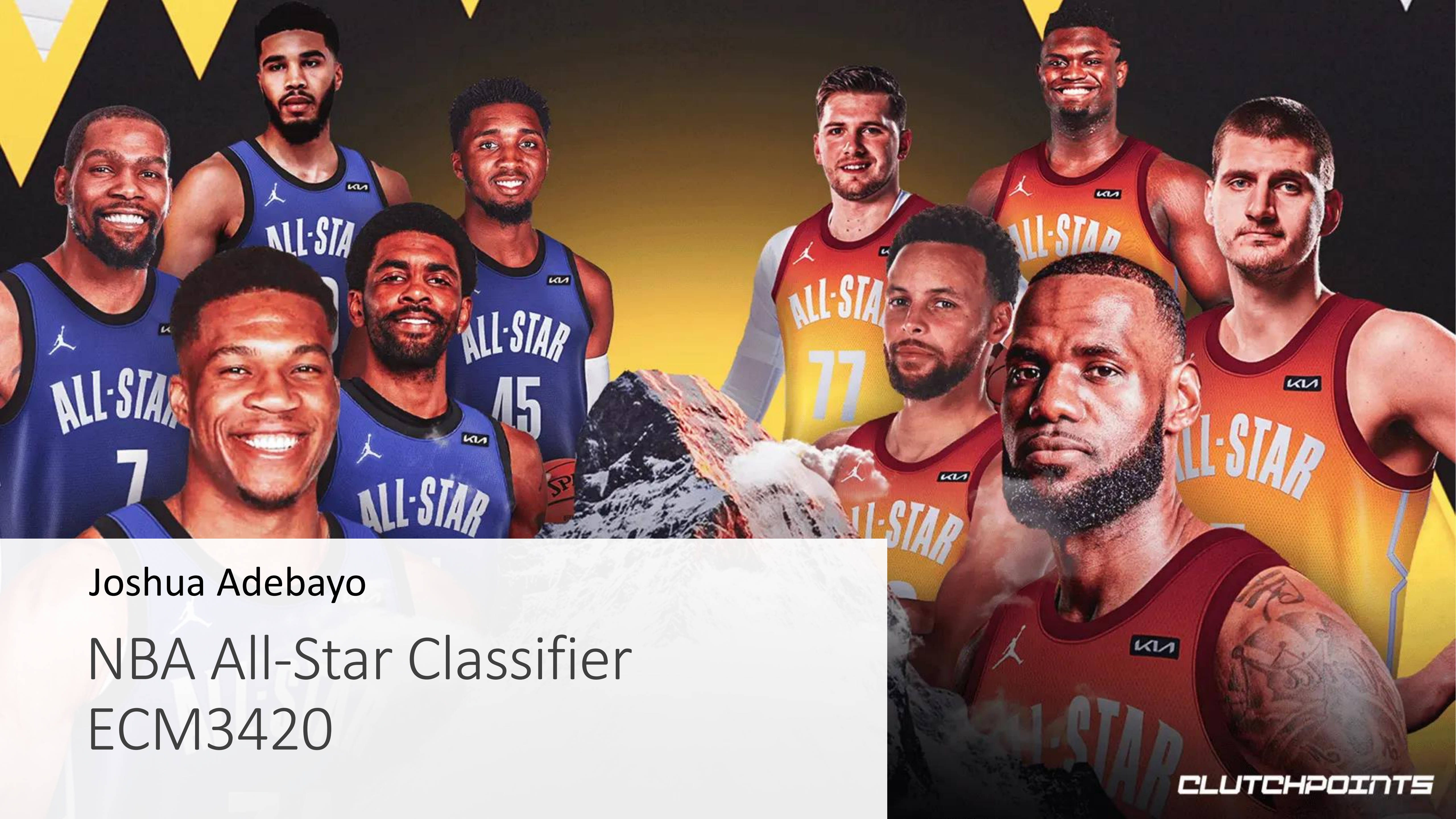


Joshua Adebayo

NBA All-Star Classifier
ECM3420

Structure

Intro to the All –Star game

The Dataset

ML techniques

Results

Conclusion

NBA All Star Game

Since 1951 , the NBA hosts an All-Star at the midpoint of the competitive season.

~24 players

A-S Game selection

1951 - 1980:

Media voting (100%)

1980 - 2017:

Fan voting (100%)

Since 2017:

- Fan voting (50%)
- NBA players (25%)
- Media (25%)

Why does selection matter?

$100,000 cash prize

Contract incentives

$1.2 MILLION	$1.3 MILLION	$1.5 MILLION

Research Question :
Can Machine Learning be used to select the NBA All-Star teams?

The dataset

"NBA Advanced Stats 2002-2022" – Kaggle (https://www.kaggle.com/datasets/owenrocchi/nba-advanced-stats-20022022)

- **12211 data entries originally**
- **4135 from 2017 – 2022**
- **141 All Stars, 3994 others**

```
AS_NBA.head()
```

Out[14]:

	Unnamed: 0	year_name	Pos	Age	Tm	G	MP	PER	TS%	3PAr	...	OWS	DWS	WS	WS/48	OBPM	DBPM	BPM	VORP	year
8076	8076	2017-Álex Abrines	SG	23	OKC	68	1055	10.1	0.560	0.724	...	1.2	0.9	2.1	0.096	-1.3	-0.4	-1.6	0.1	2017
8077	8077	2017-Quincy Acy	PF	26	TOT	38	558	11.8	0.565	0.529	...	0.5	0.5	0.9	0.082	-1.5	-0.6	-2.1	0.0	2017
8078	8078	2017-Quincy Acy	PF	26	DAL	6	48	-1.4	0.355	0.412	...	-0.2	0.0	-0.1	-0.133	-10.3	-4.1	-14.3	-0.1	2017
8079	8079	2017-Quincy Acy	PF	26	BRK	32	510	13.1	0.587	0.542	...	0.6	0.5	1.1	0.102	-0.6	-0.2	-0.9	0.1	2017
8080	8080	2017-Steven Adams	C	23	OKC	80	2389	16.5	0.589	0.002	...	3.3	3.1	6.5	0.130	-0.2	0.0	-0.2	1.1	2017

5 rows × 29 columns

```
NBA.head()
```

	Age	MP	PER	TS%	3PAr	FTr	ORB%	DRB%	TRB%	AST%	...	TOV%	USG%	OWS	DWS	WS	WS/48	OBPM	DBPM	BPM	VORP
8076	23	1055	10.1	0.560	0.724	0.144	1.9	7.1	4.5	5.5	...	8.3	15.9	1.2	0.9	2.1	0.096	-1.3	-0.4	-1.6	0.1
8077	26	558	11.8	0.565	0.529	0.353	3.9	18.0	11.0	4.9	...	9.7	16.8	0.5	0.5	0.9	0.082	-1.5	-0.6	-2.1	0.0
8078	26	48	-1.4	0.355	0.412	0.176	4.6	15.2	9.7	0.0	...	9.8	20.0	-0.2	0.0	-0.1	-0.133	-10.3	-4.1	-14.3	-0.1
8079	26	510	13.1	0.587	0.542	0.373	3.8	18.2	11.1	5.4	...	9.6	16.5	0.6	0.5	1.1	0.102	-0.6	-0.2	-0.9	0.1
8080	23	2389	16.5	0.589	0.002	0.392	13.0	15.4	14.2	5.4	...	16.0	16.2	3.3	3.1	6.5	0.130	-0.2	0.0	-0.2	1.1

5 rows × 22 columns

Dimensionality reduction - PCA

The core concept of Principle Component Analysis identifies linear combinations of the dataset that account for the greatest amount of variance in the data.

1. Data Standardization - Before applying PCA, standardize the data by subtracting the mean and scaling to unit variance. This ensures that each feature contributes equally to the analysis.
2. Covariance Matrix Calculation - The covariance matrix is computed based on the standardized data. The covariance matrix expresses how each feature in the data set varies with every other feature.
3. Eigendecomposition - The next step involves calculating the eigenvectors and eigenvalues of the covariance matrix. Eigenvectors represent the directions or components with the highest variance, while eigenvalues indicate the magnitude of variance in those directions.
4. Selection of Principal Components - The eigenvectors are ranked by their corresponding eigenvalues in descending order. The eigenvectors with the highest eigenvalues (principal components) capture the most variance in the data.
5. Projection - The selected principal components are used to create a new subspace by projecting the original data onto these components. This results in a lower-dimensional representation of the data.

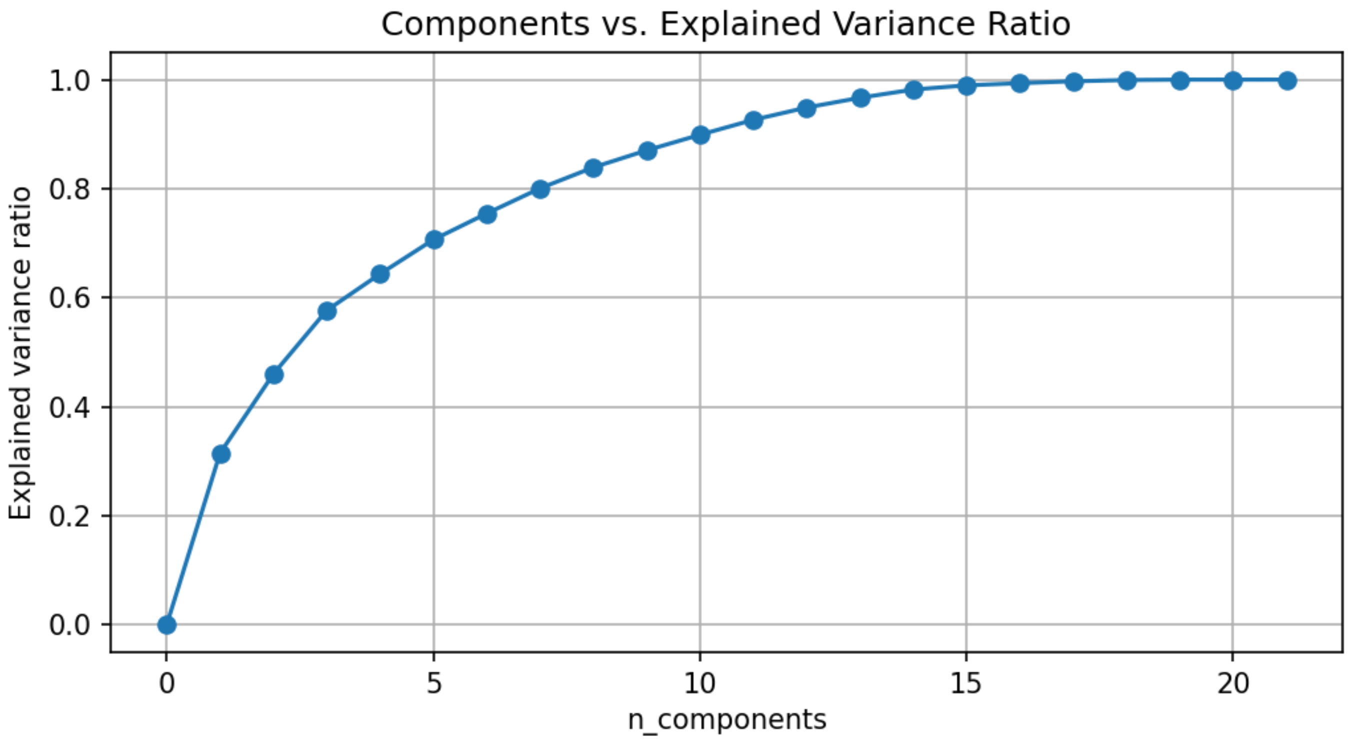
Alternatives

t-SNE or UMAP – They are very good a visualising high dimensional data

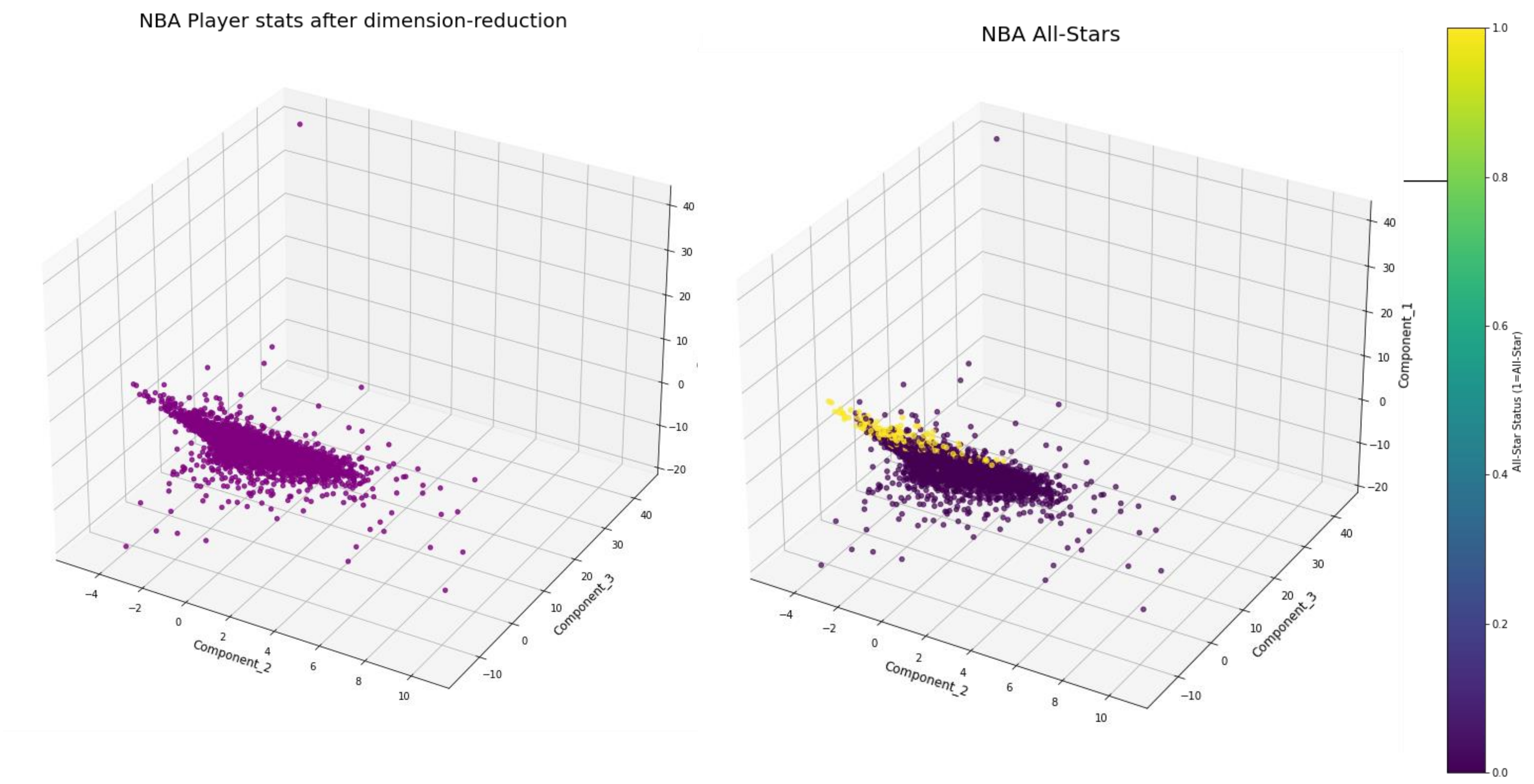
The relationship between clusters is lost

Additional problem with the hyperparameters

Dimensionality reduction - PCA



- The core concept of Principle Component Analysis identifies linear combinations of the dataset that account for the greatest amount of variance in the data.
- How many parameters is enough?
- After 3 components the explained variance added is minimal

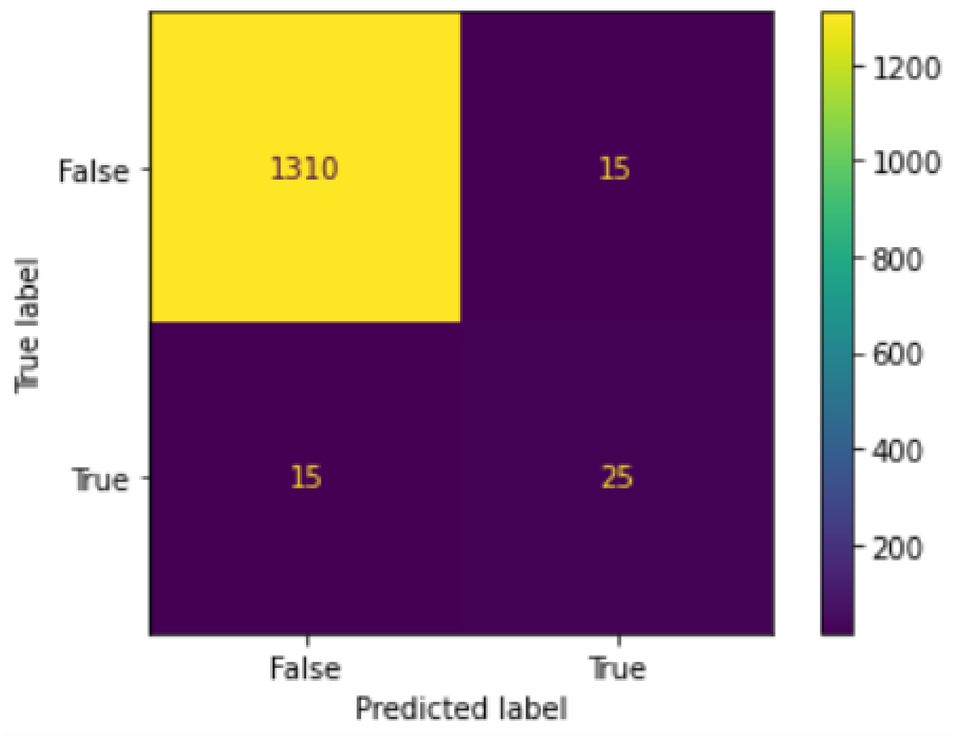


NBA Player stats after dimension-reduction
NBA All-Stars
Component_1
Component_2
Component_3
All-Star Status (1=All-Star)

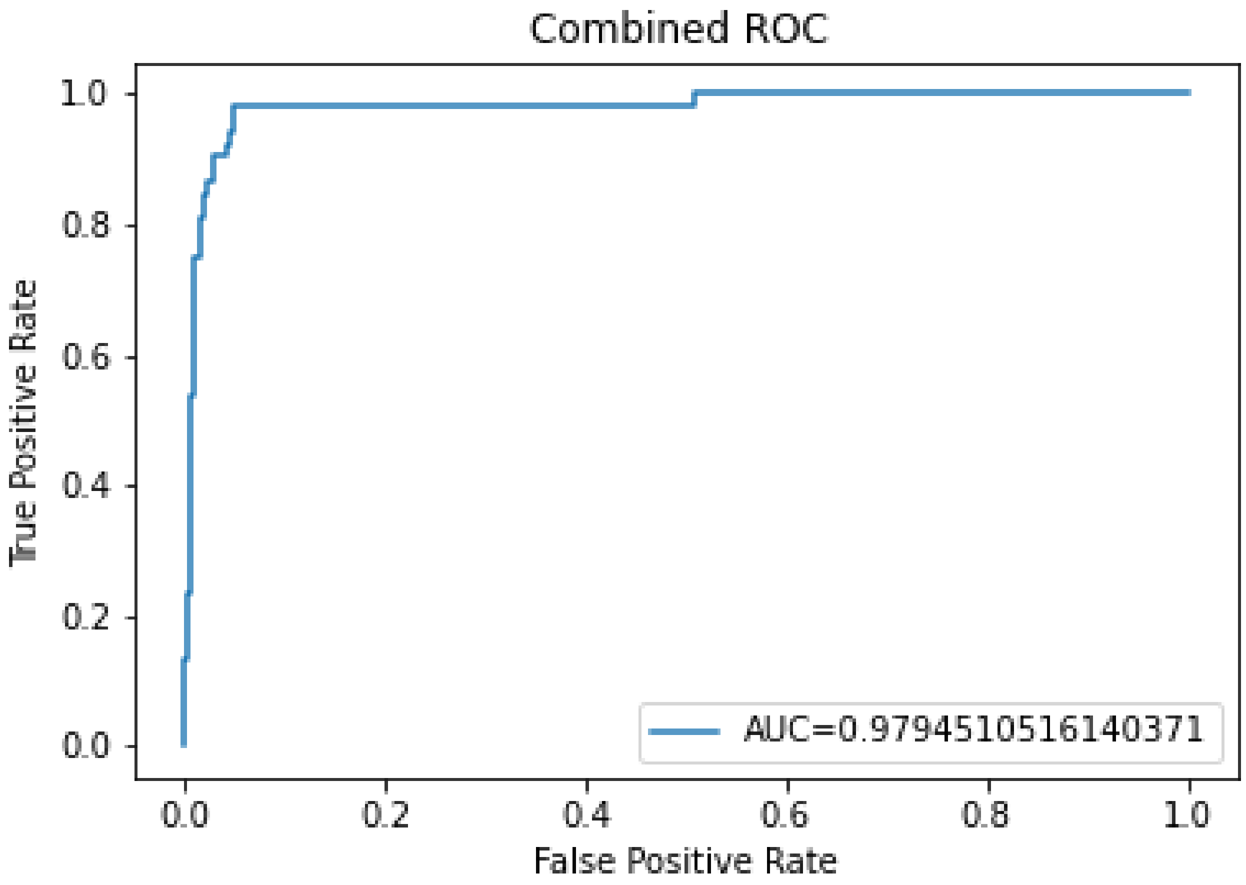
Model – Multivariate Logistic Regression

Model to select players for the All-Star game based on their stats, the 3 components from PCA

Results



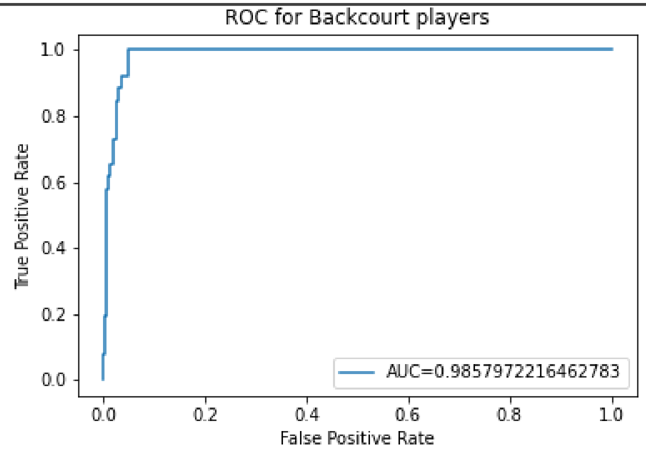
Accuracy of all players: 0.978021978021978
Precision of all players: 0.625
Recall of all players: 0.625
F1-score of all players: 0.625

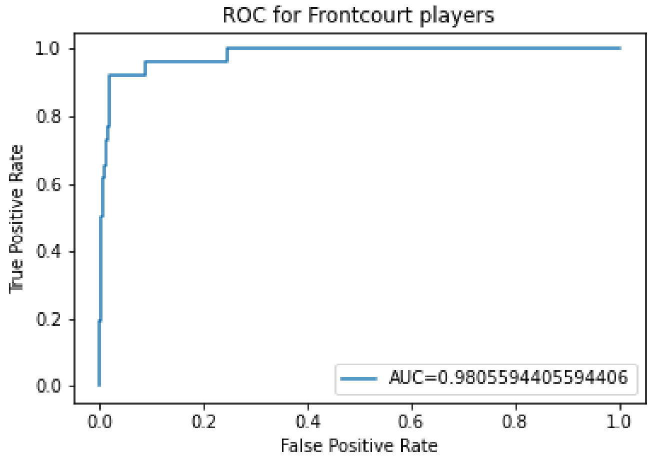

Combined ROC
True Positive Rate
False Positive Rate
AUC=0.9794510516140371
0.0
0.2
0.4
0.6
0.8
1.0

ROC Analysis

Area Under Curve suggests that the model is almost a perfect classification.

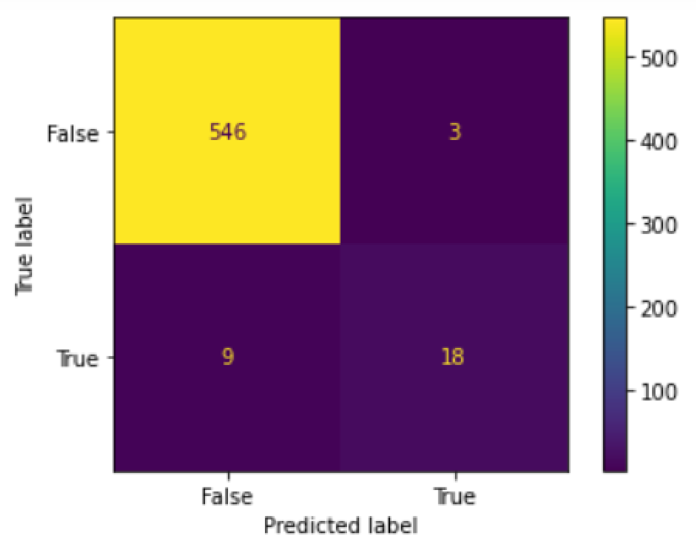
When front and backcourt players are separated the model is slightly better.



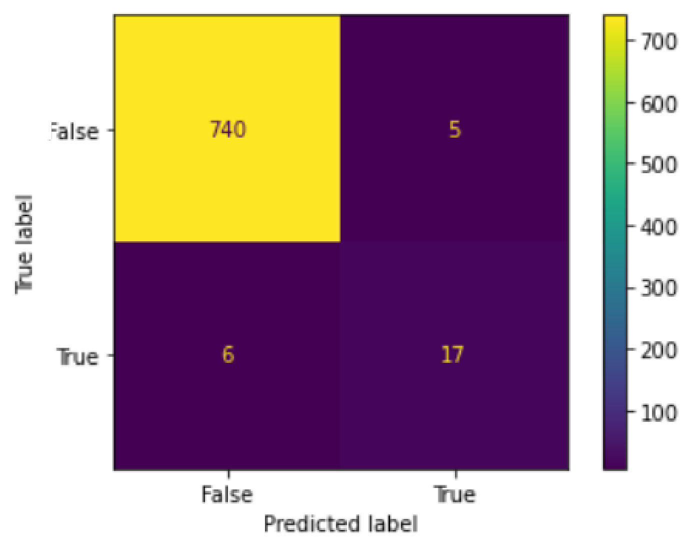


```
Accuracy ONLY Frontcourt players: 0.9774305555555556
Precision ONLY Frontcourt players: 0.9090909090909091
Recall ONLY Frontcourt players: 0.45454545454545453
F1-score ONLY Frontcourt players: 0.6060606060606061

Accuracy ONLY Backcourt players: 0.9765625
Precision ONLY Backcourt players: 0.6666666666666666
Recall ONLY Backcourt players: 0.43478260869565216
F1-score ONLY Backcourt players: 0.5263157894736841
```



Frontcourt



Backcourt

Analysis

Fairly good model

Use of a Naive Bayesian model?

Logistic Regression makes a prediction for the probability using a direct functional form where as Naive Bayes figures out how the data was generated given the results

Thoughts

I can see ML being used in NBA All-Star selection to make it fair for all players.

It disconnects the selection process from fans. Could impact viewership positivitly and negatively.

THANK YOU!