PLAYER SIMILARITY TOOL FOR RECRUITMENT

Sathish Prasad VT - 185001137

Sreyas V - 185001162

Sricharan PP - 185001163

Sriram M - 185001168

PROBLEM STATEMENT

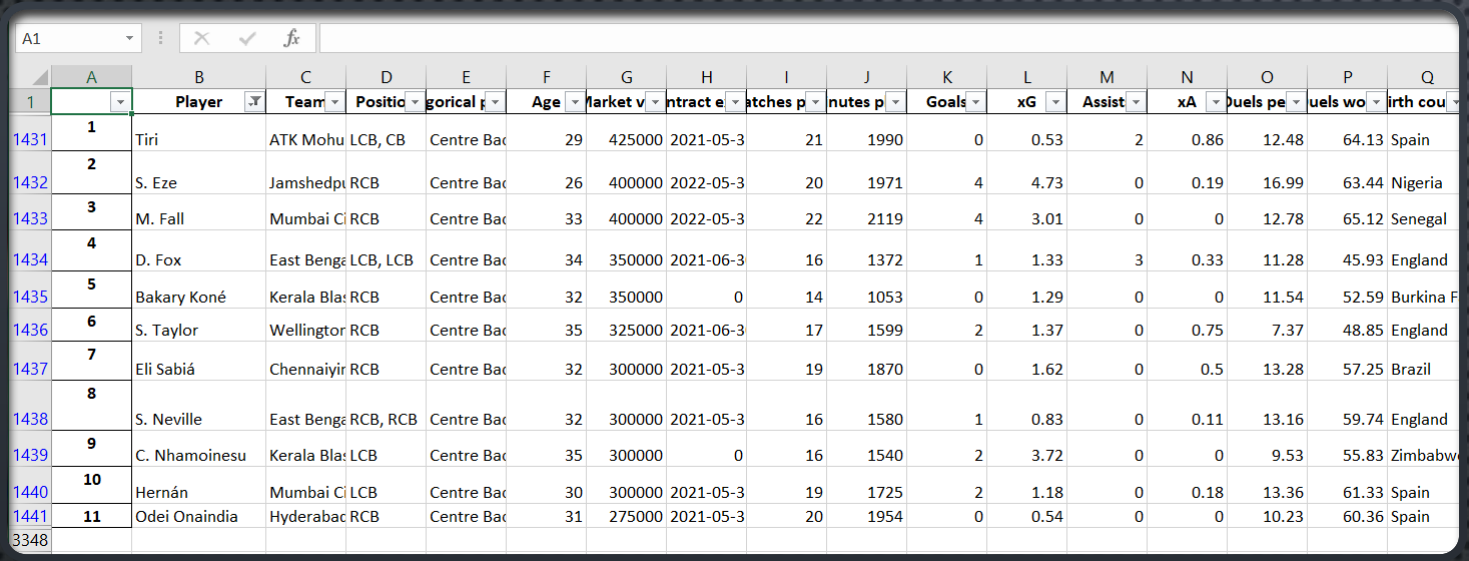
We are aiming to build a player similarity model, which can be used by football teams for scouting, talent recruitment and even in transfers. This will mainly help teams to find players with similar traits and characteristics to top players in the world. This will be a really useful tool for teams mainly in lower tier leagues and those who can't afford to bring in top players for their team.

OBJECTIVE

The main objective of this task is to find players with a similar style of play to a particular player from the dataset. The similarity can be calculated for all attributes of the player or certain select attributes provided by the user.

As another objective, we have performed interpolation of players of same/different positions and generated players with the combined features of both the selected players. This is achieved by taking the vector representation of one player, then mixing it with the vector of the second player to get the interpolated vector, and then looking for the closest players from the dataset. This feature is really helpful for teams who want to recruit multi- faceted players and also look for talents in similar areas.

DATASET



	A	Player	Team	Positio	gorical p	Age	Market v	ntract e	atches p	nutes p	Goals	xG	Assist	xA	Duels pe	uels wo	irth cou
1431	1	Tiri	ATK Mohu	LCB, CB	Centre Bac	29	425000	2021-05-3	21	1990	0	0.53	2	0.86	12.48	64.13	Spain
1432	2	S. Eze	Jamshedpu	RCB	Centre Bac	26	400000	2022-05-3	20	1971	4	4.73	0	0.19	16.99	63.44	Nigeria
1433	3	M. Fall	Mumbai Ci	RCB	Centre Bac	33	400000	2022-05-3	22	2119	4	3.01	0	0	12.78	65.12	Senegal
1434	4	D. Fox	East Benga	LCB, LCB	Centre Bac	34	350000	2021-06-3	16	1372	1	1.33	3	0.33	11.28	45.93	England
1435	5	Bakary Koné	Kerala Bla	RCB	Centre Bac	32	350000	0	14	1053	0	1.29	0	0	11.54	52.59	Burkina F
1436	6	S. Taylor	Wellingtor	RCB	Centre Bac	35	325000	2021-06-3	17	1599	2	1.37	0	0.75	7.37	48.85	England
1437	7	Eli Sabiá	Chennaiyir	RCB	Centre Bac	32	300000	2021-05-3	19	1870	0	1.62	0	0.5	13.28	57.25	Brazil
1438	8	S. Neville	East Benga	RCB, RCB	Centre Bac	32	300000	2021-05-3	16	1580	1	0.83	0	0.11	13.16	59.74	England
1439	9	C. Nhamoinesu	Kerala Bla	LCB	Centre Bac	35	300000	0	16	1540	2	3.72	0	0	9.53	55.83	Zimbabw
1440	10	Hernán	Mumbai Ci	LCB	Centre Bac	30	300000	2021-05-3	19	1725	2	1.18	0	0.18	13.36	61.33	Spain
1441	11	Odei Onaindia	Hyderabac	RCB	Centre Bac	31	275000	2021-05-3	20	1954	0	0.54	0	0	10.23	60.36	Spain
3348																	

- Our dataset consists of all the players from the top 5 football leagues (Premier League, La Liga, Bundesliga, Serie A and Ligue 1) and also the Dutch and Indian leagues. It contains the data of around 3500 players. The total number of attributes for each player is 105, regardless of his position. For example, an attacker may have zeros in goalkeeping attributes, but still those attributes will be present.
- We have split the attributes based on whether they are pertaining to one of the three basic divisions : attacking, defence and passing.

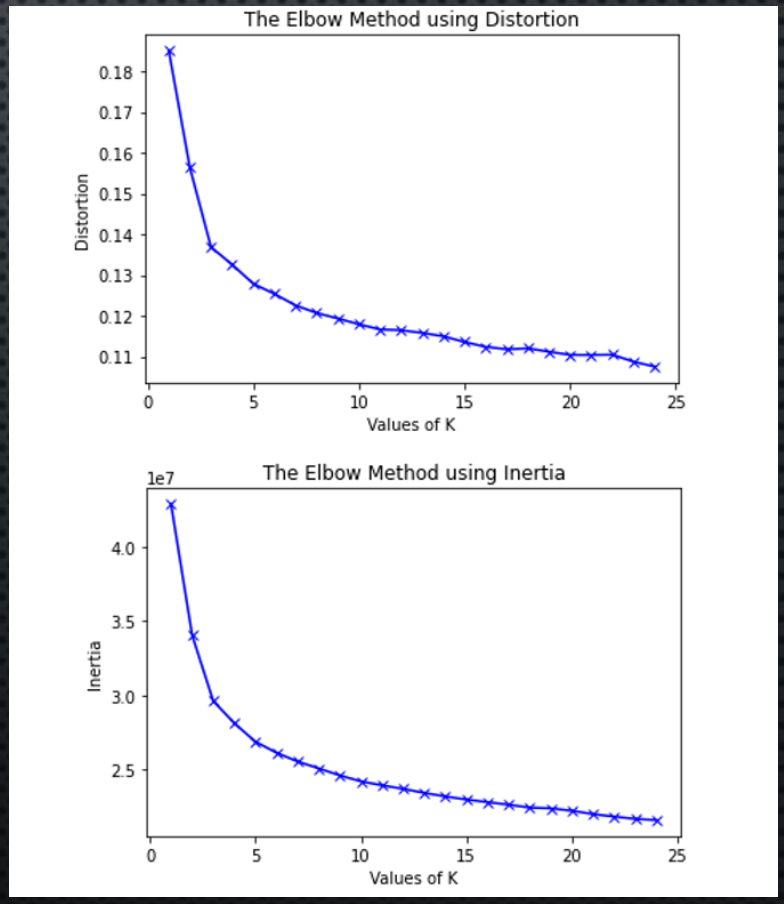
METHODOLOGY

SOME OF THE IMPORTANT MACHINE LEARNING TECHNIQUES WE USED ARE :

ELBOW METHOD

THE ELBOW METHOD IS USED TO DETERMINE THE NUMBER OF CLUSTERS NEEDED FOR EFFECTIVE PERFORMANCE OF THE MODEL. THE INTUITION IS THAT INCREASING THE NUMBER OF CLUSTERS WILL NATURALLY IMPROVE THE FIT (SINCE THERE ARE MORE PARAMETERS TO USE) BUT THAT AT SOME POINT THIS CAN LEAD TO OVERFITTING. THE ELBOW MODEL REFLECTS THIS.

DISTORTION (VARIANCE) : IT IS CALCULATED AS THE AVERAGE OF THE COSINE DISTANCES FROM THE CLUSTER CENTERS OF THE RESPECTIVE CLUSTERS.
INERTIA : IT IS THE SUM OF DISTANCES OF SAMPLES TO THEIR CLOSEST CLUSTER CENTERS.



METHODOLOGY

COSINE SIMILARITY

COSINE SIMILARITY MEASURES THE SIMILARITY BETWEEN TWO VECTORS BY CALCULATING THE COSINE OF THE ANGLE BETWEEN THE TWO VECTORS.

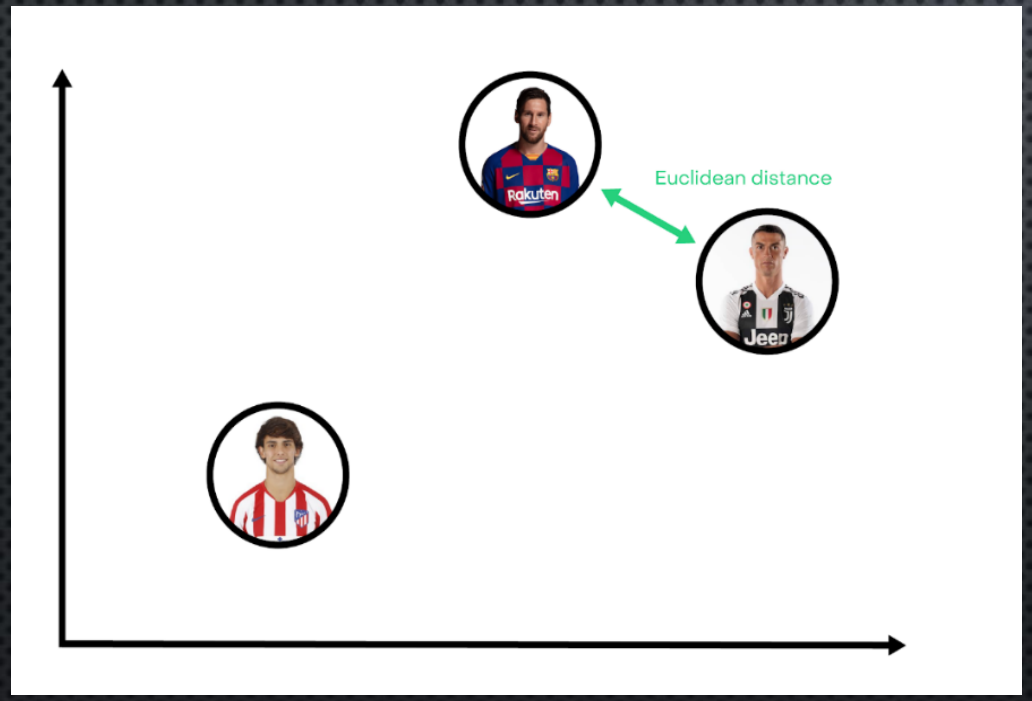
COSINE SIMILARITY IS ONE OF THE MOST WIDELY USED AND POWERFUL SIMILARITY MEASURES IN DATA SCIENCE. IT IS USED IN MULTIPLE APPLICATIONS SUCH AS FINDING SIMILAR DOCUMENTS IN NLP, INFORMATION RETRIEVAL, FINDING SIMILAR SEQUENCES TO DNA IN BIOINFORMATICS, DETECTING PLAGIARISM AND MANY MORE.

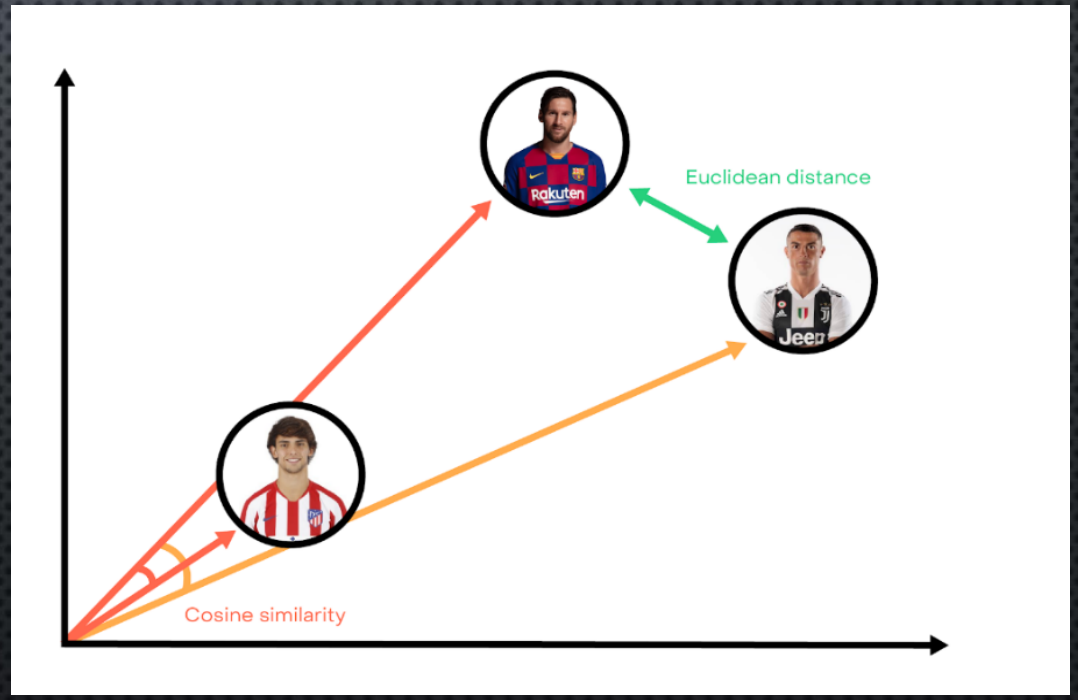
CALCULATED AS :

$$cos(\theta) = \frac{x.y}{\|x\| \, \|y\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}$$

METHODOLOGY

COSINE SIMILARITY VS EUCLIDEAN DISTANCE



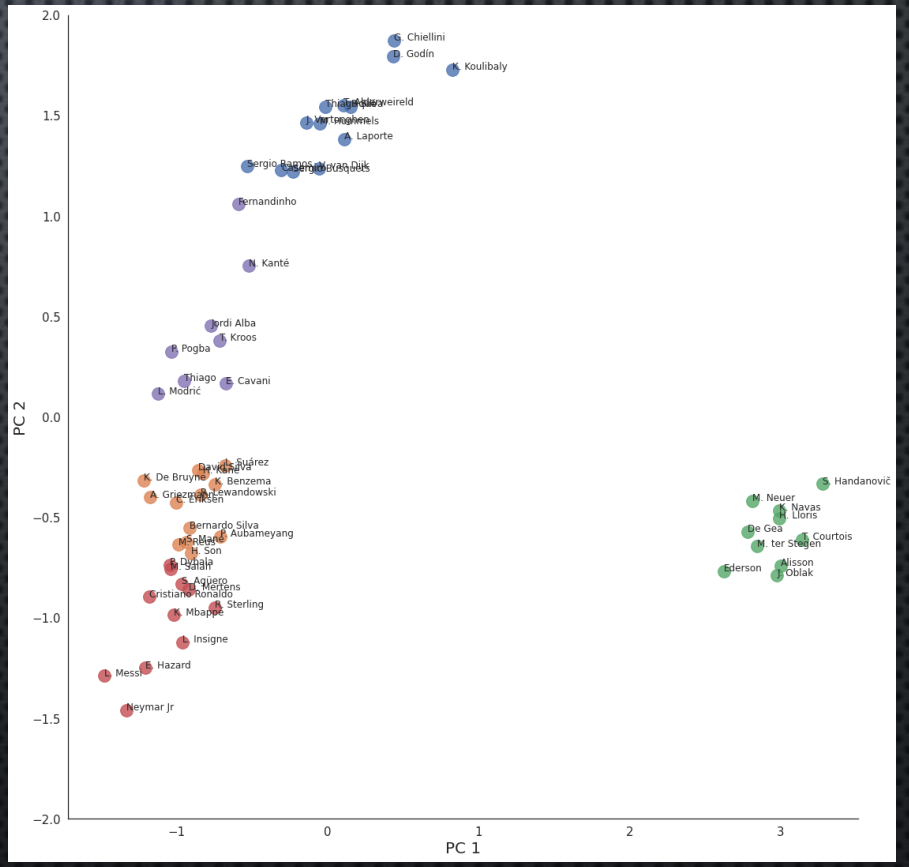


METHODOLOGY

CLUSTERING

CLUSTERING IS ONE OF THE UNSUPERVISED LEARNING TECHNIQUES. WE CAN CLUSTER (OR GROUP) OBSERVATIONS INTO THE SAME SUBGROUPS SO THAT OBSERVATIONS WITHIN A SUBGROUP ARE QUITE SIMILAR TO EACH OTHER AND OBSERVATIONS IN DIFFERENT SUBGROUPS ARE QUITE DIFFERENT FROM EACH OTHER.

WE USE THE **K-MEANS CLUSTERING** TECHNIQUE FOR THIS MODEL.

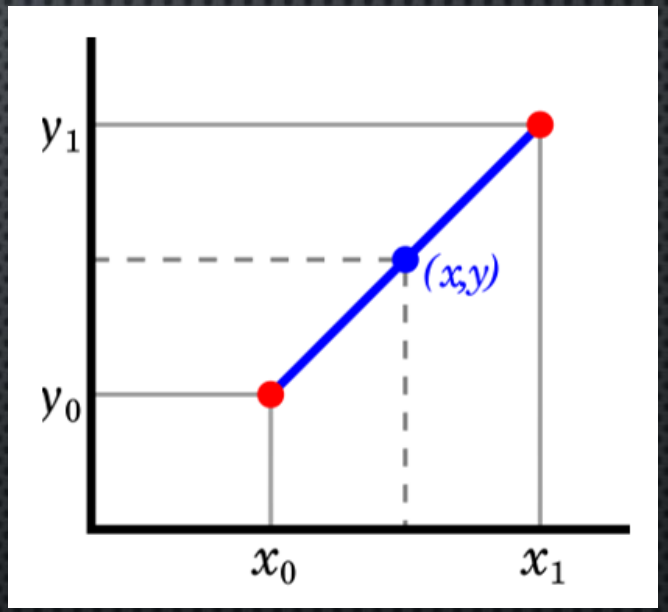


METHODOLOGY

VECTOR INTERPOLATION

MANY A TIMES WE COME ACROSS USE CASES WHERE THE WE NEED TO FIND PLAYERS WITH ATTRIBUTES OF TWO OR MORE PLAYERS, TO SOLVE THIS PROBLEM WE USE VECTOR INTERPOLATION. IN MATHEMATICS, LINEAR INTERPOLATION IS A METHOD OF CURVE FITTING USING LINEAR POLYNOMIALS TO CONSTRUCT NEW DATA POINTS WITHIN THE RANGE OF A DISCRETE SET OF KNOWN DATA POINTS.

THIS IS THE GENERAL FORMULA FOR VECTOR INTERPOLATION.



$$v = \alpha.x + (1 - \alpha).y$$

SAMPLE RESULTS