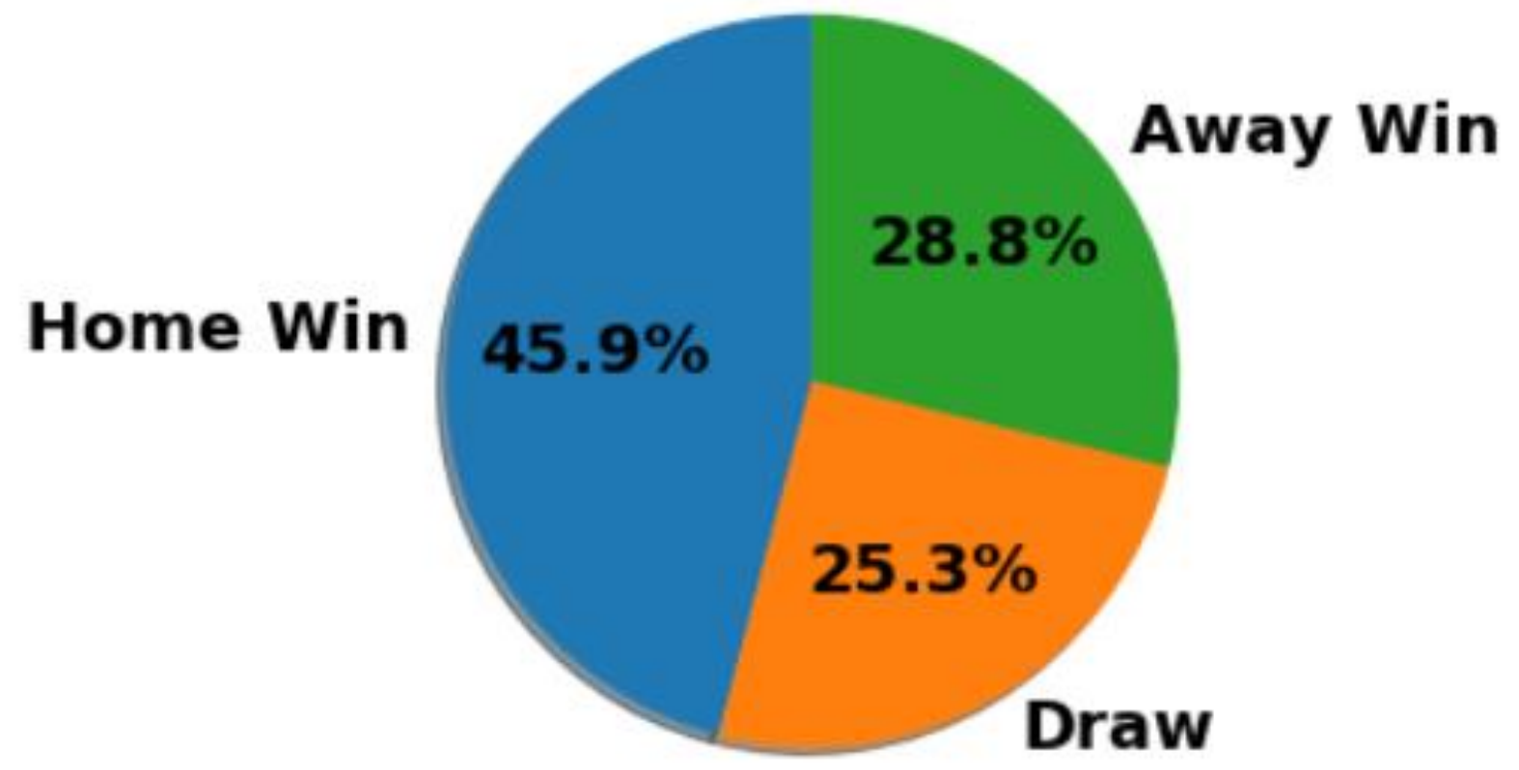
Uncovering the Great Betting Conspiracy

Mete Morris, Jiwon Lee, Jonghae Lee

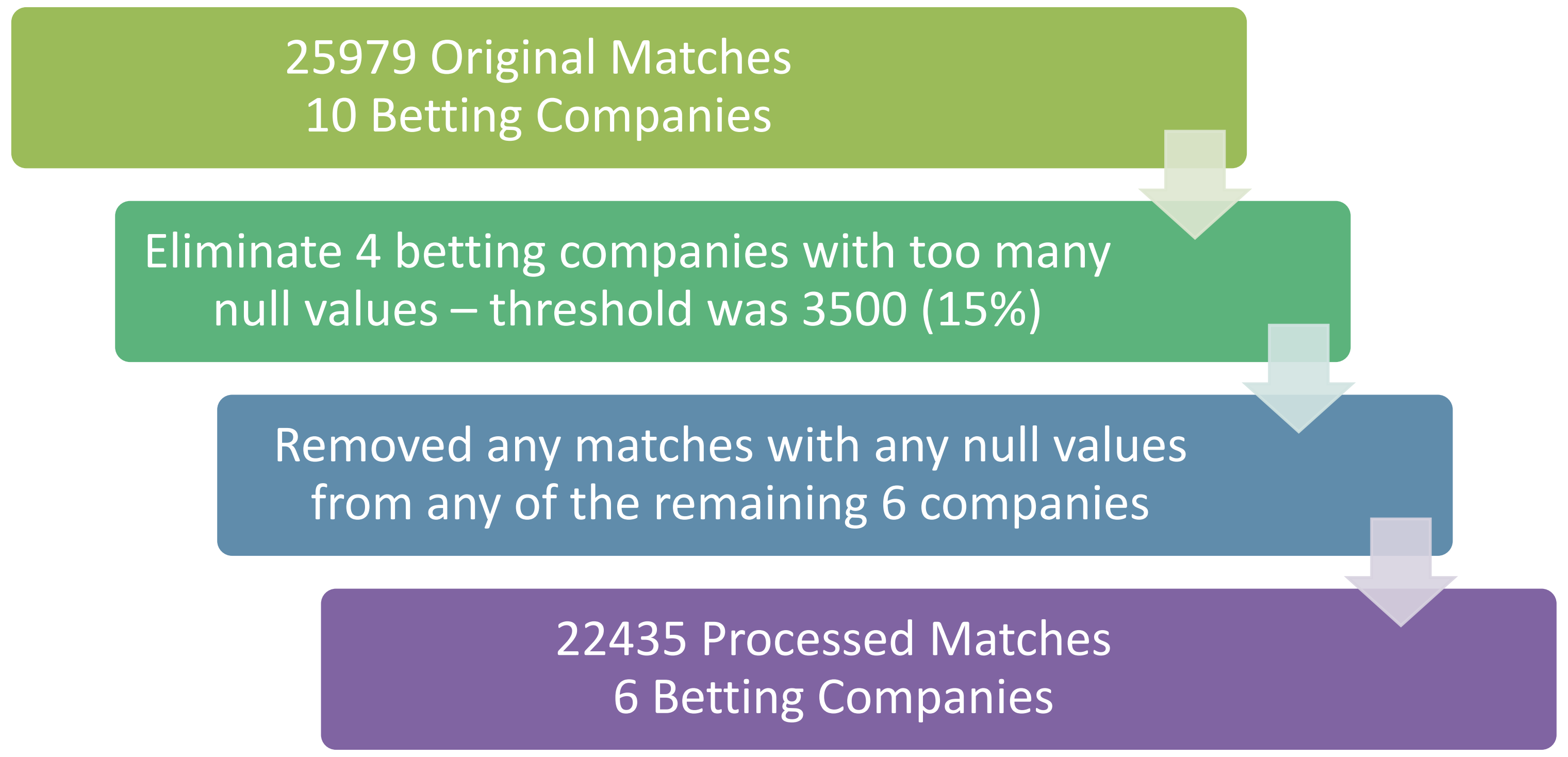
Johns Hopkins University | Whiting School of Engineering | Baltimore, MD

Introduction

Our objective is to analyze the accuracy of soccer betting companies based on empirical data from the **European Soccer Database.** We look at the betting odds of each company and compare real-life results. Furthermore, we then implement classification methods to create our own predictions about soccer match outcomes and compared them against the betting companies. In total, we looked at 6 companies and their betting odds for wins, losses, and draws for home and away teams for 22432 matches.



Data Pre-Processing

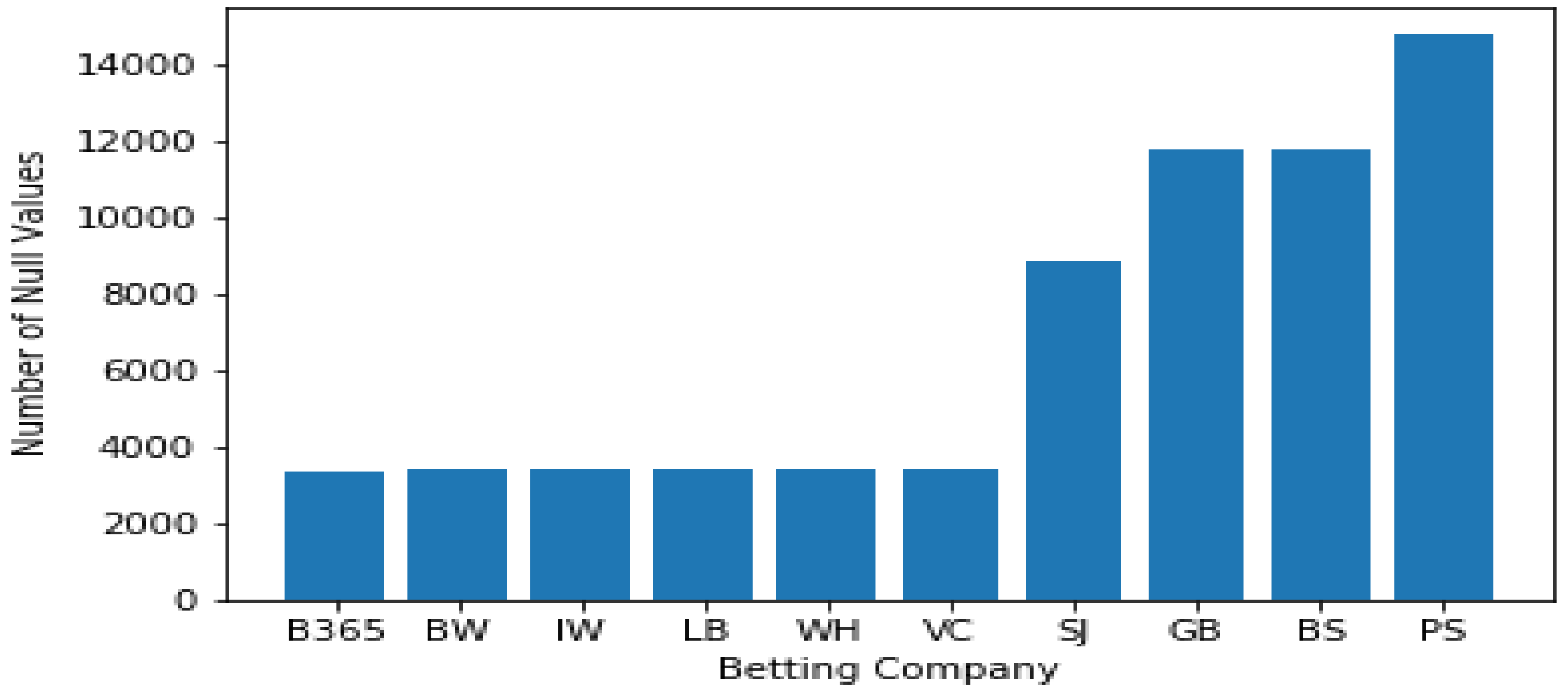


25979 Original Matches
10 Betting Companies

Eliminate 4 betting companies with too many null values – threshold was 3500 (15%)

Removed any matches with any null values from any of the remaining 6 companies

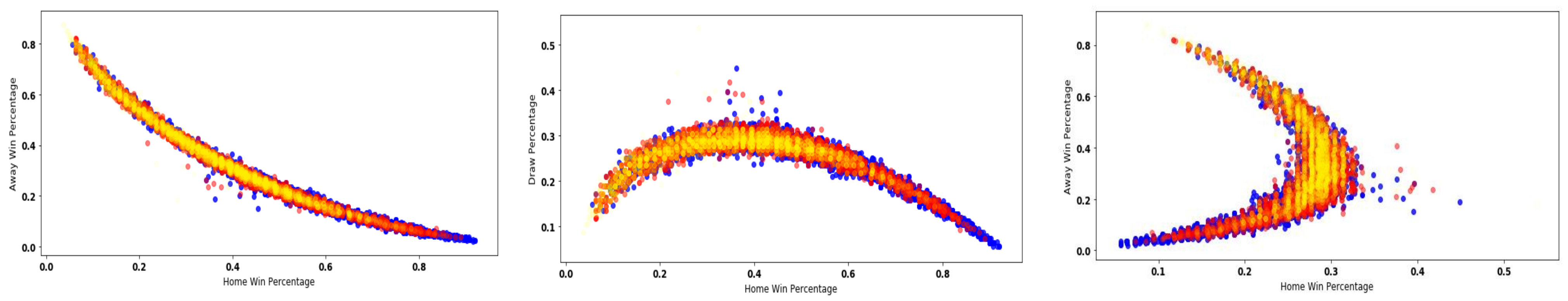
22435 Processed Matches
6 Betting Companies



From decimal odds, we calculated the percentage chance that the betting companies associated using formula (right).

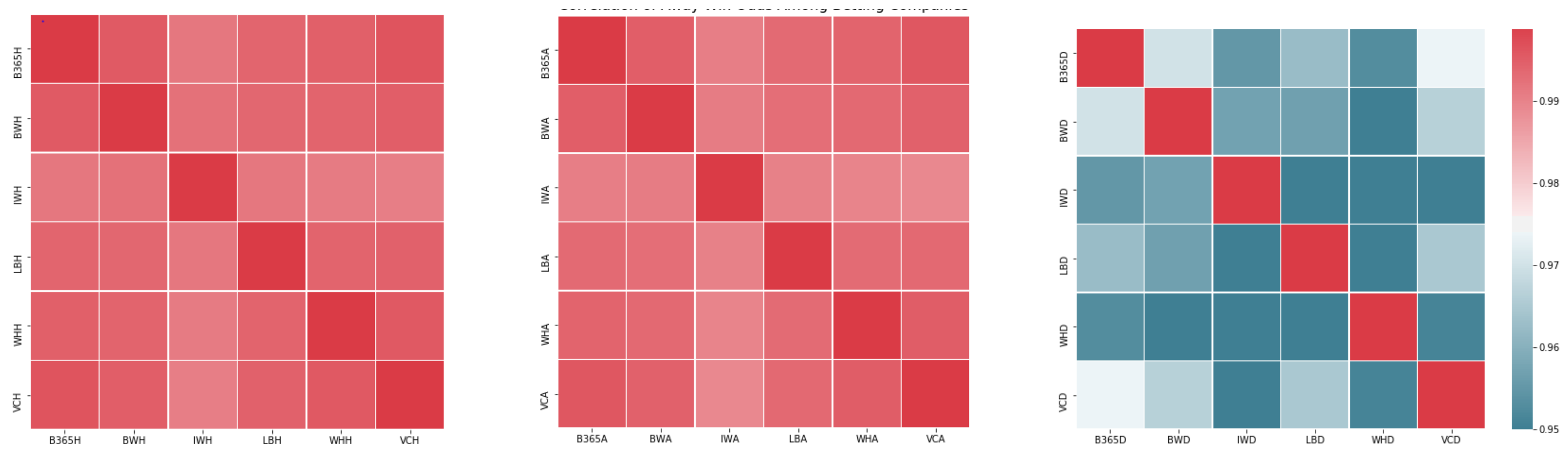
$$d_E = \frac{1}{p_E + o_E}$$

Characteristics of Betting Data



Blue = Home Win   Yellow = Away Win   Red = Draw

Correlation of Company Betting Odds



Based on the correlation matrices of betting odds of home wins, away wins, and draws from all the companies, we saw that there is very little variation in betting odds.

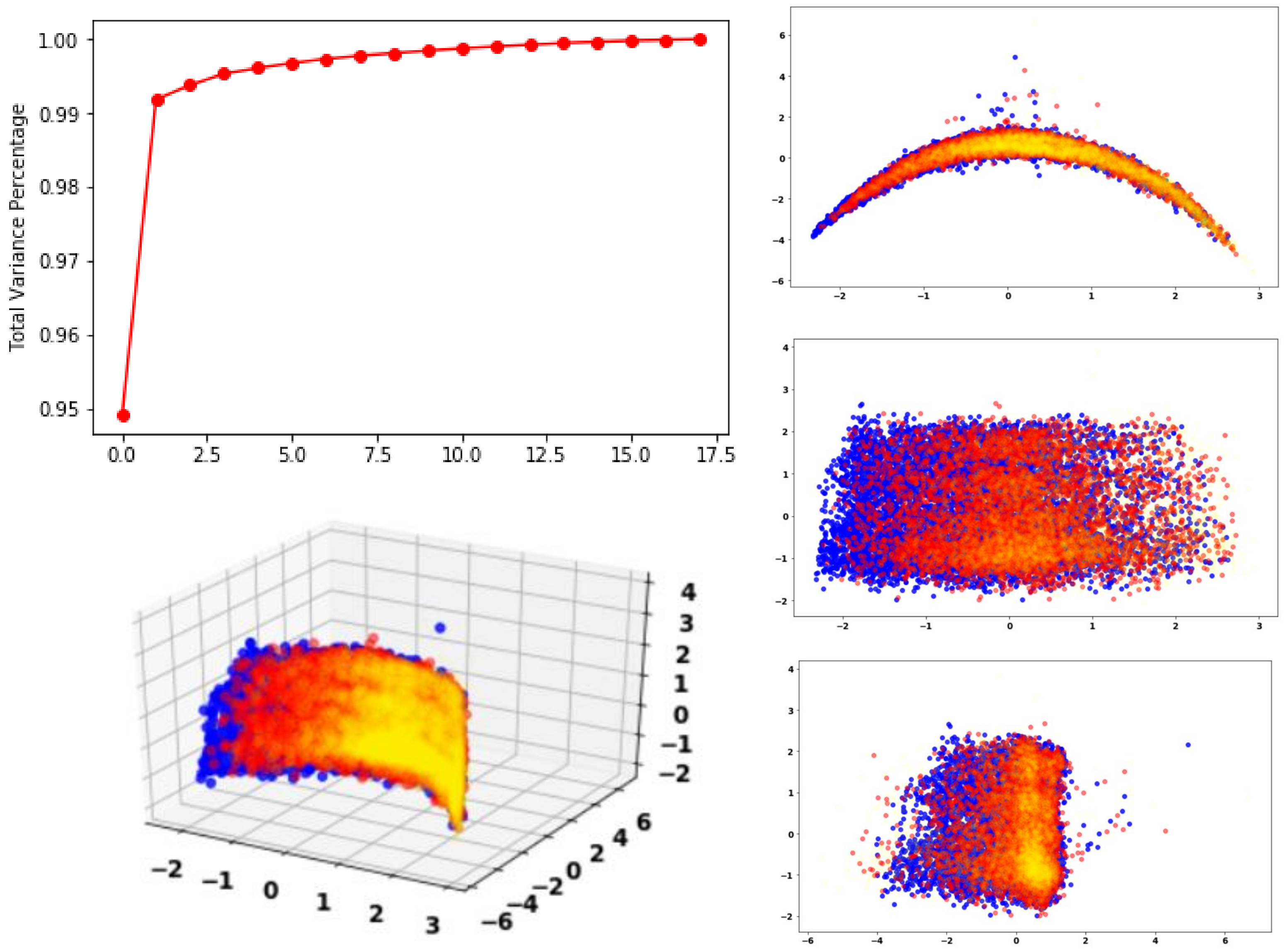
Classification Accuracy before PCA

	Bet 365	Bet & Win	Interwetten	Ladbrokes
kNN (k = 11)	0.5055	0.4672	0.4668	0.4627
Naïve Bayes	0.5195	0.4962	0.5152	0.5268
LDA	0.5579	0.5080	0.5267	0.5455
QDA	0.5473	0.5116	0.5089	0.5072
Random Forest (100 trees)	0.4361	0.4761	0.4443	0.4474
Decision Tree	0.4269	0.4065	0.4457	0.4286
Logistic Regression	0.5312	0.5303	0.5480	0.5596

	William Hill	VC Bet	Cumulative
kNN (k = 11)	0.4536	0.4840	0.4893
Naïve Bayes	0.5090	0.4892	0.4751
LDA	0.5394	0.5534	0.5105
QDA	0.5348	0.5375	0.4645
Random Forest (100 trees)	0.4457	0.4386	0.5000
Decision Tree	0.4376	0.4321	0.4289
Logistic Regression	0.5596	0.5304	0.5537

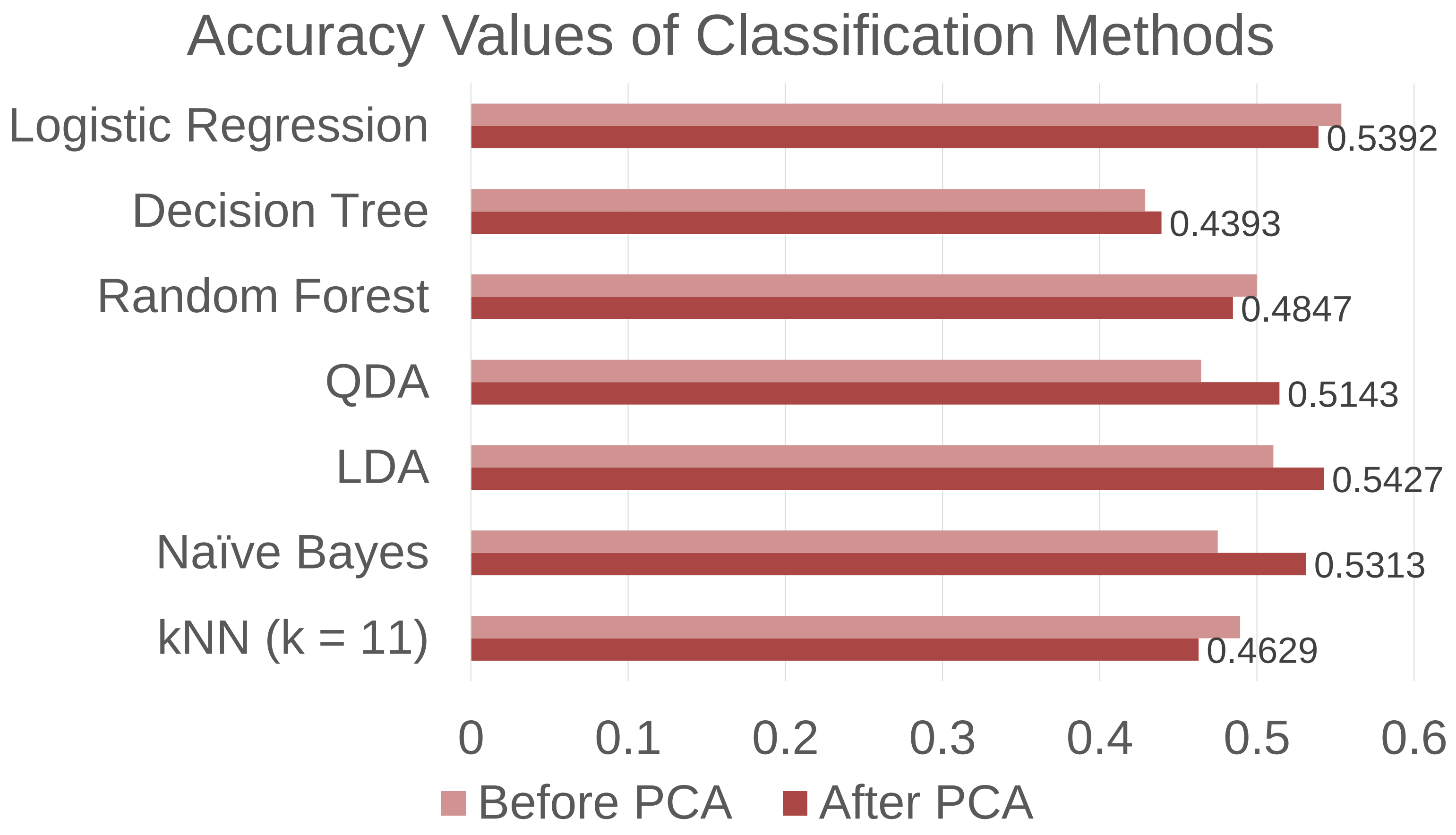
These results were based on 10-fold cross validations through 50 iterations. Among various classifiers, logistic regression and LDA performed the best. In addition, Bet 365 calculated betting odds most accurately, but the margin was almost insignificant.

Principle Component Analysis



PCA revealed 3 components represent 99% of data variance.

Classification Accuracy After PCA



Conclusion

- Over-rounding ensures betting companies a mathematical advantage over its bettors
- Betting odds are highly correlated among companies
- Soccer games are difficult to predict using only betting odds (accuracy never exceeded 56%)
- For future exploration of data, factors other than betting odds may improve prediction of matches