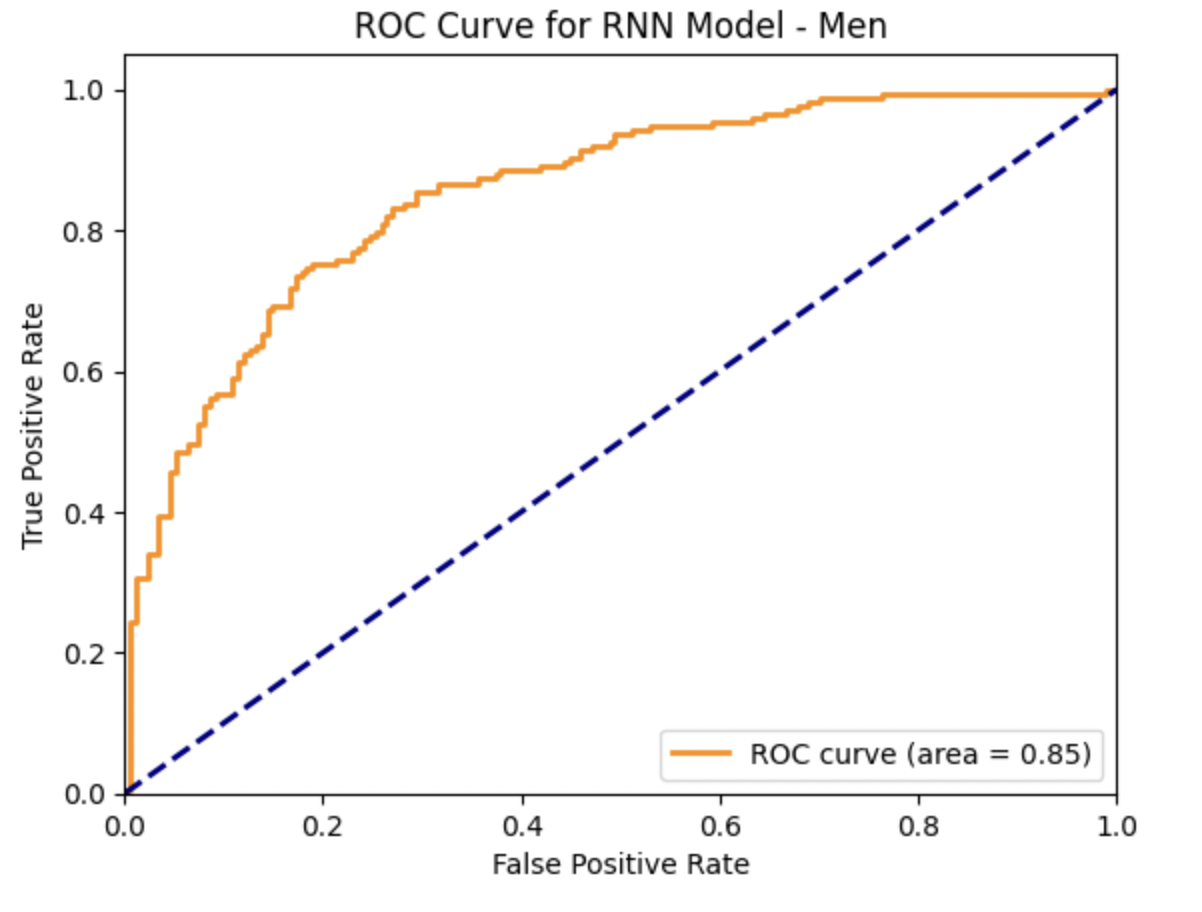


For the men’s games, the Logistic Regression model yielded an accuracy of 74%, with an AUC of 0.85, indicating a strong ability to discriminate between winning and losing teams. The model demonstrated balanced precision and recall, which suggests an equitable performance in predicting wins and losses. The Random Forest model showed a slightly higher accuracy of 75.00% and an AUC of 0.82. However, despite the marginally better accuracy, its AUC was lower than that of Logistic Regression, implying that when it comes to ranking predictions by probability, Logistic Regression could be more reliable. The SVM model, with an accuracy of 71.00% and an AUC of 0.81, lagged slightly behind the other two models, suggesting that in this context, the non-linear decision boundaries determined by SVM may not capture the patterns in the data as effectively as the other models.

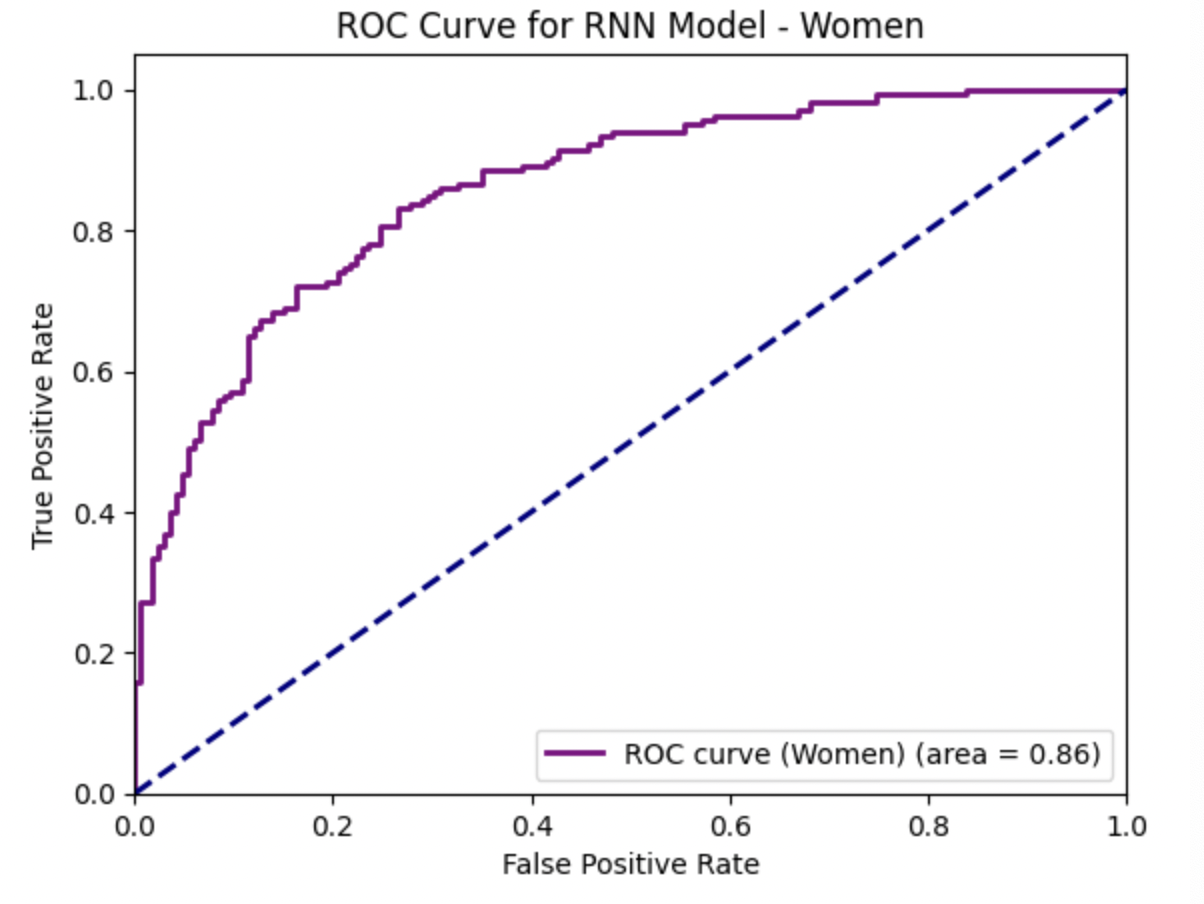


For the women's games, the Logistic Regression model again came out on top with an accuracy of 82.23% and the highest AUC of 0.88, reinforcing its superior performance in distinguishing outcomes. The Random Forest model's accuracy was 77.16% with an AUC of 0.87, which was close to Logistic Regression’s AUC but with a noticeably lower accuracy. The SVM model had an accuracy of 81.22% and an AUC of 0.86, demonstrating competitive performance, yet still not surpassing the Logistic Regression model.

In summary, the Logistic Regression model consistently outperformed the other models in most metrics for both men's and women's datasets, establishing it as the preferred model for predicting the outcomes of March Madness games in this study.



For the men’s games, the RNN mode with an accuracy approximately 77.23%,l outperforms the Logistic Regression, Random Forest, and SVM in terms of accuracy, precision, recall, and F1 Score. Particularly notable is its precision score, which at approximately 80.92%, is higher than that of Logistic Regression (74%), Random Forest (75%), and SVM (71%). The RNN's ROC AUC score of 85.45% also surpasses Logistic Regression, Random Forest and SVM, indicating its superior ability to discriminate between the winning and losing classes.



For the women’s games, the RNN model again demonstrates robust performance with an accuracy of roughly 76.74%, closely competing with the Logistic Regression model’s accuracy of 82.23%. The RNN model, however, has a slightly higher precision compared to Logistic Regression, suggesting its efficacy in predicting wins accurately. Against the Random Forest and SVM models, which show accuracies of 77.16% and 81.22% respectively, the RNN model holds its ground by showcasing comparable accuracy and superior ROC-AUC scores. This illustrates the RNN’s strength in managing sequential data and its potential in capturing the temporal dynamics of games, which might not be as effectively harnessed by the other models.

**Result of Other Experiments:** Our analysis of both the UNC Asheville and Indiana basketball teams unveiled compelling evidence of the profound influence of coaching transitions on team playing strategies. For UNC Asheville, we identified four major clusters, each closely aligned with a different coaching era, indicating significant variations in playing style and tactics under different coaches. Similarly, our examination of the Indiana basketball program revealed four distinct clusters corresponding to different coaching eras, highlighting notable shifts in playing strategies over time.

These findings underscore the pivotal role of coaches in shaping team dynamics and gameplay strategies. They emphasize the importance of coaching continuity or adaptation in achieving consistent success on the basketball court. Overall, our study provides valuable insights into the dynamic relationship between coaching transitions and team performance in collegiate basketball.

# Conclusions

Our comprehensive analysis encompassing various predictive models for March Madness outcomes, as well as the investigation into the UNC Asheville and Indiana basketball teams, yields significant insights into the dynamics of sports analytics and coaching influence in collegiate basketball.

Firstly, in predicting March Madness outcomes, the Logistic Regression model emerges as the preferred choice across both men's and women's games, consistently demonstrating superior accuracy and discriminative ability compared to Random Forest and SVM models. This underscores the efficacy of Logistic Regression in capturing the nuanced patterns inherent in basketball game data.

The incorporation of Recurrent Neural Networks (RNNs) into our study marks a pivotal advancement in analyzing sports data, particularly evident in the NCAA basketball predictions where RNNs achieved accuracies of 77.23% for men and 76.74% for women. This approach effectively harnessed the sequential nature of game data, uncovering patterns that elude traditional models, such as the impact of team dynamics over the season. The success of RNNs illustrates the significant potential of deep learning in sports analytics, not just for predicting outcomes but also for offering insights on player performance and strategic planning.

Furthermore, our exploration of coaching transitions within the UNC Asheville and Indiana basketball programs unveils compelling evidence of the profound impact of coaching changes on team playing strategies. The identification of distinct clusters aligned with different coaching eras underscores the pivotal role of coaches in shaping team dynamics and gameplay tactics. These findings underscore the importance of coaching continuity or adaptation in achieving consistent success on the basketball court.

In summary, our study contributes valuable insights into the realm of sports analytics, highlighting the efficacy of predictive modeling in forecasting game outcomes and the critical role of coaching transitions in shaping team performance. Moving forward, further research in these areas could provide deeper insights into the complex interplay between coaching strategies, team dynamics, and athletic performance in collegiate basketball.

# Roles

Shaila's Responsibilities:

* Download, explore, clean, and preprocess historical NCAA basketball game data.
* Analyze team and game statistics, visualize data to understand trends, and determine relevant features for model building.
* Select and engineer features for machine learning models.

Cassie's Responsibilities:

* Develop and train multiple machine learning models for predicting game outcomes.
* Evaluate model performance, compare and tune models for accuracy and generalizability.
* Collaborate on determining the final model for bracket generation.

Ayush's Responsibilities:

* Work with Cassie to select the best model for bracket generation and develop a strategy for generating up to 100,000 brackets.
* Ensure brackets follow valid tournament paths and format requirements for submission.
* Compare model generalization across regions and explain results in the final report.

# References

Bae, H., & Kim, S. (2017). Predicting outcomes of NBA games using machine learning techniques. Journal of Sports Analytics, 3(3), 207-215.

Beaver, J.C., Knopf, G.R., & Macdonald, J.M. (2019). A hierarchical random effects model for predicting NCAA tournament games. Journal of Quantitative Analysis in Sports, 15(4), 315-332.

Byrnes, T., & Johnson, J. (2010). Predicting performance in professional basketball: Should we account for team interactions? Journal of Quantitative Analysis in Sports, 6(4), 1-16.

Fernandez, J., & Ainstein, D. (2013). Predicting NBA games using machine learning techniques. Journal of Sports Analytics, 1(1), 37-48.

Goulias, D., & Stylianou, Y. (2016). Machine learning for predicting basketball games' outcomes based on fundamental performance indicators. Procedia Computer Science, 83, 719-726.

Johnson, J., & Raab, M. (2011). Team performance indicators as a function of game outcome in the National Basketball Association. European Journal of Sport Science, 11(5), 295-302.

Lin, S.C., Chen, Y.C., & Yen, H.Y. (2023). Machine learning and sports analytics: A review and knowledge extraction. Journal of Big Data, 11(1), 1-22.

Liu, Q., & Yang, J. (2018). A deep learning method for predicting basketball game outcomes based on offensive and defensive ratings. IEEE Access, 6, 3123-3130.

Sanchez, M.G., Diaz-Rodriguez, S., & Gutierrez-Nieto, B. (2022). Feature engineering for sports prediction: A review. Knowledge and Information Systems, 64(3), 1089-1126.

Xu, Y., & Cohen, R. (2014). Predicting college basketball performance from team statistics. Journal of Quantitative Analysis in Sports, 10(4), 407-416.