**EE-551 Final Project Report**

**Ye Fang - 10431002**

**Introduction**

k-nearest neighbors is a simple algorithm that stores all available cases and classifies new cases by a majority vote of its k neighbors. This algorithm segregates unlabeled data points into well-defined groups. In a match, there are 5 players on the court for one team. They play different positions. There are 5 positions, Point Guard (PG), Shooting Guard (SG), Small Forward (SF), Power Forward (PF) and Center (C). This project is about data analysis. Through the players stats, we can predict which position the players are acting. Imagine that we have a list of players stats of the rookies (the freshmen in NBA). We can easily find the most suitable position for a rookie.

**1. Data Processing**

In the original dataset, there are many roles I don’t need to and also there are some null values which will affect the results. So, I first select the column “Height”, “Weight”, “Pos”, “PTSPM”, “ASTPM”, “STLPM”, “BLKPM”, “DREBPM”, “OREBPM”, “AST/TOV”, “STL/TOV” and then I use the “data.dropna()” to ignore all the null values. Finally, I create a new dataset that I will used to analyze.

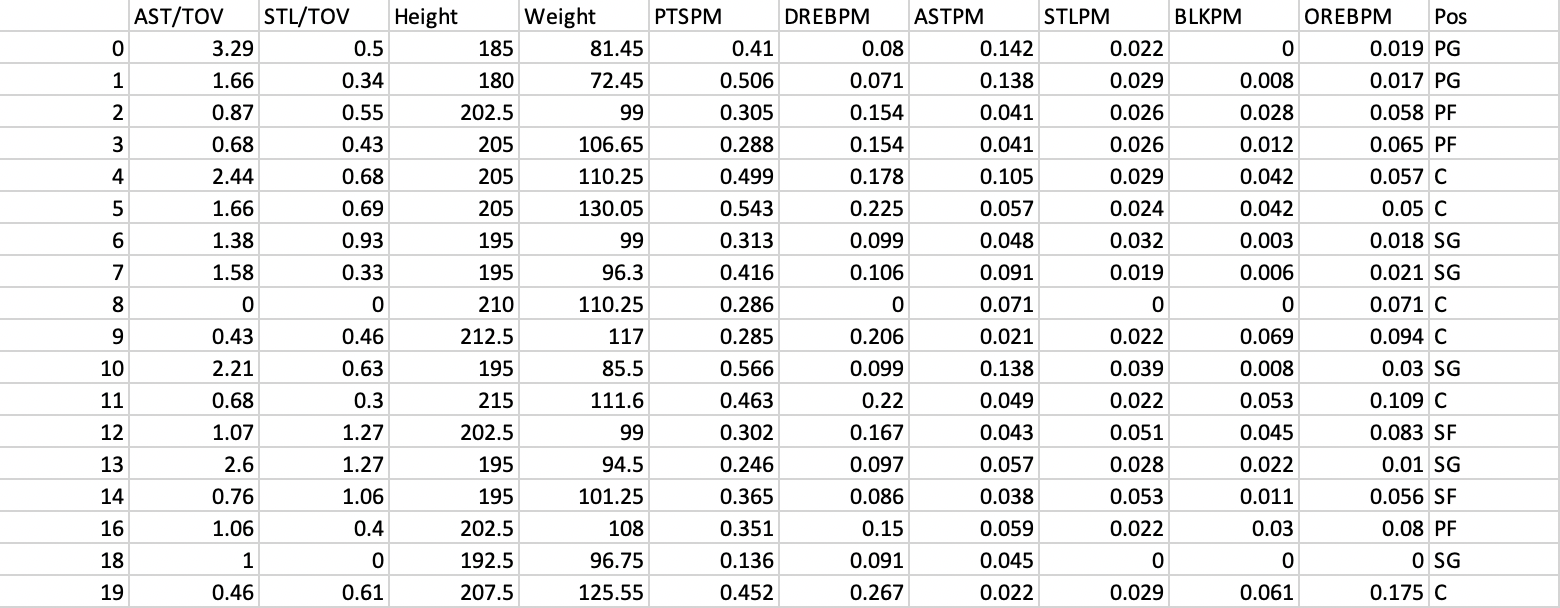


Figure 1. The dataset which used to analyze

**2. Design and Implement kNN algorithm**

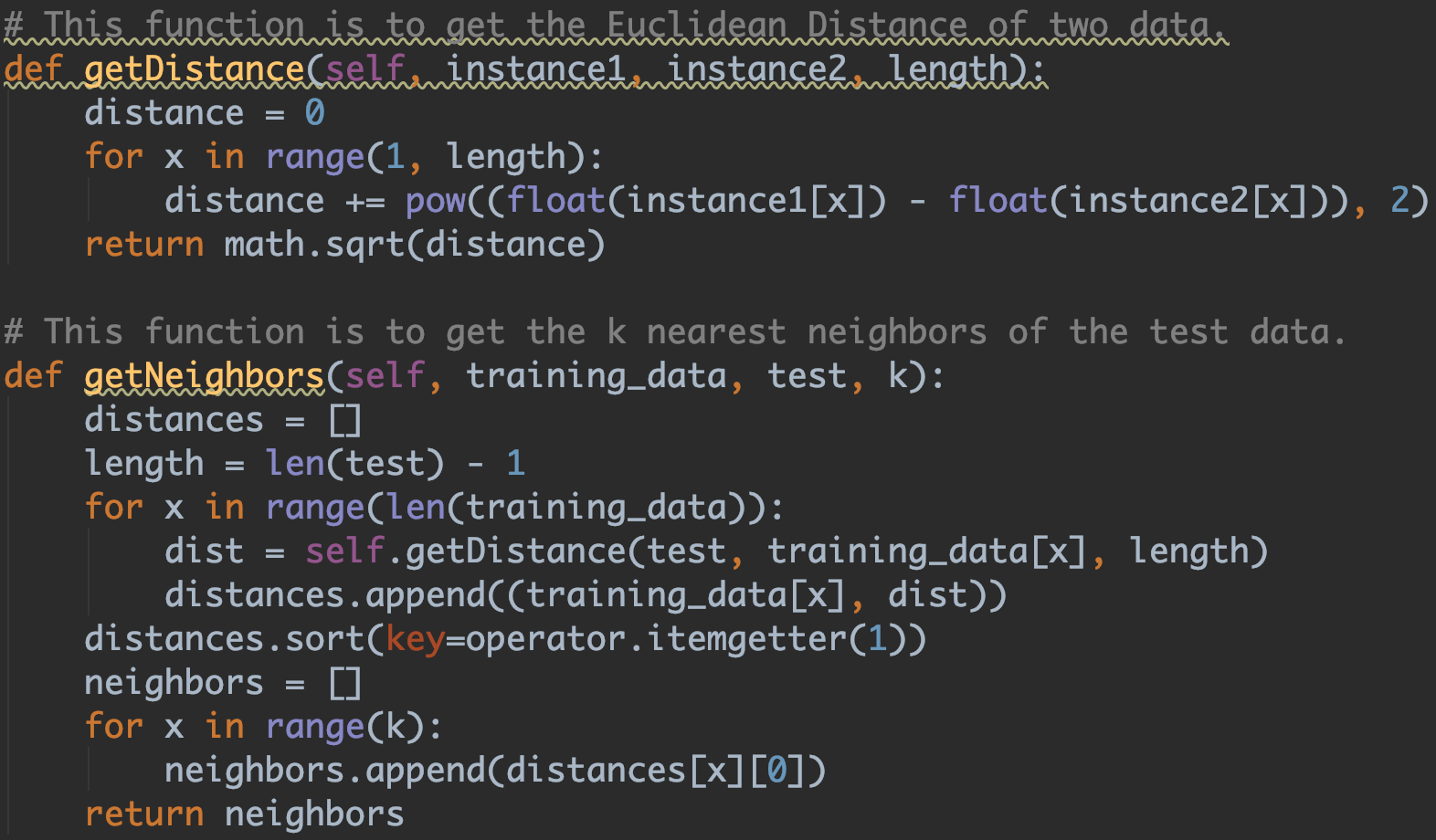


Figure 2. Function ‘getDistance’ and ‘getNeighbors’

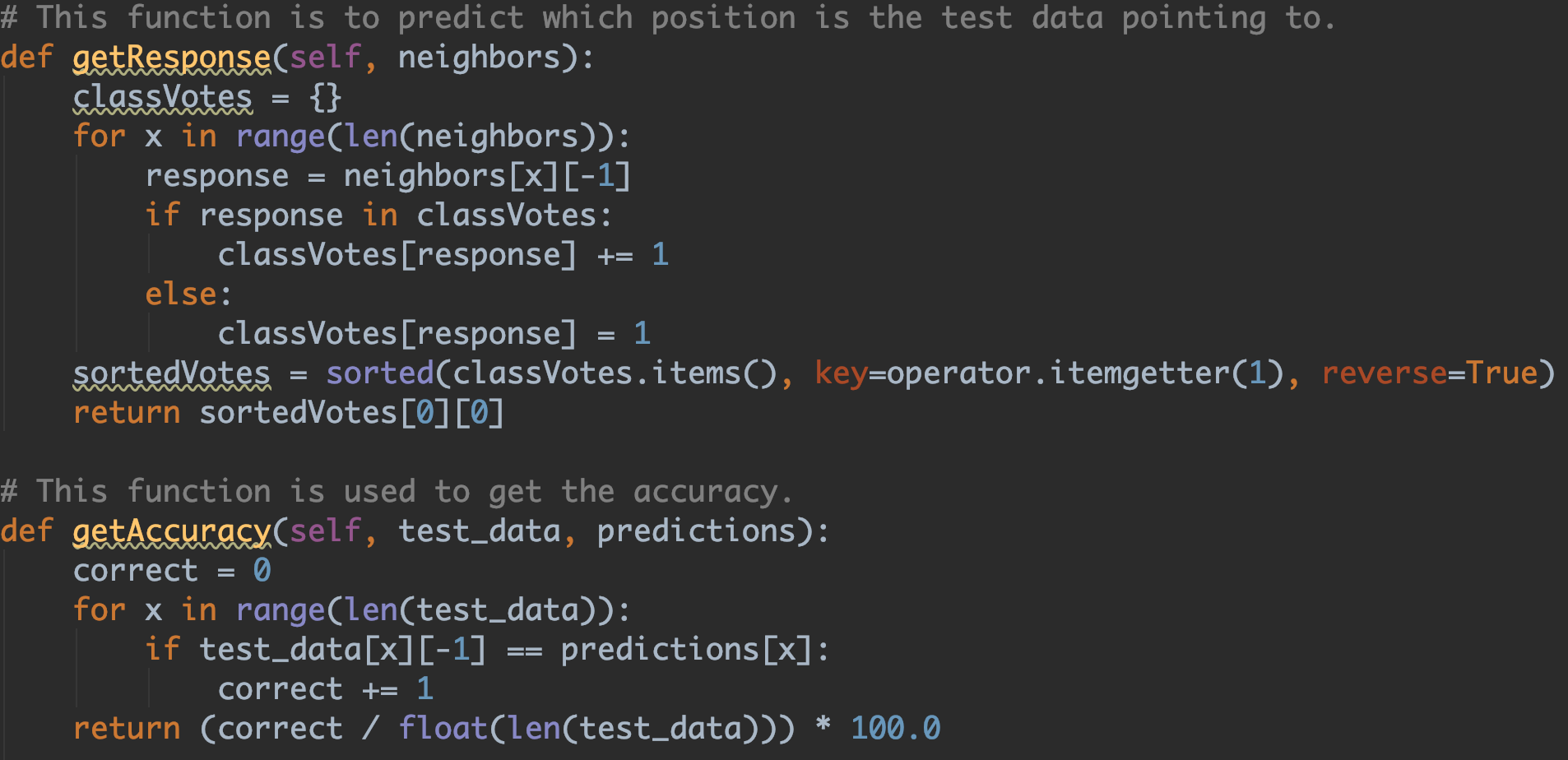


Figure 3. Function ‘getResponse’ and ‘getAccuracy’

**3. Results and Conclusion**

In this project, I test the dataset in different k values.

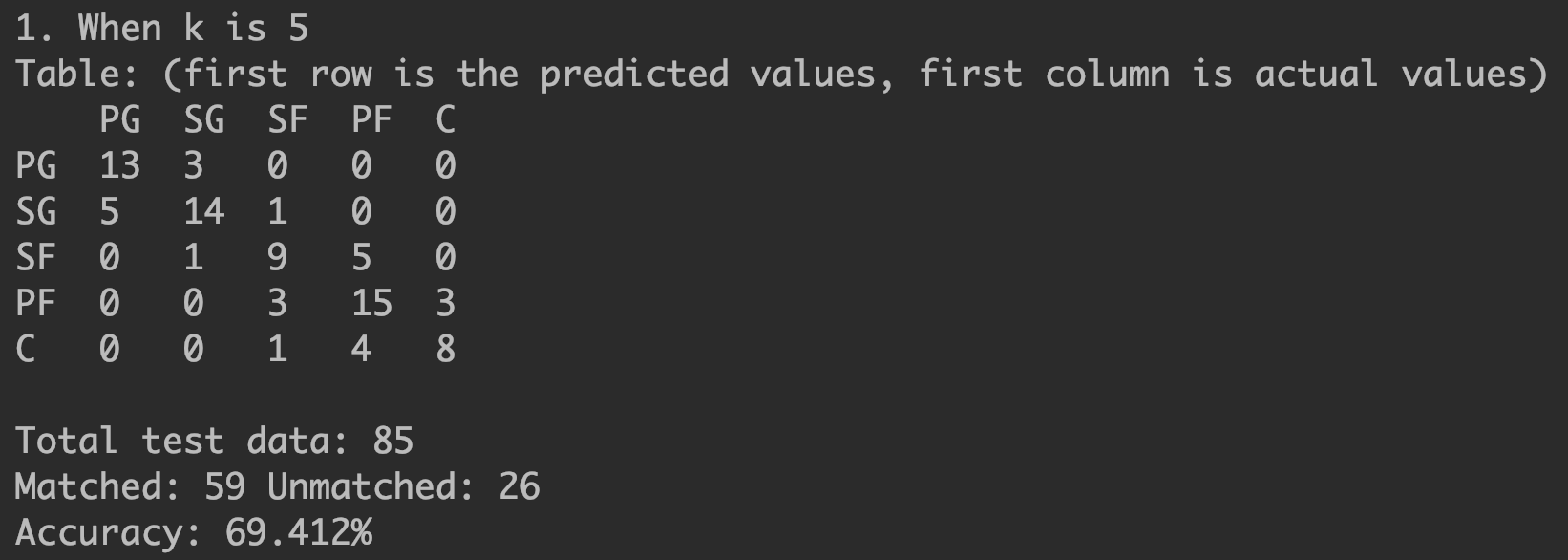


Figure 4. the result of k=5

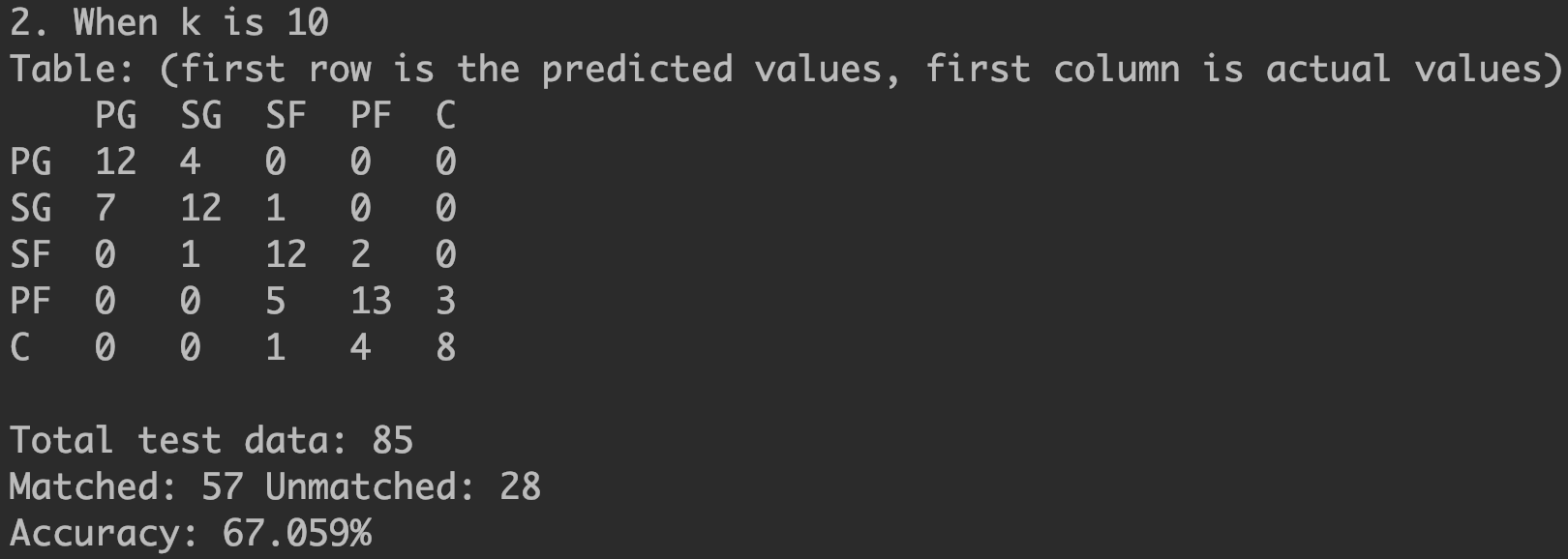


Figure 5. the result of k=10

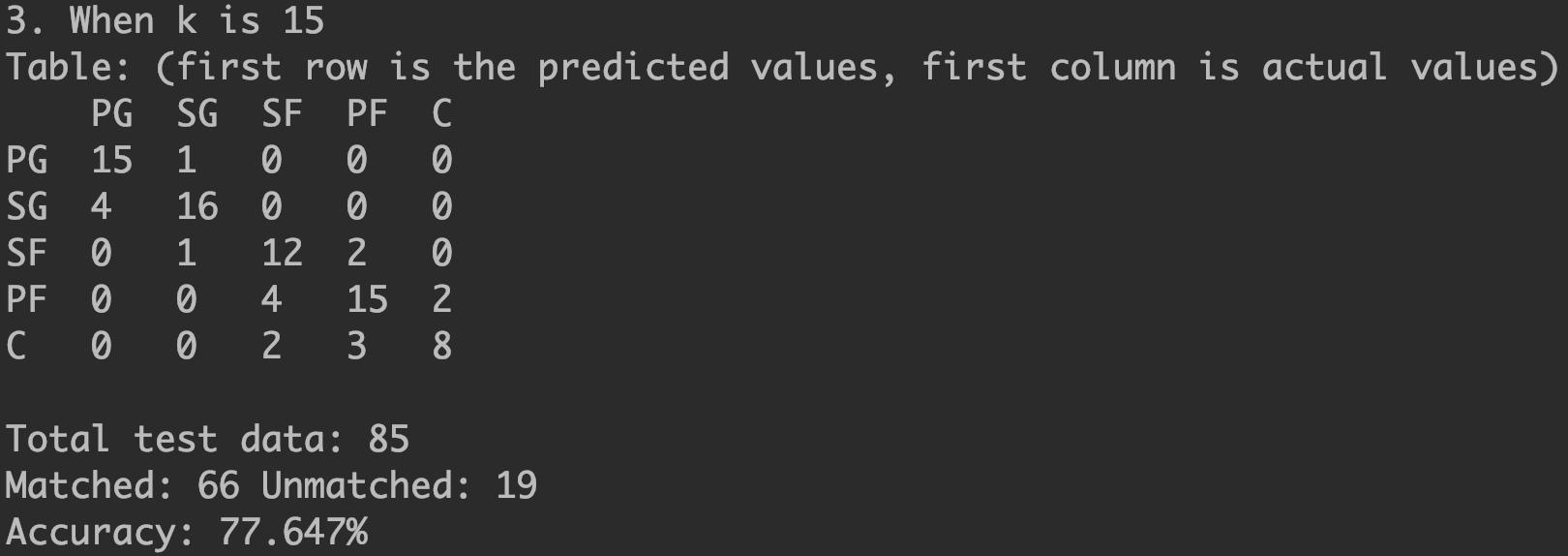


Figure 6. the result of k=15

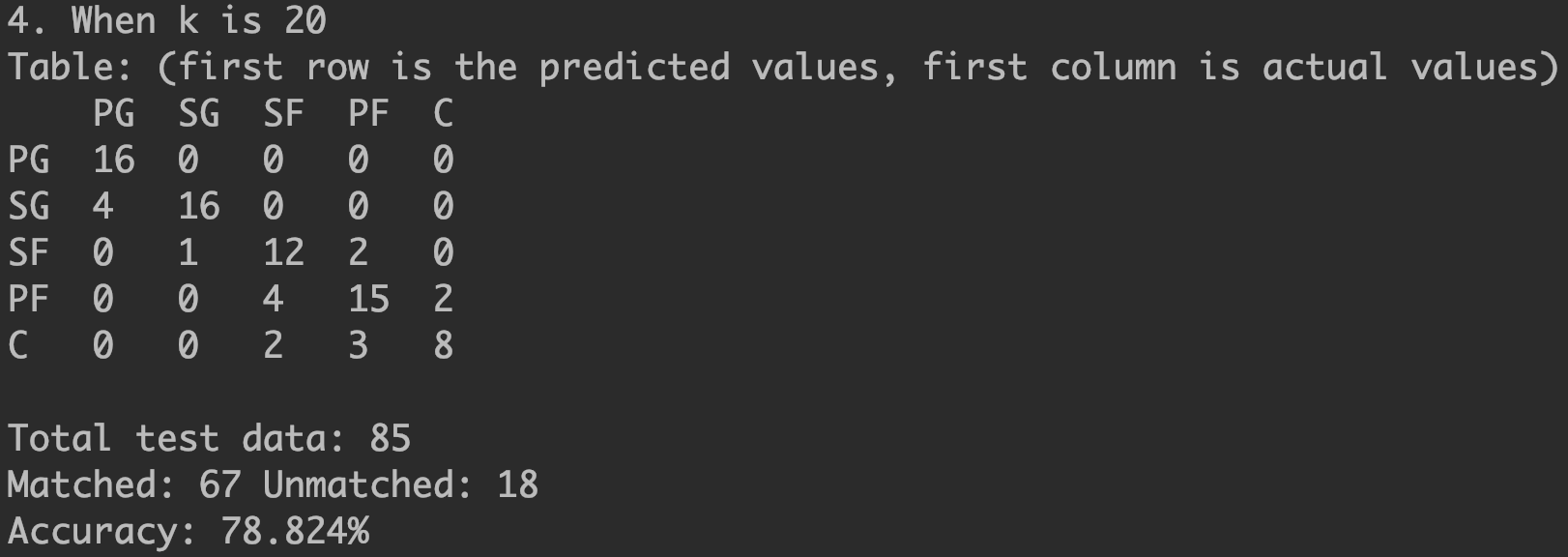


Figure 7. the result of k=20

As we can see in Figure 4, 5, 6, 7, the first row is the predicted values, and the first column is the actual values. However, you may want to know what cause the wrong prediction. As we know, some positions may have the similar data. For example, SG and SF may be hard to recognize, they may have the similar data on points or rebounds. Due to this situation, the accuracy will not be really high, which means that some players may be predicted to another position. Though the correct rate isn’t very high, we can still save much time on finding the most suitable position for a player.

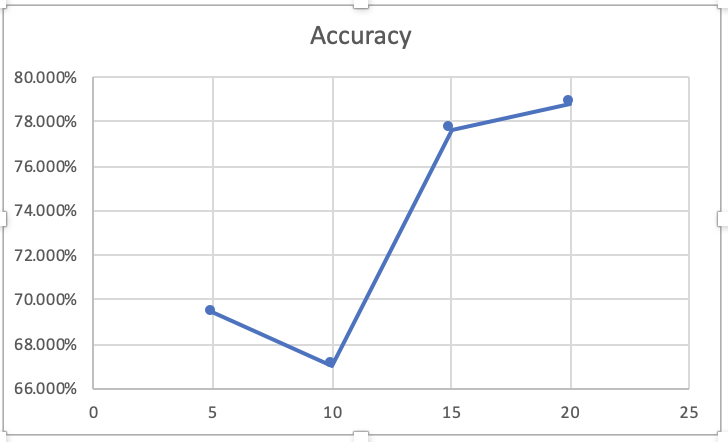


Figure 8. the tendency of accuracy when k value changes

According to Figure 8, we can see that when k value becomes bigger, the accuracy of prediction becomes higher. So, we draw a conclusion that, when k is 20, the accuracy of prediction is the best. Thus, we choose k is 20.

**4. Practical Used**

1) Imagine that we have a list of players stats of the rookies (the freshmen in NBA).  
2) We can easily find the most suitable position for a rookie.