MULTIVARIATE ANALYSIS

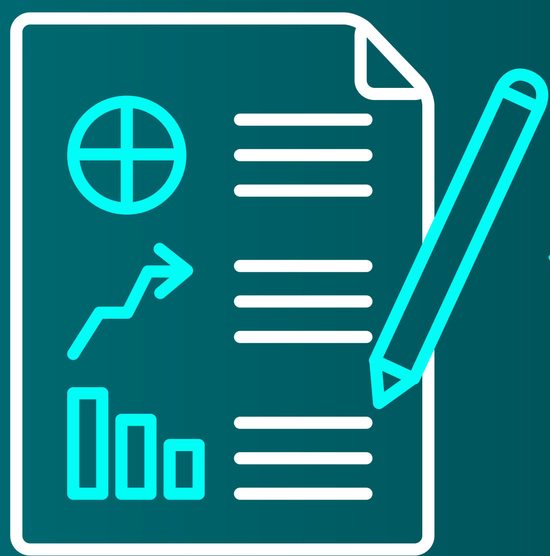
CLASSIFICATION STUDY OF NBA PLAYER PROFILES

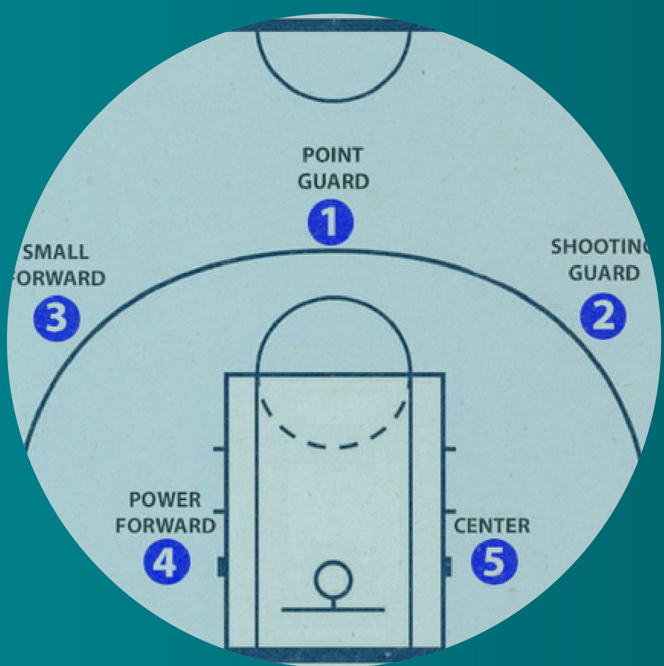
Renato Vivar Orellana

Data Science Engineer

WHY THIS STUDY?

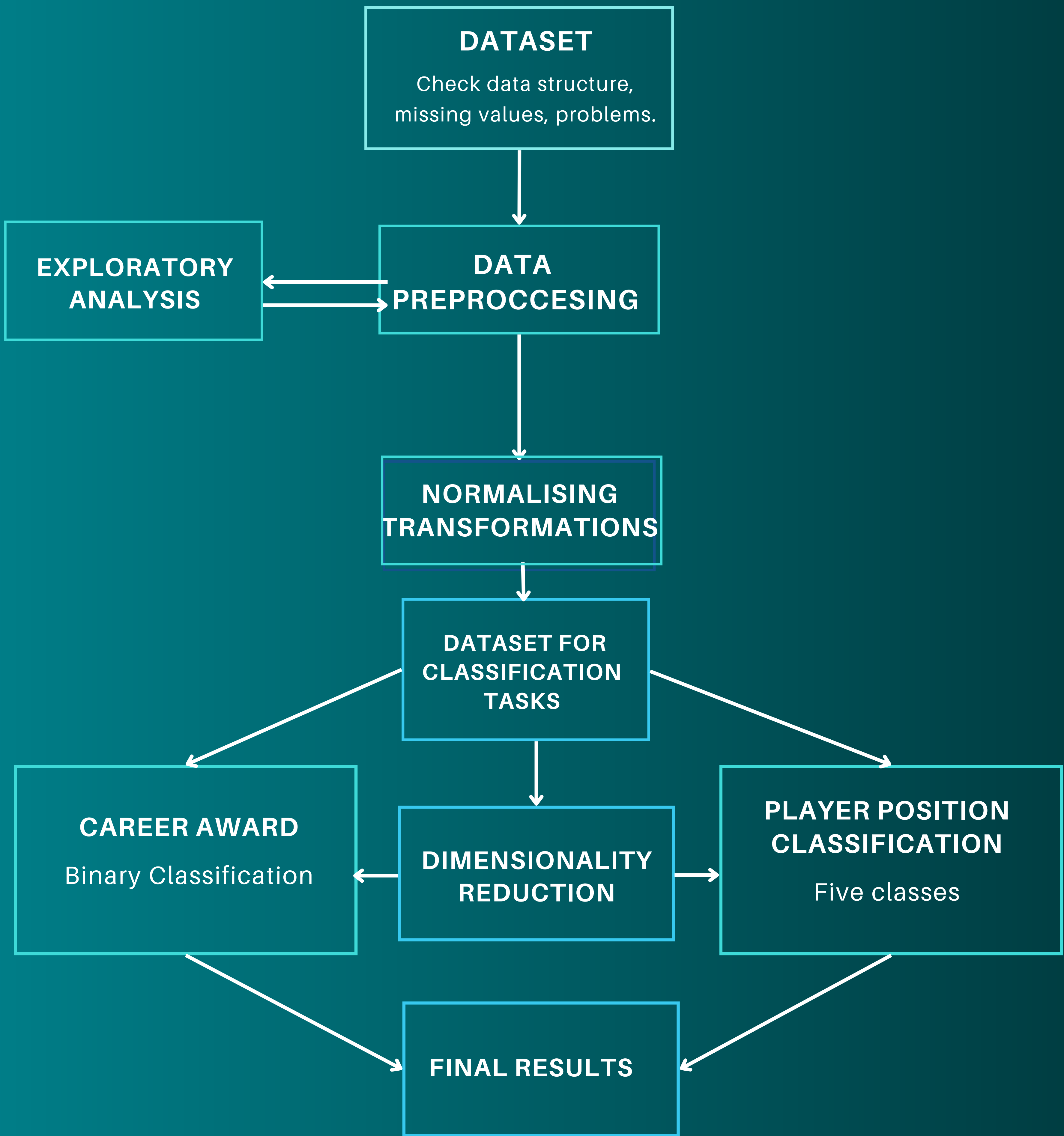
- Understand what performance leads to major NBA recognition
- Classify player positions using career stats
- Use interpretable statistical models





Renato Vivar Orellana
Data Science Engineer

METHODOLOGY

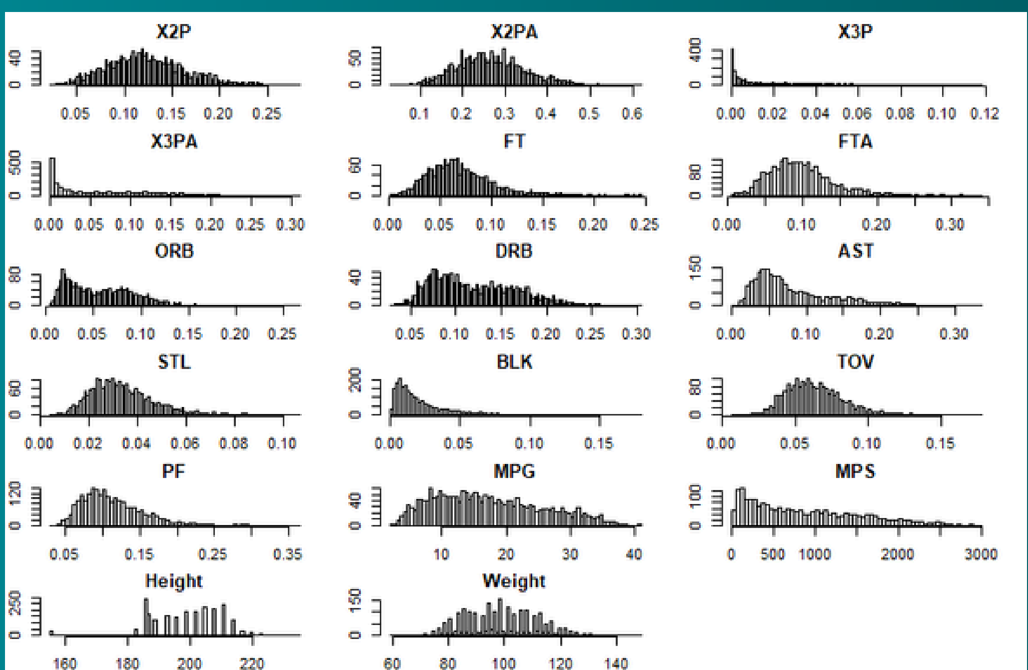


Renato Vivar Orellana
Data Science Engineer

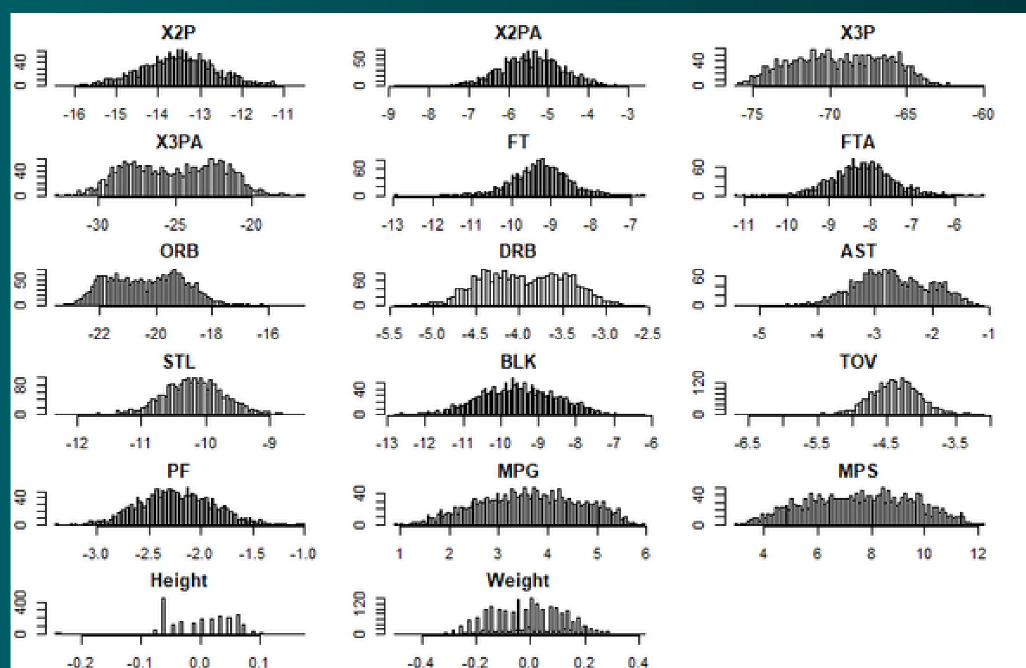
PREPROCESSING STEPS AND FEATURE TRANSFORMATION

- Aggregate per-player lifetime stats
- Drop players with <60 minutes or <10 games
- Normalize using per-minute stats (e.g., X2P/MP)

- Applied log-like transform to improve normality
- Constructed features: MPG, MPS, action rates
- Dropped categorical or poorly-behaved features



Before Transformation



After transformation

Renato Vivar Orellana

Data Science Engineer

WHY DO WE USE THE TRANSFORMATIONS?

- To make the predictors better resemble a multivariate normal (MVN) distribution — a key assumption for methods like LDA and QDA.

WHAT TRANSFORMATION WAS APPLIED?

- A log-like monotonic transformation of the form:

$$f(x) = c(x-1) + log(x)$$

- ***log(x)***: Reduces positive skewness (common in count-based sports stats).
- ***c(x−1)***: Adds a linear scaling factor to adjust for the level of skewness in each variable.
- ***c*** is not fixed: It was tuned per feature by minimizing the Anderson-Darling (AD) test statistic (a test for normality).
- The transformation ensures positivity (by adding 1 to counts), which is important before applying log.

WHY NOT JUST USE LOG(X)?

- Because not all features had the same skewness — a fixed transformation would not equally improve all distributions. So, a data-adaptive approach was used:
- c ∈ (0,10^5] was optimized via Brent's method (a global minimization algorithm).
- Some features (e.g., X2P, STL) passed the AD test after transformation.

WHAT ABOUT HEIGHT AND WEIGHT?

- These were transformed with log(x/x), effectively centering around 0. This removes strict positivity without changing interpretability or affecting results (translation-invariant).

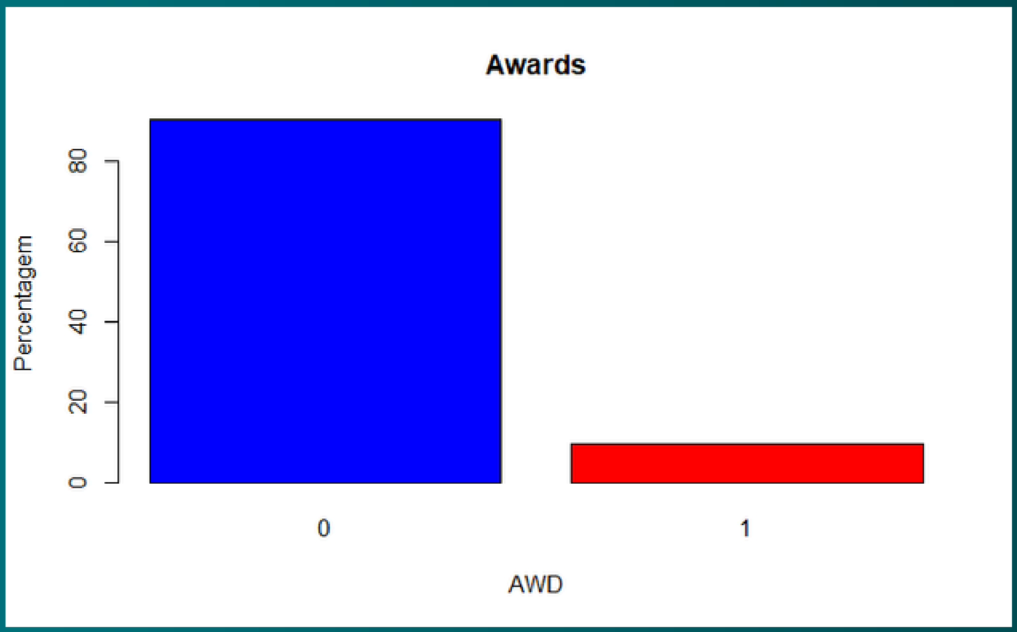
RESULTS OF THE TRANSFORMATIONS

- Features became more symmetric and less skewed.
- PCA retained more variance post-transformation (e.g., 33.7% vs. 31.0% in PC1).
- Improved compatibility with Gaussian-based methods like LDA/QDA.

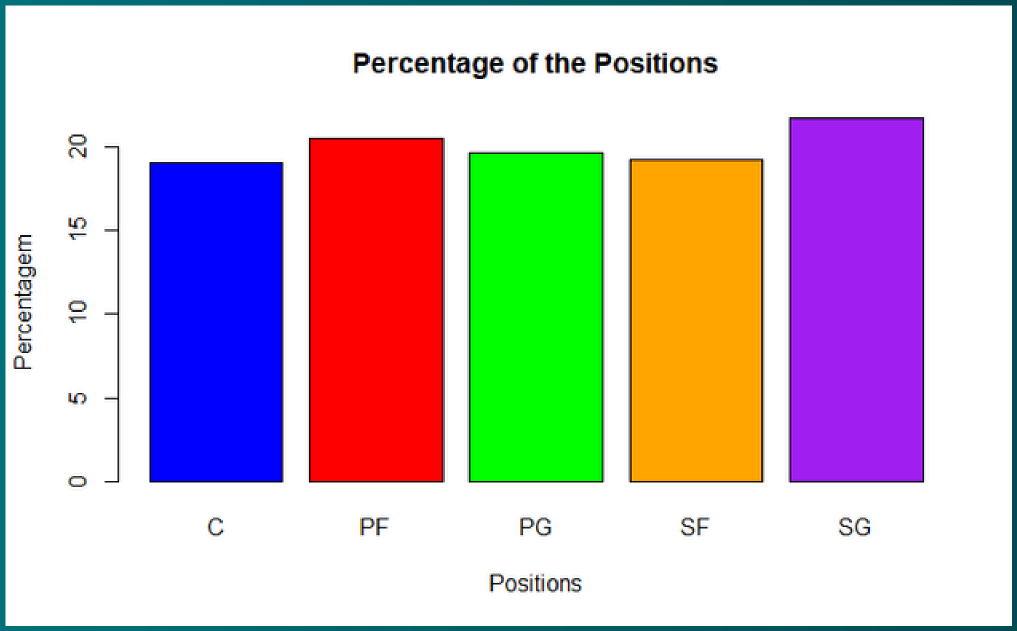
Renato Vivar Orellana
Data Science Engineer

WHAT ARE WE PREDICTING?

- AWD: Award recipient? (Yes/No, 9.74% positive).
- Pos: Player position (C, PF, SF, SG, PG).



Award Proportion (Imbalanced).



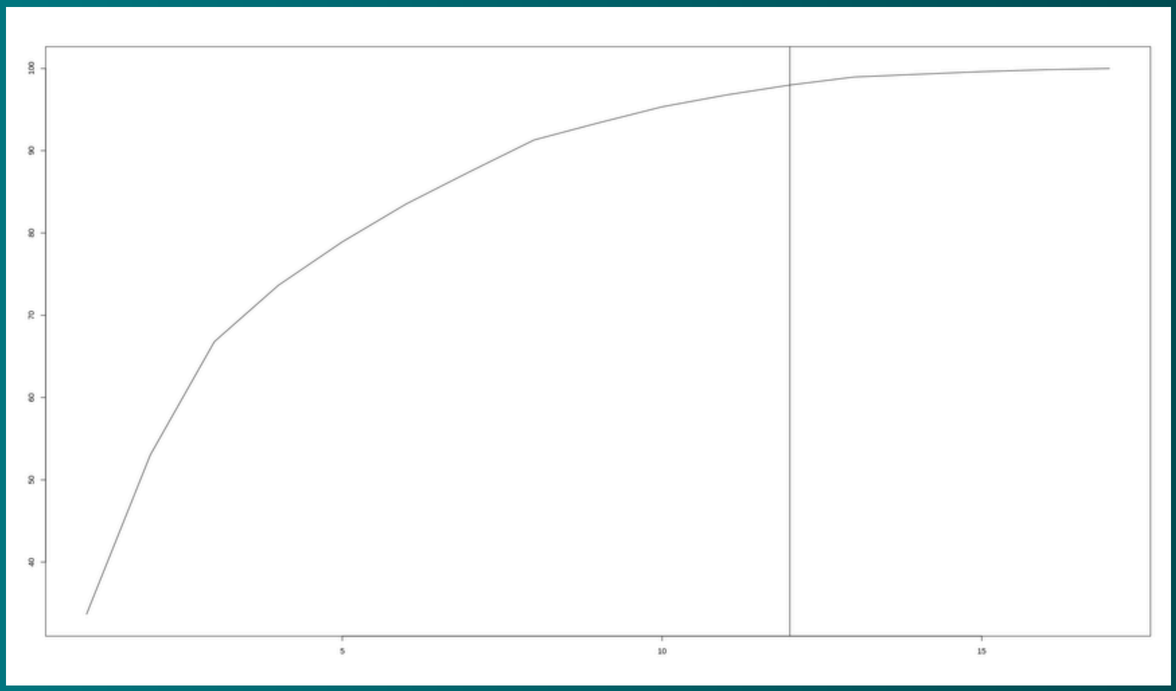
Positions Proportion

Renato Vivar Orellana

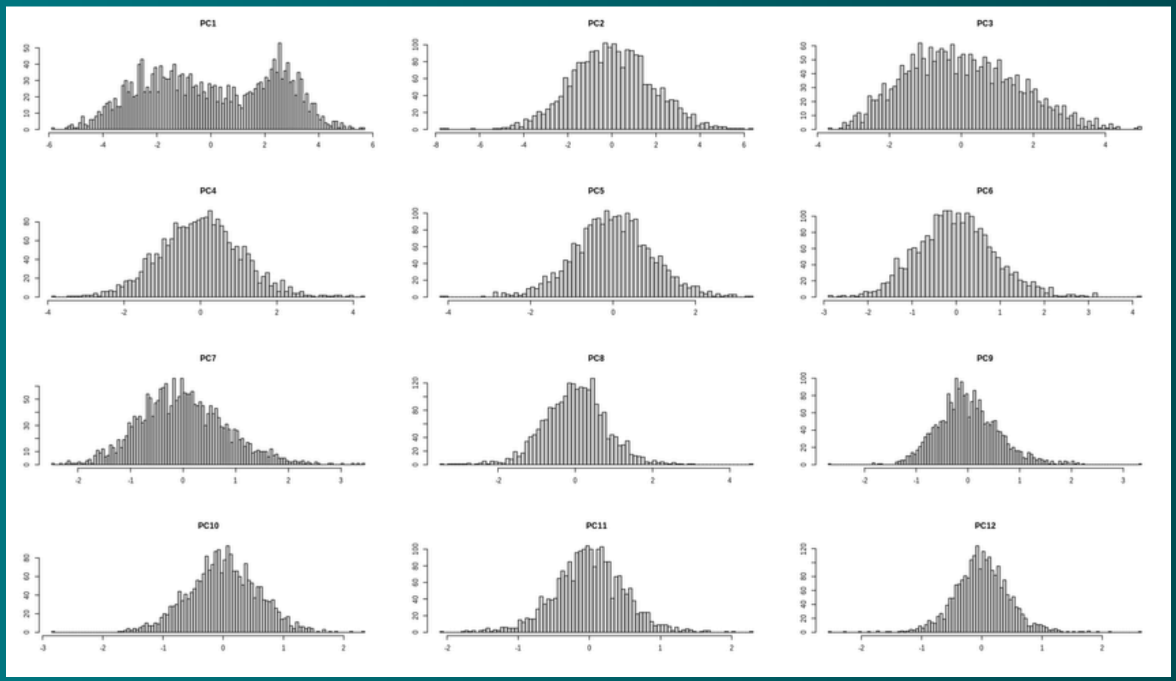
Data Science Engineer

PRINCIPAL COMPONENT ANALYSIS (PCA)

- Reduced to 12 PCs, preserving 98% of variance
- Improved performance and interpretability
- Used in both transformed and untransformed datasets.



Cumulative Var% with v-line at 12 PCs



histograms of the retained PCs.

Renato Vivar Orellana

Data Science Engineer

CLASSIFICATION MODELS USED

LDA – Linear Discriminant Analysis

- Assumes features are normally distributed within each class.
- Assumes equal covariance across classes (linear boundaries).
- Works well when features are continuous and classes are well-separated.
- Best performer in this study.
- Use when interpretability matters.

QDA – Quadratic Discriminant Analysis

- Similar to LDA but allows different covariances per class.
- More flexible boundaries (curved decision surfaces).
- May overfit with small or high-dimensional data.
- Use when classes may have different spreads.

k-NN – k-Nearest Neighbors

- Non-parametric, no training needed.
- Classifies based on the majority class among k closest data points.
- Sensitive to feature scaling and high dimensionality.
- Simple, but suffers with noisy or sparse data.

Naive Bayes

- Assumes all features are conditionally independent given the class.
- Fast and simple, works well with text or categorical data.
- Assumption often violated in real datasets like this one.
- Lightweight, but less accurate with correlated features.

Evaluation:

- All methods were compared using:
- Stratified 10-fold cross-validation
- F1-score, Accuracy, Confusion Matrix

Renato Vivar Orellana
Data Science Engineer

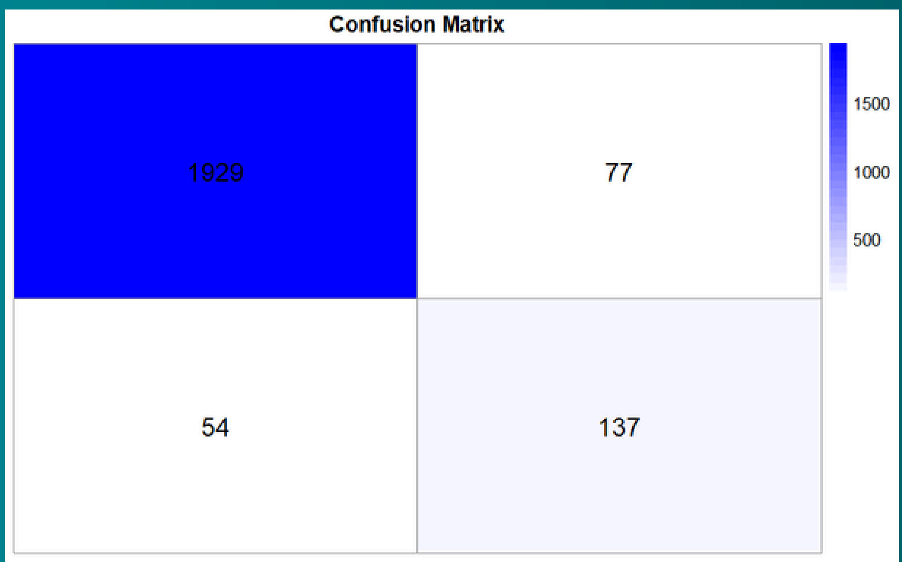
RESULTS: AWARD PREDICTION (AWD)

Metric / Method	LDA	QDA	kNN	Naive Bayes
Accuracy	0.9404	0.8812	0.9294	0.8994
F1-Score	0.8215	0.7501	0.7382	0.7765

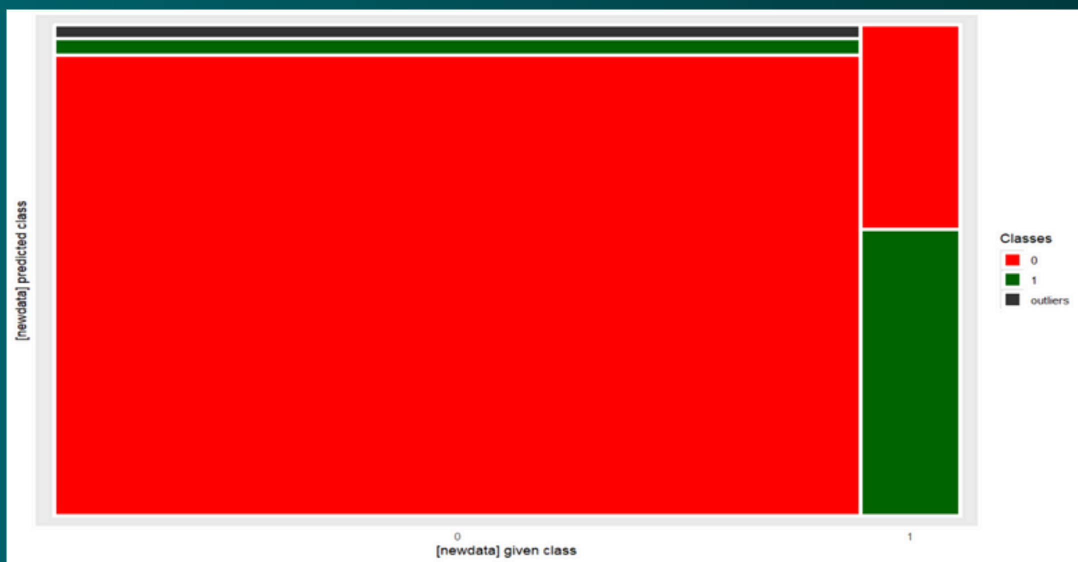
Classifcation Results without Dimension Reduction

Metric / Method	LDA	QDA	kNN	Naive Bayes
Accuracy	0.9335	0.9099	0.9030	0.9281
F1-Score	0.7870	0.7601	0.6975	0.7664

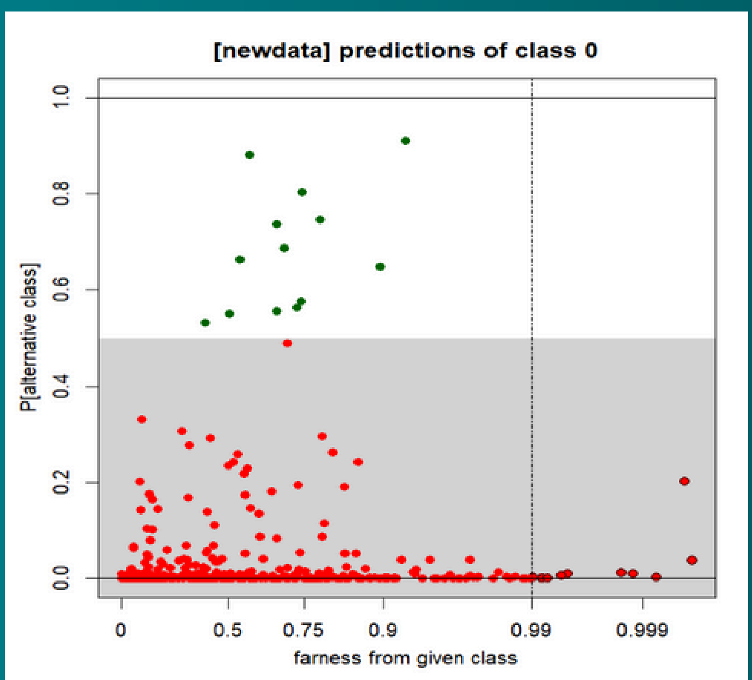
Classifcation Results with PCA



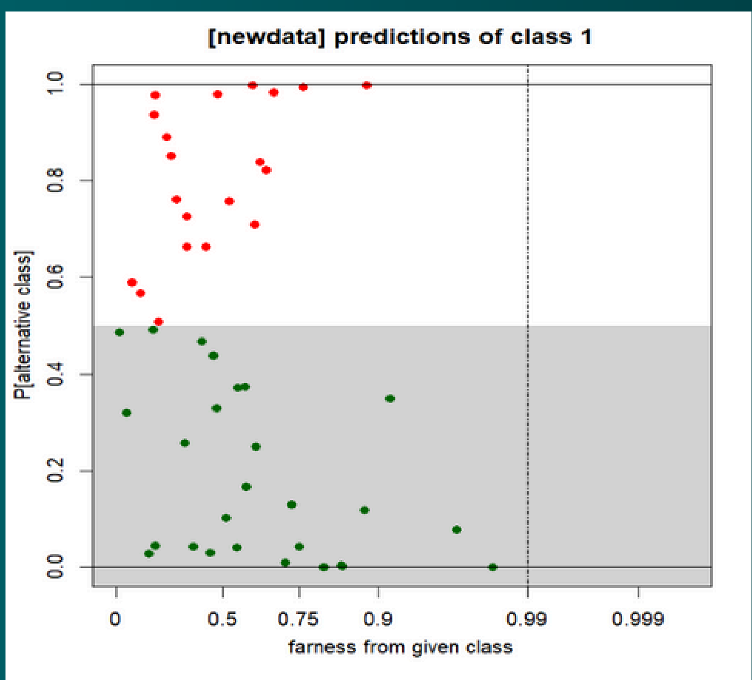
Confusion Matrix for LDA and Full Dataset



Stacked mosaic plot of a classification with LDA and Full Dataset



Classmap for class 0.



Classmap for class 1.

Renato Vivar Orellana

Data Science Engineer

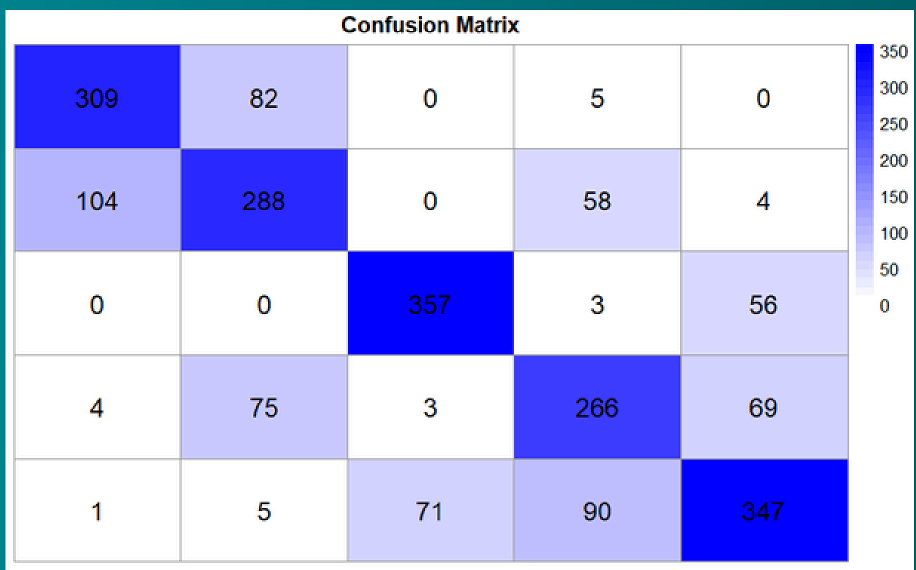
RESULTS: PLAYER POSITION PREDICTION

Metric / Method	LDA	QDA	kNN	Naive Bayes
Accuracy	0.7132	0.6941	0.7069	0.6982
F1-Score	0.7140	0.6886	0.7060	0.6941

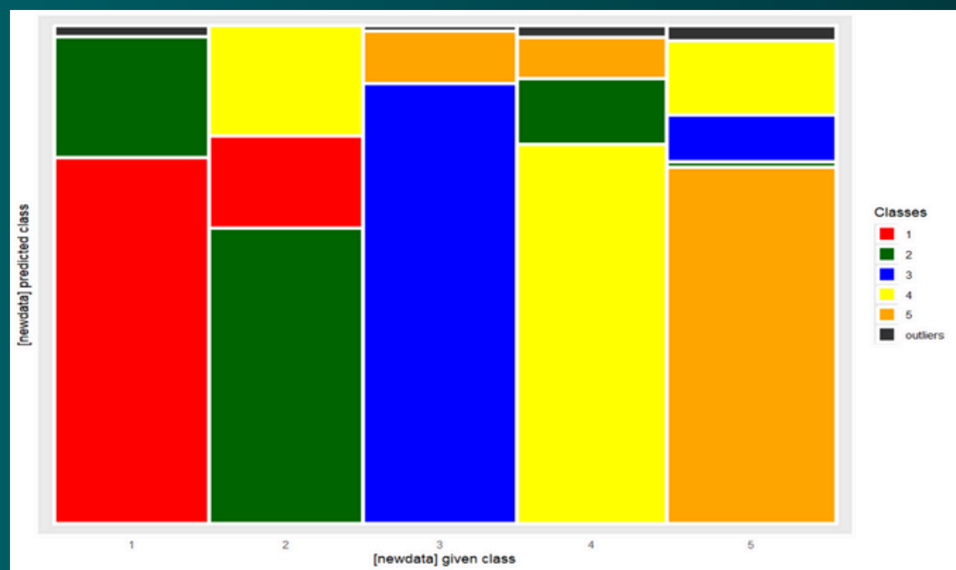
Classifcation Results without Dimension Reduction

Metric / Method	LDA	QDA	kNN	Naive Bayes
Accuracy	0.7069	0.7212	0.6441	0.6468
F1-Score	0.7072	0.721	0.6398	0.6415

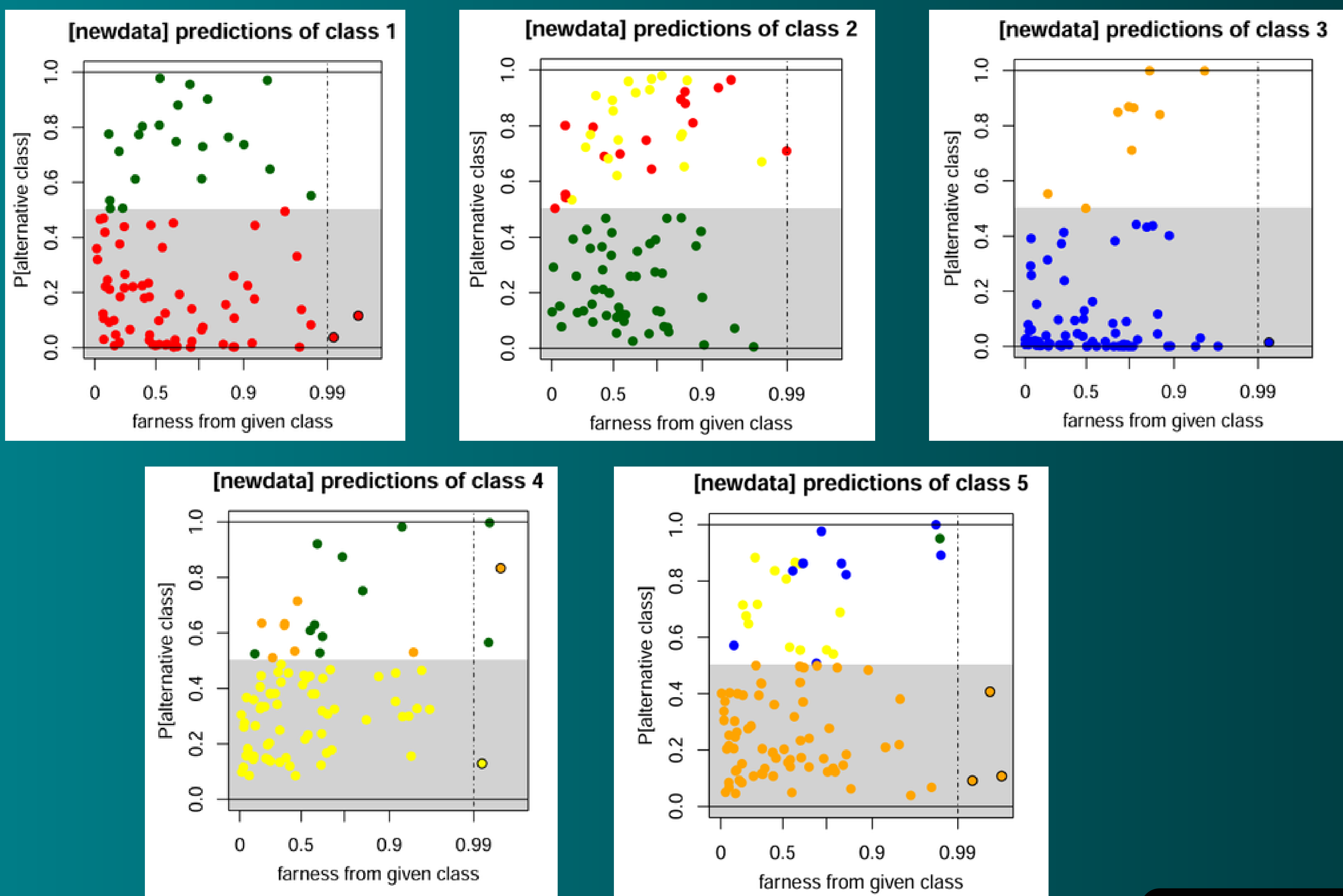
Classifcation Results without Dimension Reduction



Confusion Matrix for LDA and Full Dataset



Stacked mosaic plot of a classification with LDA and Full Dataset



Classmaps

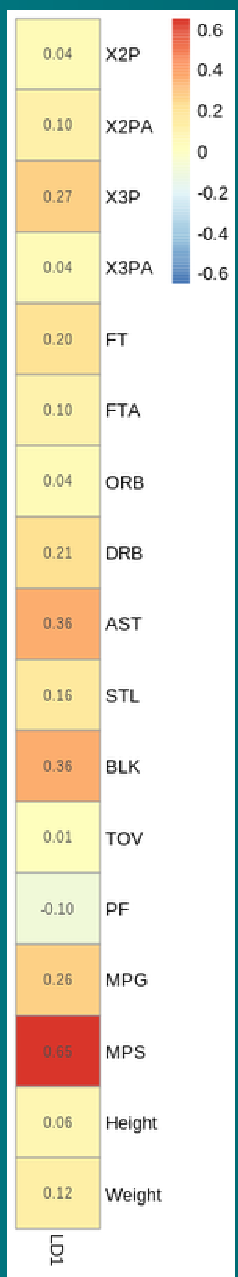
Renato Vivar Orellana

Data Science Engineer

RESULTS: CLASSIFICATION (LDA)

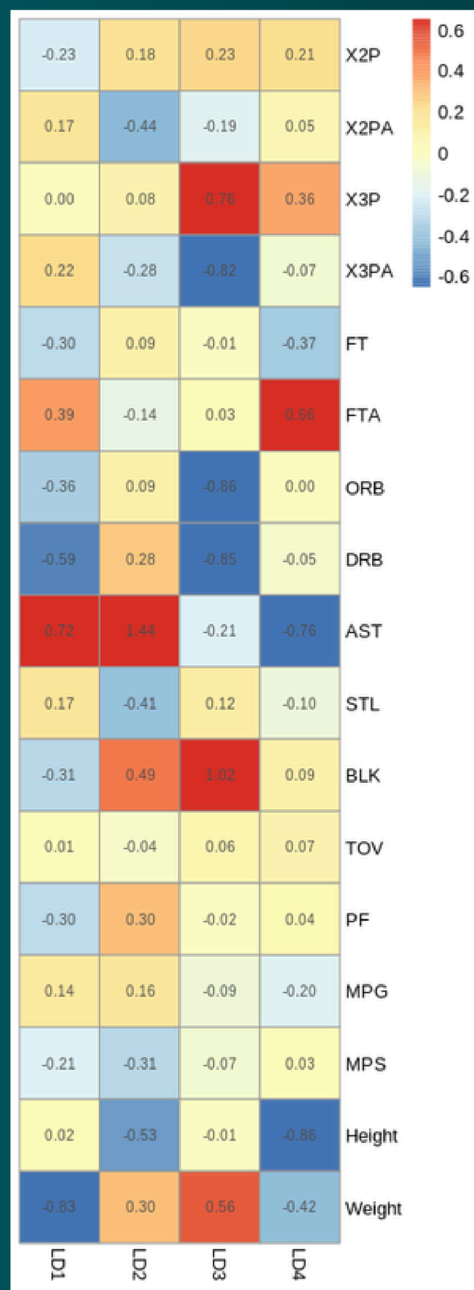
- Importance of playing a significant number of matches and minutes for recognition.
- Higher useful action rates (e.g., assists, rebounds) increase likelihood of receiving awards.
- Anthropometric parameters do not impact award classification

- First linear discriminant highlights assists and rebounds as key classification features.
- Centers and Forwards (SF-PF) typically have fewer assists and more rebounds than guards
- Weight is a significant factor; Centers and Forwards generally weigh more than guards.
- Centers likely incur more fouls, leading to more FT executions due to physical play in tight spaces.



LDA discriminant coefficients Awards

Renato Vivar Orellana
Data Science Engineer



LDA discriminant coefficients Positions

CONCLUSIONS & FUTURE WORK

- Best performance classification methods: LDA and QDA.
- Handling Imbalance: downsampling.
- Features were constructed to better distinguish between aspects of the target classes
- Technical departures from model assumptions were addressed without losing much transparency
- PCA did not improve results but confirmed existence of a reduced efficient subset of feature info
- The data appears sufficient for establishing the relationships of interest despite being basic
- Future work and recommendations: include additional features like advanced statistics (e.g., plus-minus, win shares) and contextual data (e.g., team performance); apply machine learning methods that target prediction performance rather than parsimony

Renato Vivar Orellana
Data Science Engineer