DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

REVIEW OF SPORTS DATA MINING TECHNIQUES

by

Sezer BAYSAL

September, 2019

İZMİR

REVIEW OF SPORTS DATA MINING TECHNIQUES

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In partial Fulfilment of the Requirements for the Master of Science in
Statistics, Statistic Program

by

Sezer BAYSAL

September, 2019

İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"REVIEW OF SPORTS DATA MINING TECHNIQUES"** completed by **SEZER BAYSAL** under supervision of **ASST. PROF. DR. ENGİN YILDIZTEPE** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst. Prof. Dr. Engin YILDIZTEPE

Supervisor

Assoc. Prof. Dr. Özgür Özkaya	Assoc. Prof. Dr. A. Fırat Özdemir
(Jury Member)	(Jury Member)

Prof.Dr. Kadriye ERTEKİN

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

I would like to express my deep gratitude to Asst. Prof. Dr. Engin YILDIZTEPE for his patient guidance, enthusiastic encouragement, and useful critiques. Without his guidance and persistent help, this thesis would not have been possible.

I would like to acknowledge my thesis committee members, Assoc. Prof. Dr. Abdullah Fırat ÖZDEMİR and Assoc. Prof. Dr. Özgür ÖZKAYA, for their constructive comments and suggestions.

My grateful thanks are also extended to Burak DİLBER and Hasan SAĞIROĞLU for their positive manners during the stringent research process.

Finally, wish to express my special thanks to my parents, Nediret BAYSAL and Yücel BAYSAL, for their support and encouragement throughout my study.

Sezer BAYSAL

REVIEW OF SPORTS DATA MINING TECHNIQUES

ABSTRACT

Sports data mining is the usage of data collected on players, teams, and games for performance evaluation, player selection, score-outcome prediction, and strategy development with data mining tools and techniques. Special performance measures developed for each sports branch have an important role in sports data mining and sports statistics. Performance measures calculated for the team sports and the players can be used to predict the expectation of winning. The Pythagorean Expectation developed for this purpose was first used in baseball games. The Pythagorean Expectation, which attracts the attention of other sports branches, is adapted to other team sports with two possible outcomes such as basketball. However, Pythagorean Expectation studies are limited for sports which have three possible outcomes such as football.

In this thesis, it is aimed to investigate sports data mining studies related to various sports branches. In this context, performance measurements and sports data mining studies related to many sports branches are examined and a detailed literature review is performed. Also, it is aimed to suggest a new approach to computation of Pythagorean Expectation for football. In the application section, for the teams competing in fifteen different European football leagues, 2017/2018 season-end rankings and points are predicted using the proposed approach. The data of the last five seasons of the selected European football leagues is used as training data set. All calculations are performed in R.

Keywords: Data mining, sports statistics, Pythagorean Expectation, Christodoulou method, football, prediction of match result

SPOR VERİ MADENCİLİĞİ TEKNİKLERİNİN İNCELENMESİ

ÖZ

Spor veri madenciliği, oyuncu, takım ve oyun bazında toplanan verilerin, veri madenciliği araç ve teknikleri ile performans çözümleme, oyuncu seçimi, skor-sonuç tahmini ve strateji geliştirme amacı ile kullanılmasıdır. Her bir spor dalı için geliştirilen özel performans ölçülerinin spor veri madenciliği ve spor istatistiklerinde önemli bir rolü vardır. Takım sporları ve oyuncular için hesaplanan performans ölçüleri, kazanma beklentisini tahmin etmek için kullanılabilir. Bu amaçla geliştirilen Pisagor Beklentisi, ilk olarak beyzbol oyunlarında kullanılmıştır. Diğer spor dallarının da ilgisini çeken Pisagor Beklentisi, basketbol gibi iki olası sonuçlu diğer takım sporlarına uyarlanmıştır. Ancak futbol gibi üç olası sonuçlu sporlar için Pisagor Beklentisi üzerine yapılan çalışmalar kısıtlıdır.

Bu tezde, çeşitli spor dalları ile ilgili spor veri madenciliği çalışmalarının incelenmesi amaçlanmıştır. Bu bağlamda, birçok spor dalına ilişkin performans ölçüleri ve spor veri madenciliği çalışmaları incelenmiş ve detaylı bir literatür taraması yapılmıştır. Ayrıca, futbol için yeni bir Pisagor Beklentisi hesaplama yaklaşımı önerilmesi amaçlanmıştır. Uygulama bölümünde, on beş farklı Avrupa futbol liginde mücadele eden takımlar için 2017/2018 sezon sonu puanları ve sıralamaları önerilen yaklaşım kullanılarak tahmin edilmiştir. Seçilen Avrupa futbol liglerine ait son beş sezon verisi, eğitim veri kümesi olarak kullanılmıştır. Tüm hesaplamalar R’da yapılmıştır.

Anahtar kelimeler: Veri madenciliği, spor istatistiği, Pisagor Beklentisi, Christodoulou metodu, futbol, maç sonuç tahmini

CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

CHAPTER ONE

INTRODUCTION

Collecting and storing data have been easier and cost-effective in parallel with the progress in technology. Herewith, large amounts of data are generated in many different areas and used for different purposes (Baysal & Yıldıztepe, 2019b). The gain of previously unknown and potentially useful information from data is called as knowledge discovery.

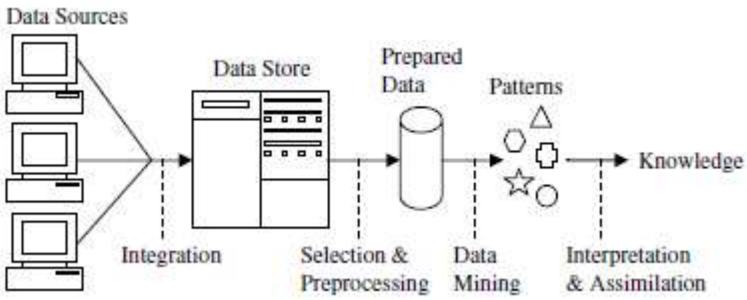

Figure 1.1 The knowledge discovery process (Bramer, 2007)

While knowledge discovery is a process from start to finish, the process consists of mainly two stages. First stage is the process of gathering data from different sources and selection. Then, knowledge discovery process is completed through the models developed with data mining techniques, which are the most critical part of the process (Bramer, 2007).

Sports data mining is defined as the use of data for performance evaluation, player selection, outcome-point prediction and strategy development by data mining tools and techniques. Sports data mining may be useful for different social environments. Audience can benefit from sports data mining as follows (Schumaker, Solieman, & Chen, 2010).

- The decision-makers of sports organizations can take more scientific and unbiased decisions by sports data mining compared to traditional methods.

- Sports managers and talent hunters from different professional sports organizations can take advantage of the analysis of various player and competitor data to maximize their investment.
- Managers in schools and universities could use data mining to better analysis and train their student-athletes and determine areas for improvement.
- The competent authorities may use the analysis of team and player activities to identify possible irregularities and frauds.
- The players of sports clubs can better understand their strengths and weaknesses and those of their challengers to gain a competitive edge.
- Students and lecturers who are interested in the academic level with sports science can gain more data-centric and scientific approach.
- Sports fans can analyze the performances of their favorite teams and players.

Sports data mining is rapidly spread and adopted due to demonstrating team player performance and helping talent hunters to discover new talents. Also, the popularity of sports data mining has increased due to the studies conducted on predicting the outcomes of sports events. In sports data mining, it is necessary to define sabermetric first. The term "Sabermetric" is derived by Bill James from the SABR, which is an abbreviation of the Society for American Baseball Research Association, founded in 1971 (Society for American Baseball Research [SABR], n.d.). Sabermetric depends on the idea of creating new statistics that better measure individual and team performances compared to the traditional statistical methods in baseball (Baysal & Yıldıztepe, 2019b). Although the idea had been proposed earlier, it was introduced by Bill James at the end of the seventies in the annual “Baseball Abstracts” booklets published by himself. James rapidly pronounced his name and increased popularity with his unusual ranking methods and new statistical performance measurements called sabermetrics. The transition from traditional statistics to sabermetrics is the result of queries and solutions on the performance criteria introduced by Bill James. James (1985) described the sabermetrics of which he developed in his later books (James, 1985). James (1980) developed Pythagorean Expectation (PE), a performance measurement metric that predicts the game-winning rate of teams in baseball (James, 1980). The PE has been widely used for baseball in subsequent years (Lee, 2015;

Tung, 2012; Valero, 2016). Several attempts have been carried out to create statistical measures similar to sabermetric in football. However, the analysis of game activity and game-based events in football are much more difficult than baseball. Because the performances of the players in football are more dependent on each other compared to baseball. The roles of the players in baseball have been set sharply; the pitcher hits the ball, the batter meets the coming ball by his bat. In football, teams can attack and defend with various strategies and the number of players. Therefore, sabermetric performance measurements were not generally used in the data mining studies performed in football. However, the success of sabermetric for measuring the team-based performance is the reason for developing sabermetric-style statistics for football. Football team managers can analyze their teams and opponents and develop strategies for future matches using the sabermetric-statistics.

In this thesis, a review of the literature is performed to examine the sports data mining techniques. Also, two new approaches are presented. The first approach is an adaptation of PE for football. The second approach is the usage of the Christodoulou method based on the PE to increase the accuracy rate of the Christodoulou method. In the second section, a detailed literature review for sports data mining is given. The PE for football and the usage of the Christodoulou method based on PE are both given in the third section. The application is presented in the fourth section which consists of two subsections. In the first subsection, it is aimed to predict the points and the ranking of the teams by using the proposed approach based on the scored and conceded goals at the end of the season. In the second subsection, it is aimed to predict the match results for a particular season by using the Christodoulou method based on PE approach. In the last section, the findings obtained in the application are interpreted.

CHAPTER TWO

LITERATURE REVIEW

In this section, some of the studies on sports statistics and sports data mining are examined chronologically in sub-headings for specific sports branches.

2.1 Baseball

In baseball, the pitcher is the player who is responsible for throwing the ball to the batter's box. The batter (hitter) is a player whose turn it is to face the pitcher. The batter tries to reach certain areas called base in the field after hitting the ball. Reaching the base/bases is possible with walking or home run. Home run is made by throwing the ball into the off-site and the value of this event is equal to four reached bases. Home run is the best possible result in favor of batter. For example, the batter can reach one, two or three base or can makes home run. So performance measurements in baseball were developed to measure the ability of the hitter who is responsible for reaching the bases and hitting the ball.

Baseball made achievements with Bill James to develop sabermetrics. But the first studies are based on previous years. Henry Chadwick, an English statistician who announced his name with cricket, adapted Batting Average which specific to cricket to baseball at the ends of the 19th century (Thorn, 2013). One of the firstborn and most common tools to measure a batter's skill at the plate, Batting Average is calculated by dividing a player's hits by his total at-bats. The batting average is a beneficial statistic for measuring the player's ability on the plate but does not take into account some features, namely, the formula ignores the type of batting (with a double, triple or home run being more precious than a single). Batting Average is calculated as follows, where the number of at-bat of the batter is the number of balls thrown to batter.

$$Batting\ Average = \frac{the\ number\ of\ hit\ score\ of\ batter}{the\ number\ of\ at - bat\ of\ batter} \tag{2.1}$$

In the forties, a former baseball player Branch Rickey and statistician Allan Roth invented a statistic called as On Base Percentage (OBP), which coaches use to determine how frequently players can reach a base (On Base Percentage, 2015). Lidsey (1963) developed a performance measurement metric called Linear Weights that can calculate the runs of the offensive player by assigning weights to each run value (one base, two base, three base, and home run) (Linsey, 1963). He proposed Linear Weights as an alternative to Batting Average. Linear Weights (LW) is calculated as follows. In the formula, while 1B, 2B, and 3B represent the number of reached related base, HR represents the number of home run.

$$LW = (0.41) \times 1B + (0.82) \times 2B + (1.06) \times 3B + (1.42) \times HR \tag{2.2}$$

It is an accomplishment that the batter meets the ball using the bat. However, different consequences may occur after the batting. The Slugging Percentage given in equation 2.3 is anonymous but is widely used in sites that provide information about baseball, such as baseball-reference.com (Slugging Percentage, 2017). This statistic takes into account the possible consequences when evaluating the player's performance. Slugging Percentage was developed on the inability of Batting Average. While Batting Average only takes into account the number of hits, Slugging Percentage also takes into account the possible results of the hits. Slugging Percentage is calculated as follows.

$$Slugging\ Percentage = \frac{(1B) + (2 \times 2B) + (3 \times 3B) + (4 \times HR)}{the\ number\ of\ at - bat\ of\ batter} \tag{2.3}$$

PE was proposed by Bill James as a performance measurement metric that predicts the team winning rate of baseball teams using runs scored (RS), runs allowed (RA) obtained from past games and the league constant of γ (James, 1980). The PE can be used to determine the teams which performing above and below the expectations by comparing the actual winning rate. PE for baseball is calculated as in Equation 2.4.

$$PE = \frac{RS^{\gamma=2}}{RS^{\gamma=2} + RA^{\gamma=2}} \tag{2.4}$$

Bill James developed Runs Created, which measures the team performance according to at-bat, hits, reached bases and walk statistics (James, 1982). Runs Created is used to analyze the ability of a team to reach the bases. James showed that Runs Created gives more accurate results compared to other statistics (Lewis, 2003). Runs Created is calculated as follows, where Walk is the walking of the hitter after the successful hitting, Hits is the number of hit scores of batter and Bases is the number of total reached base of batter.

$$Runs\ Created = \frac{(\,(Hits\ +\ Walks) \times \sum Bases)}{(At -\ Bats +\ Walks)} \tag{2.5}$$

Thorn and Palmer (1984) developed the On-Base Plus Slugging (OPS) statistics using the OBP and Slugging Percentage statistics (Thorn & Palmer, 1984). Owing to OPS, it is possible to analyze how much runs created by the batter with batting power and batting average statistics.

In addition to the batter, pitcher statistics were developed. Standard pitcher statistics have some flaws. For example, a high winning percentage for matches indicates that the player is a member of a good team. Thorn and Palmer (1993) developed Pitching Runs statistic, which allows measuring the performance of the pitcher independent from team success (Thorn & Palmer, 1993). It is possible to analyze the performance of the pitcher based on conceded runs by using Pitching Runs.

There are many other methods for baseball. Davenport and Woolner (1999) argued that the γ value of PE should be calculated separately for each team according to the balance of offensive and defensive power, and suggested that the γ coefficient in baseball should be calculated as in Equation 2.6 to obtain a smaller minimum root mean squared error (RMSE) value (Davenport & Woolner, 1999).

$$\gamma = 1.5 \times log\left(\frac{RS + RA}{NG}\right) + 0.45 \tag{2.6}$$

The beginnings of the 2000s, James and Henzler (2002) proposed the Win Shares (James & Henzler, 2002). Win Shares is a formula that takes into account too many features. Although its performance in the sabermetrics community is still a matter of controversy, Win Shares is a statistic that assigns credit to players about earning a game based on their performance.

Miller (2007), however, shown that the use of constant γ of PE as 1.82 reduces the standard error (Miller, 2007). While determining these coefficients, the criterion was defined as the minimum RMSE which, were obtained by using linear regression one of the data mining techniques.

Tung (2012) applied the PE by using the data set including the seasons from 1901 through 2009 and produced a confidence interval for the number of games predicted to be won (Tung, 2010). Lee (2015) applied the PE for the 2005-2014 seasons of the Korean Baseball League and compared the expected and actual game-winning numbers for each clubs (Lee, 2015). Inconsistency between expected and actual winning numbers, assuming the conditions of the teams originated due to an unusual distribution, was related to the coefficient of variation and standard deviation in the number of runs allowed. Valero (2016) predicted the match outcomes of the American Baseball League by using sabermetrics, including PE to assess the predictive capabilities of data mining methods (Valero, 2016). Valero, following the statistical analysis, showed that classification methods had more accuracy rate by comparison to others. The PE is given in detail in the fourth section.

2.2 Cricket

Cricket is another sporting field where performance measures are developed. Cricket is considered utterly rich in terms of statistics. (Croucher, 2000). John Buchanan, the coach of the Australian national team, pioneered many of the sabermetrics involving in the cricket between 1999 and 2007. The most well-known is “Marginal Wins”. The performances of players are evaluated through these statistics according to their positions and can also be compared with the opponent players (Schumaker, Solieman, & Chen, 2010). Before examining the literature in cricket, it is necessary to mention some basic terms specific to this game. Cricket is a game that played between two teams that do defense and offense in turn. Cricket consists of four innings. The bowler is a member of the defensive team who throws away the ball into the batting area. Fielders are the defensive players who are responsible for catching the ball hit by the batter as soon as possible. The aim of the attacking team is to scoring run/runs. There are two players in the field, who struggle for the offensive team. After a successful hit, offensive players change their sides reciprocally until the fielders catch the ball. The number of side changing is called the run score. In cricket, there are various ways of taking a wicket. For example, if the shooter (bowler) manages to drop one of the three vertical sticks standing behind the batter, the batter is dismissed and the bowler takes the wicket again. Both teams have eleven players and at least two players are required to attack and must be on the field at the same time. If ten players of the offensive team are dismissed, the first half ends because of the insufficient number of offensive players, the defensive team wins the right to attack.

Croucher (2000) ranked the batsmen and bowlers according to their performance using the data of the One-Day International Cricket Series organized between Australia, India, and Pakistan (Croucher, 2000). He analyzed the performances of the players according to two different skills; batting and bowling. He proposed and applied the Batting Index using the Batting Average and Batting Strike Rate statistics given below.

$$Batting\ Average\ = \frac{NRSB}{NDB} \tag{2.7}$$

Where NRSB is the number of run score of batter, NDB is the number of dismissing of batter.

$$Batting\ Strike\ Rate\ = \frac{NRSB\ per\ 100\ Ball\ Faced}{100} \tag{2.8}$$

Where NRSB per 100 Ball Faced is the number of run score of batter per 100 ball faced (received).

$$Batting\ Index = \ Batting\ Average \times \ Batting\ Strike\ Rate \qquad (2.9)$$

He also proposed and applied the Bowling Index (BI) using the Bowling Average and Bowling Strike Rate statistics given below.

$$Bowling\ Average \ = \frac{NCRS}{NWT} \qquad (2.10)$$

Where NCRS is the number of conceded run score and NWT is the number of wicket taken.

$$Bowling\ Strike\ Rate \ = \frac{NBB}{NWT} \qquad (2.11)$$

Where NBB is the number of balls bowled.

$$BI = Bowling\ Average \ \times \ The\ Bowling\ Strike\ Rate \qquad (2.12)$$

Finally, Croucher, who wanted to summarize the batting and bowling performance of the players with a single statistic, proposed and applied the All Rounder Index which is calculated as follows.

$$All\ \ Rounder\ Index \ = \ 100 \times \frac{The\ Batting\ Index}{The\ Bowling\ Index} \qquad (2.13)$$

He defined the predicted number of runs for each player in the final series matches as the number of runs per inning in previous matches. In the application section, the total number of runs was estimated for each team and match results were predicted according to the estimated values.

Allsopp and Clarke (2004) analyzed the defense and attack strength of the teams with multiple linear regression and examined the having home advantage effect on the games (Allsopp & Clarke, 2004). According to the results, Australia and South Africa were substantially above the other teams. Also, the results showed that home teams are generally advantageous over the away teams. Besides, they investigated how these factors affect the match results using the multinomial logistic regression techniques. They showed that having batters who can use both hands is a key to success for teams.

Scarf and Akhtar (2011), showed that the effect of certain factors such as home advantage effect on the match results vary from the first inning to the last inning (Scarf & Akhtar, 2011). They modeled the matches results using different models for each inning with multinomial logistic regression. They used the dataset as 397 matches played between 1997 and 2007.

Cricket is completed in periods called sessions. Akhtar and Scarf (2012), forecasted the probabilities of possible outcomes at the start of each session using a sequence of multinomial logistic regression models (Akhtar & Scarf, 2012). They investigated how the outcome probabilities (of a win, draw, or loss) and the effects of factors vary session by session. The factors fall into two categories, pre-match and in-play. The results showed that the score leading has a small effect on the match outcome for the first session, but is dominant later sessions; team strengths, ground effect, and home-field advantage are important predictors of a win for the first session. Knowing these variations in factor's effects can be useful for a team captain or management in considering a certain aggressive or defensive batting strategy for the coming session.

The Twenty20 (T20) series is a type of cricket in which opponents play more than one game against each other. Lemmer, Bhattacharjee, and Saikia (2014) developed two models which predict the T20 match results but were adversely affected by multiple matches (Lemmer, Bhattacharjee, & Saikia, 2014). The accuracies of the models were obtained as 52.7% and 56.8%, respectively. A consistency adjusted measure was developed to eliminate the inconsistency of the constructed models and the match results were re-predicted. As a result of adjusting, the model accuracies were increased to 70.9% and 76.4% respectively.

Vine (2016) determined the “lucky” and “unlucky” teams by comparing the predicted and actual number of winnings of cricket teams (Vine, 2016). Vine who used the 4-season data set of the Australian Cricket League, determined the coefficient of γ as 7.41 while adapting the PE to cricket. While determining these coefficients, the criterion was defined as the minimum (RMSE).

In the Cricket world, bowlers who fastmoving are believed to be more successful. Also, according to many commentators, the lower-order batters fail against fastballs. Malhotra and Krishna (2017) investigated whether the statements are true (Malhotra & Krishna, 2017). According to the One-way Anova test, fast bowlers were found to be more successful in strike-rate statistics when lower and middle-order batsmen were found to be more vulnerable against faster bowling.

2.3 Basketball

Performance measurement metrics developed for basketball are defined as ABPR. It is used as an abbreviation for the Association of Professional Basketball Researcher. ABPR is different from sabermetrics as it focuses on team-based statistics rather than individual player statistics. Unlike baseball, team-based performance statistics for basketball were developed and applied, as the performances of the players are more dependent on each other. Dean Oliver is the first person to introduce data mining techniques to basketball (Cao, 2012). Oliver (2004) followed a similar path to James and shared his creative performance measurement metrics with the world of basketball. (Oliver, 2004). Because of their contributions to basketball and baseball, we can talk about Bill James and Dean Oliver as the pioneers of basketball and baseball, respectively.

In basketball, the contribution of any player to match score is measured by Plus / Minus Rating (PMR). PMR originally was developed for the NHL (National Hockey League) (Fitzpatrick, 2019). But was also adapted to basketball (Rosenbaum, 2005).

For example, suppose that any player C from any team A is involved in the game between team A and B when $Skore_A$ = $Skore_B$. If he is taken from the game when the $Skore_A$: 5, and $Skore_B$: 3 the contribution of the player to the score is calculated as $\mathrm{PMR}_C: 5 - 3 = 2$. In basketball, the performances of the players depend on each other. The weakness of this statistic is that it is affected by many external factors. For example, the PMR of a player having talented teammates is affected positively, while the PMR of a player having untalented teammates is adversely affected.

Adjusted Plus/Minus Rating statistics were developed as a result of the inadequacy of Plus/Minus Rating which adapted from NHL to basketball. First, Wayne Winston and Jeff Sagarin developed it using the WINVAL software system in 2002 (Oliver, 2004). Adjusted Plus / Minus Rating is based on the idea of adjusting Plus / Minus Rating from inconsistent effects of various factors. Furthermore, it takes into account the home advantage.

Player Efficiency Rating (PER) measures the efficiency of the player per minute. PER is calculated by using various statistics that are negative and positive indicators of the reference player, such as rebound (positive) and missed shots (negative). Because PER provides information on efficiency per minute, it is possible to compare players even if they play in different periods. In addition, because the performance of the player depends on the team performance, each player's performance is adjusted according to his team's performance. Thus, the performances of the players from the low success teams will be measured more clearly. PER summarizes the performance of the players by using their various statistics (Hollinger, 2002). In addition to the performance statistics mentioned in this section of the thesis, many performance statistics specific to basketball are available. (Kubatko, Oliver, Pelton, & Rosenbaum, 2007).

Cao (2012) compiled five regular NBA Season as a dataset using Ruby and he made feature extraction by using the MySQL. In the application section of the study, four different models were created: simple logistic classifier, naive bayes, support vector machine, and artificial neural networks (Cao, 2012). With the created models, the match results were predicted by using various statistics belong to competitors. As a result of the accuracy analyzing, the best model was shown as a simple logistic classifier with 67.82% success. Miljkovic, Gajic, Kovacevic, and Konjovic (2010) have applied the data mining techniques which naive bayes and multivariate linear regression algorithms using Rapid Miner software (Miljkovic, Gajic, Kovacevic, & Konjovic, 2010). In the application section, k fold cross-validation method was used to determine the training and test sets and the outcomes of the 778 NBA games played in the 2009-10 season were predicted with 67% accuracy.

2.4 American Football

In American football, the game is started with a foot strike and completed in two halves. The game consists of four parts with 15 minutes of playtime. A match takes 60 minutes. It is forbidden to touch the ball with foot after the game starts, otherwise, the ball is given to the opponent team. To cover the player who carries the ball, it is free to take down the opponent players. The ball is delivered to all sides of the field by hand. When the attack player passes the opponent's goal line while the ball is in his hand, the attacking team wins 6 points. If the ball passes the opposing goal line by throwing, the team wins 3 points. One team has 45 players, but 11 players struggle against the opponent on the pitch. A team is divided into three teams that undertake separate tasks. These are an attack, special and defender team. After the opposing team loses the ball, the game stops and the attacking team enters the field on behalf of their teams. The attacking team aims to score. They attack until they lose the ball or score. Defenders enter the game after their team loses the ball. they are responsible for preventing the attacks of the opponent team and taking the ball. They defense until they take the ball or allow a score to the opponent. The special team consisting of players who have different characteristics and they enter the game in special cases.

The success of the statistics used in American football has not yet reached a sufficient level compared to statistics used in baseball and basketball because of insufficient data. Namely, while the American football teams play 16 games in one season, these numbers are 162 and 82 for baseball and basketball, respectively (Schumaker, Solieman, & Chen, 2010). However, there are also many performance statistics were developed for American football. Aaron Schatz is an American football analyst who has contributed a lot to developing performance statistics for American football. In this chapter, some performance statistics of Schatz's and data mining applications for American football will be mentioned.

Defense Adjusted Value Over Average (DVOA) is a performance statistics analyzing player and team performance (Schatz, 2006). DVOA is a statistic that takes into account some factors such as competitor quality, player position, remaining time to and of the match and match score. It is a value expressed in percentages. Thanks to DVOA, it is possible to compare players and teams based on their performance. For example, a team with a 30% DVOA value is 30% better than an average team, while a quarterback with -10% DVOA value is 10% worse than an average quarterback.

DPAR was developed for analyzing the individual player performance (Schatz, 2006). DPAR is based on the idea of comparing any player to his/her alternative. With DPAR, it is possible to examine the individual contribution of a player to the score. For instance, the fact that a player has a DPAR statistic of +5 means that player's team earns extra 5 points when the player in the game.

Players who are responsible for protecting the quarterback against the opponent's defenders are called offensive players in American football. The offensive line is consist of the offensive players ordered side-by-side. The offensive line acts as a walking wall in front of the quarterback (ball carrier) and has an important role in carrying the ball to the opponent's zone. Adjusted Line Yards (ALY) is a statistic to analyze the performance of offensive players (Schatz, 2006). ALY allows for the grading of the offensive players and their performance by taking into account how far the ball is carried in the opponent's field. As ALY takes into account events that are independent of the player, ALY does not summarize the player performance. For instance, if any offensive line player makes mistakes and the opponent team takes the ball, the other players of the offensive line negatively affected as ALY statistic.

College football is a kind of American football. Delen, Cogdell, and Kasap (2012) examined the past season statistics of college football teams and identified 28 features that are thought to affect the results of the matches (Delen, Cogdell, & Kasap, 2012). Then, to determine predictive capability, they predicted the matches results using the models they developed based on three different data mining techniques (artificial neural networks, decision trees and support vector machines). They divided the data set into training and testing. When evaluating the success of the models, they considered sensitivity, specificity and overall accuracy rate as criteria. According to the findings, the most successful model was determined as the model established using the C&RT, a decision tree algorithm.

Caro and Matchmes (2013) examined the success of the PE formula which adapted to American football as γ=2. To do this, they used seven-season historical data of 835 teams and made a retrospective match result prediction. (Caro & Machtmes, 2013). Leung and Joseph (2014) mentioned the Christodoulou method that is used in the prediction of matches and applied this algorithm to the American football data (Leung & Joseph, 2014). In the application section of their study, the distances of teams to each other in the American football League were calculated by data mining and similar teams were revealed by using the PE, Christodoulou method and other sabermetrics. They assigned the predicted scores to the matched teams according to the results of the matches between similar teams. They predicted the games results via predicted scores. The Christodoulou method is given in detail in the fourth section.

Bosch and Bhulai (2018) compiled the past matches results belong to the 2009-16 seasons having more than 130 features. As output, he selected "home win" which a categorical feature that can take values 0 and 1, and examined the correlation of other features with the output. Based on the machine learning and deep learning algorithms, he did retrospective matches results prediction by using ten features which most correlated with the output. In the application section, the success of the algorithms was compared. (Bosch & Bhulai, 2018).

2.5 Football

Academic studies have increased in recent years about the prediction of football matches results. It can be said that these studies started in the 1970s. Hill (1974) compared the actual score table with a predicted score table which was published before the start of the season by using particular statistics which belong to last season (Hill, 1974). In the application section, he showed that the prediction of the match results was not entirely chance. Maher (1982) in contrast to other studies suggesting that scored and conceded goals are fit to negative binomial, assumed that these statistics follow an independent Poisson distribution (Maher, 1982). Maher predicted attack and defense parameters for each team using the data set. Dixon and Coles (1997) used scored and conceded goal parameters which follow the Poisson distribution and developed a Poisson regression model that predicts future match results by applying a similar approach to Maher (1982) using the parameters they estimated based on data from previous matches (Dixon & Coles, 1997). This model implements a correction by increasing the probability of the 0-0 and 1-1 results due to the dependence of the number of mutual goals and decreasing the probability of results such as 1-0 and 0-1. The results showed that the developed model brings earnings when it is used as a betting strategy.

Offensive and defensive strength of the teams play an important role in the decision phase when betting. Rue and Salvasen(2001) estimated the offensive and defensive strength of teams and predicted the match results based on these parameters. They used the data set which compiled from the 1997-98 season of the English football divisions (Premier League and League 2) (Rue & Salvesen, 2001). They solved the problem in which interdependence of teams strengths using the Markov Chain Monte Carlo iterative simulation technique and predicted the defensive and offensive strength of all teams simultaneously. They predicted the next match results using the Bayesian Dynamic Generalized Linear Model, updating the team strengths parameters with the same method at the end of each match. Crowder, Dixon, Ledford, and Robinson (2002) modified the model of Dixon and Coles (Crowder, Dixon, Ledford, & Robinson, 2002). They modeled 92 teams competing in England football divisions by using a data set contains the 1992-97 season. Assuming that the defensive and offensive strengths of the teams are not stable game to game, they tried to measure these parameters using fast and powerful a new approach. They calculated home, away and draw probabilities on account of they represent the basic match outcomes. They predicted the game results via probabilities of these outcomes. When the predicted match results were compared with the actual results, it was seen that the modified method has superior accuracy rate to the Dixon and Coles's method (1997).

Except for the prediction of matches results, some researchers examined the effect of several factors on the matches results. Barnett and Hilditch (1993) analyzed the England football teams play their home matches on the artificial turf to examine the effect of artificial turf pitches on the match results (Barnett & Hilditch, 1993). As a result of the statistical analysis, it was determined that artificial turf pitches affect the match results in favor of the home teams. Clarke and Norman (1995) applied the least-squares algorithm to the data set contains 1981-91 seasons of England football teams to examine the effect of home advantage on performances of the teams (Clarke & Norman, 1995). They calculated the goal scored, goal conceded and earned points statistics belong to home and away games for each team by using the 10 seasons data of 94 England football teams from different divisions. It was shown that home advantage affects matches results in favor of the home team. Although there is no significant difference between divisions, it was reported that home advantage effect changes over the years. Dixon and Robinson (1998) developed a model that includes the defensive and offensive strength of the teams, home advantage, current score in the match, and the remaining time to the end of the game. (Dixon & Robinson, 1998). They analyzed the over 4000 past games of England football divisions using a developed model. They showed that during the match, the goals are not scored at random times, the probability of scoring for the last minutes in the matches is higher than the other periods. Also, it was shown that the number of goals is also dependent on the current score, except for the remaining time to the end of the match.

As the ordered regression methods are not affected by the dependence of the number of mutual goals, it has been preferred for prediction of matches results in a lot of studies. Forrest and Simmons (2000) investigated the predictive capacity of the bet tipsters (Forrest & Simmons, 2000). Firstly, they predicted the 1694 matches results by using an ordered logit regression model. After, they compared their accuracy rate with the accuracy rate of tipsters. According to the findings, it was determined that the tipsters had a lower accuracy rate compared to the developed model. When the tipsters were compared to each other according to their accuracy rate, it was determined that one of them had significantly higher prediction capability to the others. Just like Forrest and Simmons(2000), Kuypers (2000) investigated profit and loss status of betting odds belong to the England football betting markets for bettors. (Kuypers, 2000). He predicted matches results belong to the England football divisions by using an ordered probit regression model that contains several parameters.

Ordered regression models were also applied to explore the effect of several factors on team performance. Audas, Dobson, and Goddard (2002) explored the effect of management substitution in teams on the performance of the steams by applying the ordered probit regression model to the 25-seasons dataset belongs to England football divisions (Audas, Dobson, & Goddard, 2002). According to the findings, for the teams in which management substitution has occurred, performance decline was observed for the first three months. Also, variations were detected in the performance of these kinds of teams for a short time.

Goddard and Asimakopoulos (2004) applied a probit regression model to ten-season data set of England football divisions to predict the matches results (Goddard & Asimakopoulos, 2004). They investigated whether some factors (value of the match for a team, geographic distance between teams and whether the existing of a team in the cup competition) impact the results of the matches. Also, they researched how many last matches are significant for predicting the next match of a team. In the application section, home team's last nine home game statistics, away team's last four away game statistics were detected as significant factors that can be used for predicting any match score. If the match to be played is more important for any team, it is shown that this case affects the outcome of the match in favor of the team who cares more. Another finding is that the cup factor affects the results of the league matches. if the home team is out of the cup, this case affects the performance of the home team negatively. For away teams, the opposite is true. Buursma(2011) investigated the factors affecting the results of football matches and how to predict the results using these factors (Buursma, 2011). He set possible outcomes of the matches as home, away and draw on account of they represent the most basic outcomes. Also, He determined which factors and last how many games affect in predicting the results by using the Weka tool. Buursma calculated the probabilities of each possible outcomes for every match. Then he predicted the matches results in favor of high probability outcomes. When the data set is divided into training and testing, ten-fold cross-validation method is used to make a balanced distinction. In summary, it was shown that using certain factors belong to the last 20 matches, maximize the classification accuracy rate in predicting the matches results. And it is also shown that logistic regression is the best classification algorithm at predicting. Tax and Joustra (2015) analyzed the data from the past 13 years of the Netherlands football divisions league and determined which factors affect to the game result. (Tax & Joustra, 2015). They compared the predictive capability of the hybrid models with the basic models which contain only betting odds. The hybrid models include several factors that are thought to affect the game results compared to basic models. According to the McNemar test, no significant difference was found between the most accurate two models from hybrid and basic models.

Football Fans interest in the outcome of their favorite team's next games, while the coach and trainers consider being ready for the next match with the best squad and strategy. Pre-match tactics and team formation such as 4-4-2, affect the result of the match significantly. Position analysis is used to determine the best positions of the players and the most successful players for the particular position. By the gained information, the teams come to the pitch with the most appropriate formation. Memmert, Lemmink, and Sampaio (2017) analyzed the formation data of the teams Barcelona and Bayern Munchen to determine the optimum formations as a tactic for both teams (Memmert, Lemmink, & Sampaio, 2016).

Besides England leagues, there are some studies about the prediction of matches results for the Turkey Leagues. Kıran and Esme (2018) determined the previous matches similar to the matches to be predicted via the k-nearest neighbor method

(Esme & Kiran, 2018). Then, they predicted the results of the next matches for the Turkey Super League using the odds of past matches. The odds used for predicting results were created by bookmakers. They calculated odds for possible outcomes (home win, draw and home lose) for each match to be predicted. they compared the previous odds with their calculated odds and determined which matches have higher odds than calculated ones. They defined this kind of matches as profitable matches and predicted the results of these games using their odds. In the application section, the 2015/2016 match results and betting odds were compiled from a sports website as a data set and the remaining match results of the season were estimated using the matches of the first half of the league (1-17 weeks). As a result, the accuracy rate was found to be 55.6% for single-outcome prediction and 78.4% for two-way (double chance) prediction.

2.6 Other Branches

In the National Hockey League (NHL), Plus / Minus Rating is used to measure the defensive performance of players compared to other players. This statistic is also adapted to different sports such as American football (Rosenbaum 2004). Originally it was developed by the management of the Montreal Canadians team, it was used to analyze the performance of players. Later, other teams began to use it, and in 1968 it became one of the official statistics used by the NHL. Plus / Minus Rating is a statistic that is calculated per game and provides general information about player success at the end of the season. Its weakness is influenced by many factors such as the defensive and offensive strengths of the team. For example, if any team is mediocre, no matter how skilled any player of the team is, the player's Plus / Minus Rating statistic is negatively impacted (Schumaker, Solieman, & Chen, 2010)

Maszczyk et al. (2012) aimed to predict their race performance by using various features of swimmers. In this context, they used nonlinear regression and neural models (Maszczyk et al., 2012). When predicted and actual performances were compared, it was shown that neural models have better results.

Coaches can benefit from data mining techniques when planning training sessions. Li and Zhang (2012) proposed a model based on various clustering algorithms and grouped players according to their specific characteristics using training statistics (Li & Zhang, 2012). In their study, they noted that taking into account group differences would help team managers when planning training sessions.

Maszczyk et al. (2014) aimed to predict the performances of javelin throwers using their regression and neural models (Maszczyk et al., 2014). In this context, they compiled the data set of 116 athletes. Of the 116 athletes, 20 were allocated to verify the model. Before starting the statistical analysis, they first examined the suitability of the players' data for normal distribution and equal variability with normality and Shapiro-Wilk tests, respectively. The performances of the athletes were estimated using models that take into account various features. A neural model was determined to be the most successful model with the 12.68-meter absolute error.

Many factors may affect the outcome of an ice hockey match. Gu, Saaty, and Whitaker (2016) identified the factors that play an active role in the game results by applying the Support Vector Machine (SVM) algorithm to the 2014-2015 regular-season of the National Hockey League (Gu, Saaty, & Whitaker, 2016). They developed a model based on a combination of particular factors and human judgments and predicted playoff results as win-lose. The accuracy rate of the proposed model was determined as 77.5%.

Club managers and coaches strive to develop a model that determines the strategy to optimize team performance. Bunker and Thabtah (2017) proposed a framework for knowledge discovery in sports, from data collection to model building. (Bunker & Thabtah, 2017).

CHAPTER THREE

SPORTS DATA MINING TECHNIQUES

In the following sections, some of the sports statistics and sports data mining approaches are investigated in detail. In sections 3.5 and 3.6, two new approaches are proposed for football.

3.1. Pythagorean Expectation

Pythagorean Expectation (PE) was proposed by Bill James (James, 1980) as a performance measurement metric that predicts the team winning rate of baseball teams using runs scored (RS), runs allowed (RA) obtained from past games and the league constant of γ. The PE can be used to determine the teams which performing above and below the expectations by comparing the actual winning rate. PE for baseball is calculated as in Equation 3.1.

$$PE = \frac{RS^{\gamma}}{RS^{\gamma} + RA^{\gamma}} \tag{3.1}$$

PE is typically used in the middle of a season to predict the standings for the end of a season. For example, if a team wins more than the predicted in the halfway through a season, analysts claim that the team will complete the remaining half of the season with fewer winnings than the predicted (Miller, 2007). The value of constant γ in original formula has been set to 2.0 by James. Miller (2007), however, has shown that the use of constant γ as 1.82 reduces the standard error.

Various applications have been suggested for also baseball. Davenport and Woolner (1999) argued that the γ value should be calculated separately for each team according to the balance of offensive and defensive power, and suggested that the γ coefficient in baseball should be calculated as in Equation 3.2 to obtain a smaller RMSE value (Davenport & Woolner, 1999).

$$\gamma = 1.5 \times log\left(\frac{RS + RA}{NG}\right) + 0.45 \qquad (3.2)$$

Where RS is the number of runs allowed, RA is the number of runs allowed and NG is the number of games played by the team.

PE has attracted the attention in other sports branches due to its impact on baseball. Different γ values have been attained in the studies conducted using PE in sports branches such as American football, cricket, basketball and ice hockey. Some of these studies are summarized in Table 3.1 (Baysal & Yıldıztepe, 2019b).

Table 3.1 Recommended γ values for different sports branches

Sports	**γ**	**Source**
Baseball	1.82	(Miller, 2007)
American Football	2.37	(Schatz, 2003)
Basketball	14 13.91	(Oliver, 2004) (Zminda & Dewan, 1993)
Ice Hockey	1.927	(Cochran & Blackstock, 2009)
Cricket	7.41	(Vine, 2016)

3.2 Pythagorean Expectation in Football

PE in sports which have two possible outcomes (win - or - loss) which mentioned in the previous section have been applied only with the changes made in the γ coefficient. However, the teams acquire points below the predicted if the original formula is applied directly without any arrangement in football which is a game that can result in a tie (Eastwood, 2012; Hamilton, 2011).

Hamilton (2011), considering the possible tie outcome in football, predicted the points earned per game instead of predicting the winning ratios of teams using the extended Pythagorean Expectation (Hamilton, 2011). Hamilton tried to overcome the problem that PE was only applicable to sports with two possible outcomes by calculating the probability of winning and draw for each team. Hamilton calculated the predicted point per game (PPPG) for team X by using Equation 3.3 where X representing a team playing in the league and Y representing the opponents.

$$PPPG = 3 \times P(X > Y) + P(X = Y) \tag{3.3}$$

He used the least squares (LS) algorithm to express the scored and conceded goals distributions with a three-parameter Weibull distribution. However, Hamilton's method has not widely used due to intensive mathematical and statistical procedures.

Eastwood took the draw possibility into account and adopted the original PE to a football game which has 3-point for a win, 1-point for a draw, 0-points for a loss (Eastwood, 2012). Eastwood, instead of calculating the winning possibility of the teams, calculated the PPPG multiplying the average point per game (APPG) by the probability of gaining points. The equation developed by Eastwood to calculate the PPPG for each team is given below.

$$PPPG = \frac{G^{1.22777}}{G^{1.072388} + CG^{1.127248}} \times 2.499973 \tag{3.4}$$

In the Equation 3.4; G is the number of goals scored, CG is the number of goals conceded and the APPG is 2.499973.

Hamilton (2011) determined the γ value using a single season data and found RMSE value as 3.81. Eastwood (2012) obtained lower RMSE values by using the data collected from ten seasons. The adaptation of Eastwood (2012) seems like much straightforward and more practical than the adaptation formula of Hamilton (2011). However, Eastwood developed and implemented the formula only over the English Premier League data (Baysal & Yıldıztepe, 2019b).

3.3 Christodoulou Method

The Christodoulou method analyzes past games in sports, such as football and American football, to predict future match results through score prediction. The Christodoulou method uses four statistics belong to teams competing in any league based on their past results. For any team these statistics are goal scored per game (GSPG), goal conceded per game (GCPG), offensive strength (OS) and defensive strength (DS). The league statistics GSPG_{League} and GCPG_{League} are equal. OS indicates how many scores the reference team does per game in proportion to GCPG_{League}. For example, the OS of a team in which scores 60 points per game (GSPG = 60) is calculated as 60/40 in a league where GCPG_{League} is 40. DS indicates how many scores the reference team concede per game in proportion to GSPG_{League}. For example, the DS of a team in which concede 20 points per game (GCPG = 20) is calculated as 20/40 in a league where GSPG_{League} is 40. Christodoulou method aims to predict the results of the games with these statistics. For any team-A, mentioned statistics are calculated as follows.

$$
\begin{aligned}
OS_A &= \frac{GSPG_A}{GCPG_{League}} \\
DS_A &= \frac{GCPG_A}{GSPG_{League}}
\end{aligned}
\tag{3.5}
$$

$$
\begin{aligned}
GSPG_A &= \frac{Total\ goal\ scored_A}{Total\ Game_A} \\
GCPG_A &= \frac{Total\ goal\ conceded_A}{Total\ Game_A}
\end{aligned}
\tag{3.6}
$$

Result of any match can be estimated as a win, draw or lose by using estimated mutual scores of two competitors. Using the above statistics, the game score where any A and B teams match can be estimated as follows.

$$Goal_A = OS_A \times GCPG_B + GSPG_A \times DS_B$$
$$Goal_B = OS_B \times GCPG_A + GSPG_B \times DS_A \quad (3.7)$$

3.4. Leung & Joseph's Approach

Leung and Joseph (2014) developed a new approach to predict the result of a match between any teams A and B from any American football league. This approach, unlike most approaches in the literature, is based on the idea of using performance statistics of teams similar to teams A and B, respectively, instead of using performance statistics of A and B teams. Any other team may be similar to only A or B only. As a result of the determination of the teams similar to A and B teams, the prediction is made about the result of A-B match based on the past match results played between teams similar to A and B respectively. To determine the teams that are similar to a given teams A and B, their approach represents each team as a point on a 4-dimensional space representing four different statistics: Rate Percentage Index, Pythagorean Wins, Offensive Strategy, and Turnover Differential.

3.4.1 Rate Percentage Index (RPI)

A sports season is a long period in which all competitors match each other. After each competition, the statistics of the league teams are updated according to the results of the last match (win, draw or lose). The end-of-season league table provides clear and accurate information about the overall performance of the teams about the entire season. However, before all possible matches are completed, evaluating the performance of the teams through the current league table may lead to biased decisions. For example; Assume any sports branch where there are 10 competitors teams and the teams play a total of 20 games. Let compare any A and B teams based on the current league table at the end of the 3rd competition week. Let team A be the leader in the league when team B is the second. Comparing the two teams according to the current league table may lead to biased decisions if team A has reached this rank against weak teams and team B against strong teams. For this reason, Leung and Josep (2014) used RPI as part of their approach, as it measures the performance of the teams more clear. The RPI statistic for any team A is calculated as follows.

$$RPI_A = (0.25 \times WR_A) + (0.5 \times OWR_A) + (0.25 \times OOWR_A) \quad (3.8)$$

Where WR is the winning ratio, OWR is the opponent's WR and the last one, OOWR is the opponent's opponents WR. Let A any team from any American football league. Using past statistics, WR_A, OWR_A and $OOWR_A$ are calculated as follows.

$$WR_A = \frac{Total\ Win_A}{Total\ Games_A} \quad (3.9)$$

$$OWR_A = \frac{Sum\ of\ \ WR\ of\ Opponents}{Total\ Games_A} \quad (3.10)$$

$$OOWR_A = \frac{Sum\ of\ OWR\ of\ Opponents}{Total\ Games_A} \quad (3.11)$$

When calculating the OWR value for any A team, the match results played between opposing teams and the A are not taken account. Consequently, the WR of any team is calculated differently for use in the OWR function and differently for use in the RPI formula. Namely, winning ratio of any B team; is calculated as WR_B for use in formula RPI_B, is calculated as WR_{AB} for use in formula OWR_A. Furthermore, the WR of the opposing teams is added to the OWR_A formula as much as the number of games played between the opposing teams and the A-team. The same rule is applied to the *OOWR* formula. An example is given below to show how RPI is calculated for each team as follows.

3.4.2 An Example for RPI

Assume that A, B, C, and D any teams that compete in any American football league. An example of the results of the matches played between these teams is given in Table 3.2.

Table 3.2 Past games played between A, B, C and D teams

Home	**Score**	**Away**	**Score**
A	62	**B**	59
A	86	**C**	65
D	70	**A**	74
B	65	**A**	61
C	84	**D**	65
D	49	**B**	63

Using the historical match results given in the Table 3.2, RPI statistics for each team were calculated step by step as follows.

Step 1 WR Statistics;

$$WR_A = \frac{Total\ Win}{Total\ Game} = \frac{3}{4} = 0.75 \qquad WR_B = \frac{Total\ Win}{Total\ Game} = \frac{2}{3} = 0.67$$

$$WR_C = \frac{Total\ Win}{Total\ Game} = \frac{1}{2} = 0.5 \qquad WR_D = \frac{Total\ Win}{Total\ Game} = \frac{0}{3} = 0$$

Step 2 OWR Statistics;

$$OWR_A = \frac{WR_{AB} = \frac{1}{1} + WR_{AB} = \frac{1}{1}\ WR_{AC} = \frac{1}{1} + WR_{AD} = \frac{0}{2}}{4} = 0.75$$

Note that when calculating the WR values of the opposing teams, matches against team A was not taken into account. In addition, since team B and team A matched twice, team B's win ratio (WR_{AB}) was added to the function twice. The total of all win ratios was divided by the total number of games (4) played by team A.

$$OWR_B = \frac{WR_{BA} = \frac{2}{2} + WR_{BA} = \frac{2}{2} + WR_{BD} = \frac{0}{2}}{3} = 0.67$$

$$OWR_C = \frac{WR_{CA} = \frac{2}{3} + WR_{CD} = \frac{0}{2}}{2} = 0.33$$

$$OWR_D = \frac{WR_{DA}=\frac{2}{3} + WR_{DB} = \frac{1}{2} + WR_{DC} = \frac{0}{1}}{3} = 0.39$$

Step 3 OOWR statistics;

$$OOWR_A = \frac{OWR_B=0,67+OWR_B=0.67+OWR_C=0.33+OWR_D=0.39}{4} = 0.51$$

Note that since team B and team A matched twice, team B's OWR (OWR_{AB}) was added to the function twice. The total of all OWR was divided by the total number of games (4) played by team A.

$$OOWR_B = \frac{OWR_A = 0.75 + OWR_A = 0.75 + OWR_D = 0.39}{3} = 0.63$$

$$OOWR_C = \frac{OWR_A = 0.75 + OWR_D = 0.39}{2} = 0.57$$

$$OOWP_D = \frac{OWR_A = 0.75 + OWR_B = 0.67 + OWR_C = 0.33}{3} = 0.59$$

Step 4. RPI statistics;

$$RPI_A = (0.25) \times 0.75 + (0.5) \times 0.75 + (0.25) \times 0.51$$

$$RPI_A = 0.69.$$

Similarly, also for other teams, RPI statistics can be calculated as follows.

$RPI_B = 0.66$, $RPI_C = 0.43$, $RPI_D = 0.34$.

According to the results, when A and B teams are compared in terms of performance; although $WR_A = 0.75$ ve $WR_B = 0.67$, the RPI statistic of the two teams is significantly closer: $RPI_A = 0.69$ and $RPI_B = 0.66$.

3.4.3 Pythagorean Expactation for American Football

Pythagorean Expectation (James, 1980) was adapted for American football by Schatz (2003) as evaluation 3.12. Based on the Point For (PF) and Point Against (PA) statistics of the teams, PE provides an insight into the number of matches that teams must win. It was used by Leung and Joseph (2014) to determine similar teams.

$$PE = \frac{PF^{2.37}}{PF^{2.37} + PA^{2.37}} \tag{3.12}$$

3.4.4 Offensive Strategy

In American football, the offense is carried out by Passing Attempt (PA) or Rushing Attempt (RA). Offensive strategy (OS) allows to examine the pass and rush preference tendency of any team.

$$OS = Number\ of\ PA\ /\ Number\ of\ \ RA \tag{3.13}$$

3.4.5 Turnover Differential

In American football, the turnover is the ball's change between two opposing teams. So, the turnover differential of a team defined as the difference between the number of times in which the team takes away the ball and the number of times in which the team gives away the ball. It can be calculated as follows.

$$Turnover\ Differential = Take\ aways - Give\ aways \tag{3.14}$$

Leung and Joseph (2014) used the four performance statistics mentioned above (RPI, PE, Offensive Strategy and Turnover Differential) for determining the similar teams. Then, by calculating the minimum and maximum values for each, they normalized these statistics. For these statistics, 0 was obtained as minimum value despite 150, 100, 50 and 50 were obtained as maximum value for RPI, PE, OS, and TD respectively. In other words, results of the normalization process are RPI ∈ [0, 150], PE ∈ [0, 100], OS ∈ [0, 50], and TD ∈ [0, 50]. To analyze the success of the their approach, they used a data set contains 9-season statistics of past games. As a result of their application, the accuracy rate of the proposed approach was found to be 91.43%. Their approach can be explained with the following example.

3.4.6 An Example for Leung and Joseph's Approach

To explain their approach, an example was performed using the data contains match results of imaginary teams competing in any American football league. Assume that G and H are opponent teams. The teams (J, K, and L) similar to G and H were determined using the four mentioned statistics. The result of the G-H match was predicted by using the past games played between these determined teams. Past game results were given in Table 3.3. According to the *RPI, PE, Offensive Strategy and Turnover Differential* statistics the distances between the teams are $d_{J-G} = 2,\ d_{J-L} = 1,\ d_{K-G} = 4,\ d_{K-L} = 1$ and $d_{L-H} = 2$.

Table 3.3 Past matches of K,J and L which similar to G or H

Similar to G	**Past Match Score (PMS)**	**Past Match Score (PMS)**	**Similar to H**
J	24	17	L
K	10	35	L

- By analyzing the J-L match;

$$Predicted\ Score\ for\ G = \frac{J's\ score\ in\ PMS_{J-L}}{d_{J-G} + d_{J-L} + \ d_{L-H}} = \frac{24}{5}$$

$$Predicted\ Score\ for\ H = \frac{L's\ score\ in\ PMS_{J-L}}{d_{J-G} + d_{J-L} + \ d_{L-H}} = \frac{17}{5}$$

- By analyzing the K-L match;

$$Predicted\ Score\ for\ G = \frac{K's\ score\ in\ PMS_{K-L}}{d_{K-G} + d_{K-L} + \ d_{L-H}} = \frac{10}{7}$$

$$Predicted\ Score\ for\ H = \frac{L's\ score\ in\ PMS_{K-L}}{d_{J-G} + d_{J-L} + \ d_{L-H}} = \frac{35}{7}$$

If we take account the past two matches, predicted score for both teams can be calculated as below.

$Predicted\ Score\ for\ G = (\frac{24}{5} + \frac{10}{7}) \times \frac{1}{2}$ = 3.11,

$Predicted\ Score\ for\ H = \left(\frac{17}{5} + \frac{35}{7}\right) \times \frac{1}{2} = 4.2$. So, H is the predicted winner for a G-H match.

3.5 Modified Pythagorean Expectation

This section outlines the proposed approach to adapt PE to football (Baysal & Yıldıztepe, 2019b). The proposed formula, unlike baseball, is aimed to predict the expected points per game of the teams instead of winning possibilities. In order to calculate the PE, the number of goals scored and conceded by the reference team in the league have to be known. In addition, the exponential coefficient γ and the average points distributed per game in the league (APDG) should also be determined. The most important difference between recommended approach and Eastwood's formula is the usage of the γ coefficient. Eastwood's formula uses three different γ coefficients. The proposed formula is developed with a data-oriented approach. The PE equation, which calculates the expected points per game for each team, is written as follows with the determination of the required coefficients γ and APDG:

$$\text{PE} = \frac{Goal_S{}^{\gamma}}{Goal_S{}^{\gamma} + Goal_C{}^{\gamma}} \times APPG \tag{3.15}$$

$Goal_S$ represents the number of goals scored and $Goal_C$ is the number of goals conceded for any team.

Since football is not a game with two possible outcomes, the APPG value cannot be taken as 3 points. Considering that there are three possible outcomes in football, the ratios of the draw and win-loss in the leagues must be determined (Equation 3.16). The APPG is calculated by Equation 3.17 using the statistics for the total game played in the league (TGP), total win (TW), total draw (TD) and total loss (TL).

$$Ratio_{win-loss}: \frac{TW+TL}{TGP}$$
$$Ratio_{draw}: \frac{TD}{\mathrm{TGP}} = 1 - \frac{TW+TL}{\mathrm{TGP}} \tag{3.16}$$

$$APPG = 3 \times (Ratio_{win-loss}) + 2 \times (Ratio_{draw}) \tag{3.17}$$

The γ coefficient in PE (equation 3.15) for football can be predicted by simple linear regression method (SLR). The SLR provides a linear function that models the relationship between the dependent and independent variables with the least squares (LS) algorithm.

In the proposed approach, PE is calculated with the Equation 3.15 by using the various gamma values between 1 and 2 for all teams in the league then PE is multiplied by the number of matches played in order to predict end-season points. A regression model is created by using SLR where the predicted score as the explanatory variable and the actual score as the response variable. Consequently, the SLR models are generated as much as the number of γ tested. The optimum γ coefficient is determined by examining the RMSE obtained in the models and the coefficient of determination R^2.

3.6 Christodoulou Method Based on Modified Pythagorean Expectation

In football which is one of the most popular sports of today, it is possible to predict the results of the matches by analyzing the data collected about teams, players and the game. With the performance measurement metrics such as Pythagorean Expectation, the performances of the teams can be measured and the teams can be compared with each other. The outcome of any game can be predicted as a result of these comparisons.

The common characteristic of these two performance measurement metrics is that they take into account "scored" and "conceded" goals of teams per game. PE can be used to predict the end-of-season points and rankings of teams as well as being used to predict match result. By estimating the end-of-season points of the teams, real-predicted score difference information is obtained for each team. Based on this information, matches that meet certain criteria can be detected. Christodoulou method can give more successful consequence for these matches.

The algorithm that determines the games to be predicted with the Christodoulou method for the next season as follows where k= 0,1,2, … ,9 a positive threshold value determined by the researcher, d_{Home} and d_{Away} are the actual-expected score differences belongs to the last season of any team from any game: $d_{Home} \geq k \text{ and } d_{Away} \leq -k \text{ or } d_{Away} \geq k \text{ and } d_{Home} \leq -k$. This approach, in which the threshold value is selected by the researcher, can be defined as a supervised approach. In order to illustrate the proposed approach, an example is given below.

3.6.1 An Example

In this section, a simple application is made to show how the proposed method determines the matches that appropriate for prediction. In the application, last season statistics of the teams were used and the suitable matches to predict from the first two weeks of the current season were determined using different threshold values. Depending on the different threshold values, changing in the number of matches appropriate for prediction was observed. In this section, there is no prediction for matches results. Only the matches that appropriate for prediction were determined with the help of the proposed approach. The relationship between the number of predicted games and accuracy rate will be examined in section 4.2 using the real data.

Assume that A, B, C, and E are any teams competes in any football league. The expected point, actual point and the difference for these statistics are given below for mentioned team's last season.

Table 3.4 Last season statistics for A, B, C, team D

Teams	Actual Point	Expected Point	Difference $(\boldsymbol{d})$
A	77	75	+2
B	65	70	-5
C	62	65	-3
D	54	50	+4

Assume that A_{Home} - B_{Away}, C_{Home} - D_{Away} and A_{Home} - D_{Away}, B_{Home} - C_{Away} matches indicate the week 1 and week 2's matches respectively for the current season.

- For $k = 0$;

Considering Week 1's matches;

For A_{Home} - B_{Away}: Yes, $d_{Home} \geq 0$ and $d_{Away} \leq 0$

For C_{Home} - D_{Away}: Yes, $d_{Away} \geq 0$ and $d_{Home} \leq 0$

Considering Week 2's matches;

For A_{Home} - D_{Away}: Neither $d_{Home} \geq 0$ and $d_{Away} \leq 0$ nor $d_{Away} \geq 0$ and $d_{Home} \leq 0$

For B_{Home} - C_{Away}: Neither $d_{Home} \geq 0$ and $d_{Away} \leq 0$ nor $d_{Away} \geq 0$ and $d_{Home} \leq 0$

According to the proposed approach with $k = 0$, matches to be predicted are determined as A_{Home} - B_{Away}, C_{Home} - D_{Away} matches.

- For $k = 3$;

Considering Week 1's matches;

For A_{Home} - B_{Away}: Neither $d_{Home} \geq 3$ and $d_{Away} \leq -3$ nor $d_{Away} \geq 3$ and $d_{Home} \leq -3$

For C_{Home}-D_{Away}: Yes, $d_{Away} \geq 3$ and $d_{Home} \leq 3$

Considering Week 2's matches;

For A_{Home} - D_{Away}: Neither $d_{Home} \geq 3$ and $d_{Away} \leq -3$ nor $d_{Away} \geq 3$ and $d_{Home} \leq -3$

For B_{Home} - C_{Away}: Neither $d_{Home} \geq 3$ and $d_{Away} \leq -3$ nor $d_{Away} \geq 3$ and $d_{Home} \leq -3$

By using the proposed approach for $k = 3$, game to be predicted is determined as C_{Home} - D_{Away}. No matches are determined for $k = 6$. There is a trend that if the threshold value increases, the number of matches suitable for prediction decreases. Another application will be made for the proposed approach in the application section to examine the change in prediction accuracy rate in parallel to this trend and to determine the optimal threshold value.

CHAPTER FOUR

APPLICATION

In this section, two applications are performed. Firstly, a PE formula is proposed for football, based on the training set, by using the proposed method in section 3.5. The success of the PE formula is evaluated by using the test data. Secondly, the success of the Christodoulou method used depend on the PE formula proposed in the first subheading (section 4.1) is examined for the prediction of the results of the games. In this context, the results of all matches of the 2017-18 season are predicted by the Christodoulou method. Then, the difference between the actual and expected points (season-end) of the teams are examined by using the PE formula and 2017-2018 season data, and the results of the matches (meet certain criteria) of the mentioned season are predicted by the Christodoulou method. The observed prediction accuracy rate for all games is compared to observed accuracy rate for games that meet certain criteria. Thus, the accuracy rate of the proposed approach given in section 3.6 is measured.

4.1 Real Data Application – 1

In this section, in line with the proposed approach in section 3.5, a PE formula for football is developed. To develop a formula appropriate for all European football, the data compilation process was made using fifteen top division leagues of Europe. These leagues are thought to represent European football. According to the proposed approach, a PE formula was developed for football using the training data set. Accuracy rate of the developed formula was analyzed using the test data set. For the SLR, the success criterion was determined as minimum RMSE and maximum coefficient of determination (R^2), then linear models were developed for different γ values during the determining the best formula stage.

The data of fifteen European football leagues belong to the six seasons between 2012-2013 and 2017-2018 seasons used in the study were compiled from the mackolik.com website (Mackolik, 2018). The top division leagues used in the application belong to countries of Turkey, Italy, Germany, Spain, France, Holland, England, Belgium, Austria, Croatia, Denmark, Czech Republic, Portugal, Romania, and Scotland. The league tables used in the study include the number of games played for each team (G), the number of wins (W), the number of draws (D), the number of losses (L), the goals scored (S), the goals conceded (C) and the end of season points (P). Play-offs, canceled games, and cup games have not been included in the data used. The league tables belong to past five years (2012 – 2017) of fifteen European football leagues were used as training data in the application (Baysal & Yıldıztepe, 2019b). The data for 152 teams played during five seasons (2012-17) in the league (never dropped out) were used in the training data set. The data of 244 teams in the 2017-2018 season were separated as test data. All calculations are performed in R statistical programming language. A small excerpt of the data is shown in Table 4.1.

Table 4.1 An example of a league table

Teams	**Country**	**G**	**W**	**D**	**L**	**S**	**C**	**P**
Athletic Bilbao	Spain	190	84	43	63	263	233	295
Pandurii Targu Jiu	Romania	154	64	38	52	222	192	224
Nice	France	190	83	46	61	252	220	295
Zulte Waregem	Belgium	150	68	40	42	241	209	244
AZ Alkmaar	Holland	170	72	40	58	299	265	256
Schalke 04	Germany	170	74	40	56	259	222	262

Firstly, the win-loss and draw ratios were calculated by using Equation 3.16:

$$Ratio_{win} = 0.7471$$

$$Ratio_{draw} = 0.2528$$

APPG was computed with Equation 3.17 as follows:

$$APPG = 3 \times (0.7471) + 2 \times (0.2528)$$

$$APPG = 2.7471$$

The most successful results were obtained in the $1 \leq \gamma \leq 2$ range in our preliminary study. Therefore, PE of each team was calculated with Equation 3.15 using 2.7471 as APPG for eleven different γ coefficients between 1.0 and 2.0. The calculated PE values for each team are multiplied by the number of games played, and eleven distinct points are predicted for the total points of the teams for the five seasons. Eleven simple linear regression models were created to find the most appropriate γ value, where the predicted points were the independent variable (x_i) and the actual points were the dependent variable (y_i). The RMSE and R^2 values for the obtained models are shown in Table 4.2 and Figure 4.1.

Table 4.2 RMSE and R^2 values of models obtained with different γ coefficients

γ	RMSE	R^2
1.0	11.21613	0.974433
1.1	10.55446	0.977361
1.2	10.28528	0.978501
1.3	10.31085	0.978394
1.4	10.54836	0.977387
1.5	10.93303	0.975708
1.6	11.41728	0.973508
1.7	11.96756	0.970893
1.8	12.56064	0.967937
1.9	13.18052	0.964694
2.0	13.81608	0.961207

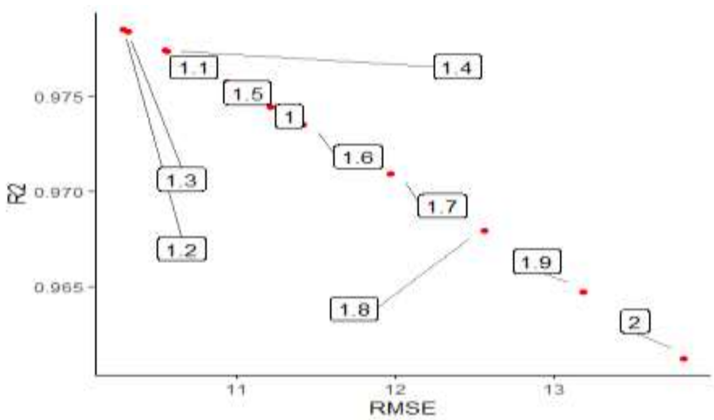

Figure 4.1 RMSE and R^2 values of models obtained with different γ coefficients

The results revealed that the lowest RMSE and the highest R^2 values were obtained when γ=1.2. The recommended PE of European football leagues can be calculated by the Equation 4.1 using the specified value of the coefficient where $Goal_S$ and $Goal_A$ are goal scored and goal conceded respectively. The γ value was recommended as 1.3 in another study using twelve European leagues (Baysal & Yıldıztepe, 2019a).

$$PE = \frac{Goal_S^{1.2}}{Goal_S^{1.2} + Goal_C^{1.2}} \times 2.7471 \tag{4.1}$$

The SLR used with the LS algorithm is a parametric statistical method that requires some assumptions. Thus, initially, the required assumptions must be checked. For this purpose, normality, and independence of the residuals obtained by the model were examined. Secondly, variance homogeneity was investigated. The results of the analyses showed that the assumptions were satisfied. Required tests for the validity of the model and coefficients were also performed. The significance level of all hypothesis tests was accepted as 0.05.

4.1.1 Results

The data of the 2017-2018 season were used for evaluation. In the first stage, the differences between the predicted and actual points were examined to measure the success of the proposed approach (Table 4.3).

Table 4.3 An example for end of season actual points and predictions

Teams	**Actual Point**	**Predicted Point**	**Difference**
Milan	64	59.65	-4.35
Saint-Etienne	55	50.58	-4.42
CFR Cluj	59	54.54	-4.46
RB Leipzig	53	48.40	-4.60
Lazio	72	67.37	-4.63
Real Madrid	76	71.10	-4.90
Utrecht	54	48.80	-5.20

In the evaluation, the margin of error was considered only a match. So, predictions with less than three-point difference from the actual point value were considered successful. The success rates calculated for 15 European leagues are presented in Table 4.4. The overall success rate for all leagues was 40%.

Table 4.4 Success rates of leagues obtained in point prediction

League	Success Rate
Germany	56%
Czech Republic	56%
Romania	56%
Belgium	50%
England	44%
Denmark	33%
France	33%
Crotia	33%
Spain	33%
Italy	33%
Turkey	33%
Netherland	28%
Austria	22%
Scotland	22%
Portugal	11%

The points gained at the end of the season determine the team standings in the league. The predicted points by the PE of the teams in 2017-2018 end-of-season were used to measure the success of the proposed approach based on the standings, and the teams were ranked based on the points predicted among the teams in their leagues. Additional evaluations were performed for the first, first four and the last three teams in the leagues. The first four ranking are considered important for the league success of a team and qualifying to the European cups. The first ranking is significant in terms of it indicates the champion team of the league. Also, because of the relegation of failed teams, the last three ranking of a league are critical as well. When calculating the ranking-based success ratio (for the league-wide and first four), predictions that predict the season end ranking of a league exactly or predict the ranking by only one difference are accepted as successful. However, calculations for the first and last three ranking of the leagues were made without any margin of error since these rankings are critical. The ranking-based success ratios are given in Table 4.5.

Table 4.5 Success rates of leagues by ranking

League	**Success rate for ranking**	**Success rate for only first four ranking**	**Prediction result for first ranking**	**Succes rate for last three ranking**
Crotia	100%	100%	FALSE	33.33%
Scotland	100%	100%	TRUE	33.33%
Austria	90%	100%	TRUE	100%
Romania	86%	100%	TRUE	33.33%
Denmark	79%	100%	TRUE	100%
England	75%	100%	TRUE	0.00%
Netherland	72%	100%	FALSE	33.33%
Portugal	72%	100%	TRUE	66.66%
Italy	70%	100%	TRUE	0.00%
France	65%	100%	TRUE	33.33%
Spain	65%	100%	TRUE	33.33%
Czech Republic	63%	100%	FALSE	0.00%
Germany	61%	100%	TRUE	66.66%
Belgium	44%	50%	TRUE	66.66%
Turkey	39%	75%	FALSE	33.33%

The League-wide success ratios for the ranking were higher than 50%, except for two leagues (Belgium and Turkey). The rankings for Croatia and Scotland leagues were predicted exactly.

4.2 Real Data Application – 2

In this section, it was aimed to evaluate the success of the proposed approach in 3.6 with an application. In this context, firstly, the result predictions for all games were made with the Christodoulou method. Then, only for the games determined by the approach, the predictions were made using the same algorithm. The success of the proposed approach was examined by comparing the accuracy rates of the Christodoulou method obtained in both cases. The actual-expected season-end points of the teams who competed in the 2017-18 season and the game results of the same teams for the mentioned season were used as data set. As stated in Section 3.6, it is more appropriate to predict current season's games based on the last season actual-expected season-end points of the teams. However, to prevent data loss since the relegated teams (in the 2016-17 season) did not take part in the 2017-18 season, only the 2017-18 season was used in the application.

At first, predictions were made for 4213 games (all games), accuracy rate observed as 52.5%. Following, in line with the proposed approach, the results of the games that provide the appropriate conditions (based on the threshold values) were predicted considering the actual-expected season-end points of the teams. A small excerpt of the data was shown in Table 4.6. To introduce the application, a small part of the application was given in Table 4.7. In this small part, the threshold value was taken as 3 to determine the games to be predicted. Predictions were made for the matches which provide the appropriate condition ($\mathrm{d_{Home}} \geq 3$ and $\mathrm{d_{Away}} \leq -3$ or $\mathrm{d_{Away}} \geq 3$ and $\mathrm{d_{Home}} \leq -3$). R.S, P.S, and d notations were given in Table 4.7 indicate real score, predicted score, and actual-expected season-end points of the teams. R statistical programming language was used in the application.

Table 4.6 An example for 2017-18 match results from fifteen europan leagues

Home Team	Score	Score	Away Team
Karabükspor	0	7	Fenerbahçe
Malaga	0	1	Eibar
St. Liege	0	0	Gent
Ajax	2	0	Feyenoord
Chelsea	2	0	Everton
Udinese	2	6	Juventus
Betis	3	6	Valencia

Table 4.7 An example for the results of the proposed approach (k= 3)

Home Team (HT)	Diffrence (d) of HT	R.S for HT	P.S for HT	P.S for AT	R.S for AT	Difference (d) of AT	Away Team (AT)	Prediction of match result is
Feirense	-3.6	3	2	2	0	6.4	Boavista	False
Austria Vienna	-4.6	0	4	4	0	4.3	Admira	True
Alkmaar	9.9	3	6	2	2	-4.3	Willem	True
Chelsea	5.3	3	3	1	1	-3.3	Newcastle	True
Dinamo Zagreb	7.1	5	5	1	1	-4.5	Istra 1961	True
Everton	4.0	2	3	3	1	-3.4	Leicester	False

4.2.1 Results

After the analysis, the accuracy rate of the proposed approach was examined. When the proposed approach was not applied, the accuracy rate of the Christodoulou method observed as 52.5% for 4213 games. According to the proposed method, the number of predicted matches decreases when the accuracy rate increases. Table 4.8 and figure 4.2 show the accuracy rate versus the number of predicted matches.

Table 4.8 Accuracy rate and number of predicted matches according to the threshold values

Threshold value (*k*)	**Number of predicted matches**	**Accuracy rate**
1	1839	0.5753127
2	1339	0.6109037
3	932	0.6341202
4	524	0.6793893
5	307	0.7328990
6	169	0.7692308
7	58	0.7586207
8	43	0.8139535
9	17	0.9411765

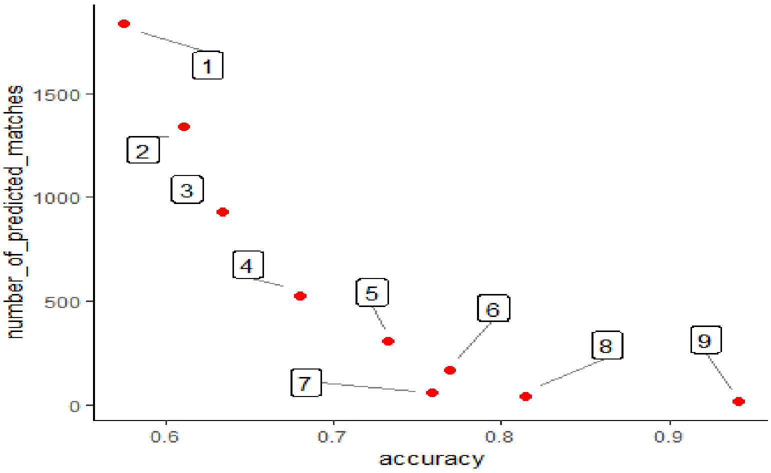

Figure 4.2 Accuracy rate and number of predicted matches according to the threshold values

CHAPTER FIVE
CONCLUSION

In this thesis, it is aimed to investigate the studies that are suggested for sports statistics and sports data mining. In this context, a literature review is performed by examining the formulas and sports data mining studies related to many sports discipline. Also, it is aimed to suggest new approaches that will contribute to sports statistics for football. For this purpose, two novel approaches are proposed using the background obtained in the literature review section. The accuracy of the proposed approaches are examined under two sub-sections in the application section. In the related literature, PE is adapted to many sports branches. However, a simple and useful adaptation for football could not be found. Therefore, firstly, a simple PE formula is proposed in the application section by using a data-driven approach for football, which one of the most popular sports in Turkey. Secondly, the Christodoulou method, a method similar to PE, is investigated. It takes into account the teams' goal scored and goal conceded statistics like PE. The approach uses these two methods together is suggested as a result of the examination of the Christodoulou method. It is assumed that the accuracy rate of the Christodoulou method would increase as a result of this usage. In order to determine the games to be predicted according to the proposed approach, the actual-expected season-end points of the matched teams are used as a criterion. The accuracy rate of the Christodoulou method is significantly increased for the games that meet the criterion.

In this study, a PE calculation approach is proposed for European football leagues. In section 4.1, the 2017-2018 end-of-season points of 244 teams playing in European leagues are predicted in order to measure the success of the proposed approach. The success of PE, which is proposed with two different approaches according to the ranking and point, is evaluated by using the points predicted. More successful results are obtained with ranking based prediction. In the study, relatively low success ratios are obtained for Turkey and Belgium leagues for the league-wide predictions. However, high success ratios are obtained for Croatia, Scotland, Austria and Romania, where there are fewer teams in the league compared to the other countries (Baysal &

Yıldıztepe, 2019b). Also, while ranking-based accuracy is performed, the first and last three rankings are predicted for each league without any margin of error. When the results are examined, it can be sad that the PE formula is quite successful in predicting the league champions. Nevertheless, the success rate in predicting the last three rankings of the leagues is mediocre.

In this study, in addition to proposing a PE formula for football, an approach is proposed for increasing the success of the Christodoulou method in predicting the game result. To examine the success of the proposed approach, matches results of 244 teams competing in European leagues are predicted using the Christodoulou method in section 4.2. Firstly, the prediction is made for 4213 games unconditionally, the accuracy rate of the Christodoulou method observed as 52.5%. Then, the results of games that meet the conditions set by the proposed approach (for k =1,2,3,...,9) are predicted by different accuracies. In line with the proposed approach for $k = 1$, 1839 out of 4213 games meet the criterion. For these games, the prediction accuracy rate observed at 57.5%. When the threshold value k is selected as 9, the accuracy rate of the Christodoulou method came near to excellent (94%), but there is a dramatic decrease in the number of predicted games. There is a trend that when the threshold value increase, the number of predicted games decrease. The mentioned trend has a flaw only for k = 6. The prediction accuracy rate for k = 6 is higher (76.9%) than k = 7 (75.8%). Therefore, the best threshold value k to be used in the proposed approach is determined as 6. In this case, the algorithm that determines the matches to be predicted with the Christodoulou method can be written as $d_{Home} \geq k = 6 \; and \; d_{Away} \leq k = -6 \; or \; d_{Away} \geq k = 6 \; and \; d_{Home} \leq k = -6$. It is possible to have a financial gain source through betting on the predicted games. The profit factor can be used as another criterion for determining the threshold value. In the second application, the optimum threshold value is determined by examining the number of predicted games and the accuracy rates. If the profit factor, accuracy rate, and the number of predicted games are evaluated together, a better threshold value can be determined for the proposed approach. In future studies, it is planned to examine these three factors together to develop a better approach.

In the study, the γ value in PE formula is calculated as 1.2 for European leagues. But γ value may change based on the specific leagues. Hence, specific γ values must be calculated for every league to make accurate predictions. Further studies are planned to determine the γ values for Asian and South American football leagues.

REFERENCES

Audas, R., Dobson, S., & Goddard, J. (2002). The impact of managerial change on team performance in professional sports. *Journal of Economics and Business*, *54*, 633-650.

Akhtar, S., & Scarf, P. (2012). Forecasting test cricket match outcomes in play. *International Journal of Forecasting*, *28*, 632-643.

Allsopp, P. E., & Clarke, S. R. (2004). Rating teams and analysing outcomes in one-day and test cricket. *Journal of the Royal Statistical Society. Series A*, *167*, 657–667.

Barnett, V., & Hilditch, S. (1993). The effect of an artificial pitch surface on home team performance in football (soccer). *Journal of the Royal Statistical Society. Series A*, *156*(1), 39-50.

Baseball-Reference. (2015). *On Base Percentage*. Retrieved May 15, 2018, from https://www.baseball-reference.com/bullpen/On_base_percentage

Baseball-Reference (2017). *Slugging Percentage*. Retrieved March 06, 2018, from https://www.baseball-reference.com/bullpen/Slugging_percentage

Baysal, S., & Yıldıztepe, E. (2019a). Futbol takımlarının sezon sonu puanlarının tahmini. *21. Akademik Bilişim Konferansı*. Ordu.

Baysal, S , Yıldıztepe, E . (2019b). Prediction of season-end point for football using pythagorean expectation. *Journal of New Theory, (27)*, 63-73.

Bosch, P., & Bhulai, S. (2018). *Predicting the winner of NFL-games using machine and deep learning*. Retrieved November 5, 2018, from https://www.semanticscholar.org/paper/Predicting-the-winner-of-NFL-games-using-Machine-Bosch-Bhulai/bcd94e514ac1ed34622810faea2914669071f641

Bramer, M. (2007). *Principles of data mining*. London: Springer.

Bunker, R. P., & Thabtah, F. (2017). A machine learning framework for sport result prediction. *Applied Computing and Informatics*, *15*(1), 27-33.

Buursma, D. (2011). *Predicting sports events from past results towards effective betting on football matches*. Retrieved October 15, 2018, from https://www.semanticscholar.org/paper/Predicting-sports-events-from-past-results-Towards-Buursma/5e22c4362df3b0accbe04517c41848a2b229efd1

Cao, C. (2012). *Sports data mining technology used in basketball outcome prediction*. Retrieved August 13, 2018, from https://arrow.dit.ie/scschcomdis/39/

Caro, C., & Machtmes, R. J. (2013). Testing the utility of the Pythagorean Expectation formula on division one college football: An examination and comparison to the Morey model. *Journal of Business & Economics Research*, *11*, 537-542.

Clarke, S. R., & Norman, J. M. (1995). Home ground advantage of individual clubs in English Soccer. *Journal of the Royal Statistical Society . Series D, 44*, 509-521.

Cochran, J. J., & Blackstock, R. (2009). Pythagoras and the National Hockey League. *Journal of Quantitative Analysis in Sports, 5*(2), 1-13.

Croucher, J. S. (2000). Player ratings in one-day cricket. *Proceedings of the Fifth Australian Conference on Mathematics and Computers in Sport,* 95-106. University of Technology, Sydney .

Crowder, M., Dixon, M., Ledford, A., & Robinson, M. (2002). Dynamic modelling and prediction of English football league matches for betting. *Journal of the Royal Statistical Society. Series D, 51*, 157-168.

Davenport , C., & Woolner, K. (1999, June 30). *Revisiting the Pythagorean Theorem: Putting Bill James' Pythagorean Theorem to the test*. Retrieved December 12, 2018, from https://www.baseballprospectus.com

Delen, D., Cogdell, D., & Kasap, N. (2012). A comparative analysis of data mining methods in predicting NCAA bowl outcomes. *International Journal of Forecasting, 28*, 543-552.

Dixon, M. J., & Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society. Series C, 46*, 265-280.

Dixon, M. J., & Robinson, M. E. (1998). A birth process model for association football matches. *Journal of the Royal Statistical Society. Series D, 47*, 523-538.

Eastwood, M. (2012). *Applying the Pythagorean Expectation to football: Part two*. Retrieved September 5, 2018, from https://pena.lt/y/2012/12/03/applying-the-pythagorean-expectation-to-football-part-two/

Esme, E., & Kiran, M. S. (2018). Prediction of football match outcomes based on bookmaker odds by using k-Nearest Neighbor algorithm. *International Journal of Machine Learning and Computing, 8*(1), 26-32.

Fitzpatrick, J. (2019, April 11). *The definition and purpose of the Plus/Minus statistic in Hockey*. Retrieved May 19, 2019, from https://www.thoughtco.com/what-is-the-plus-minus-statistic-2779372

Forrest, D., & Simmons, R. (2000). Forecasting sport: The behaviour and performance of football. *International Journal of Forecasting*, *16*, 317-331.

Goddard, J., & Asimakopoulos, I. (2004). Forecasting football results and the efficiency of fixed-odds betting. *Journal of Forecasting*, *23*(1), 51-66.

Gu, W., Saaty, T. L., & Whitaker, R. (2016). Expert system for Ice Hockey game prediction: Data mining with human judgment. *International Journal of Information Technology & Decision Making*, *15*, 763-789.

Hamilton, H. H. (2011). *An extension of the Pythagorean Expectation for association football*. Retrieved September 10, 2018, from https://www.degruyter.com/view/j/jqas.2011.7.2/jqas.2011.7.2.1335/jqas.2011.7.2.1335.xml

Hill, I. D. (1974). Association football and statistical inference. J*ournal of the Royal Statistical Society. Series C*, *23*, 203-208.

Hollinger, J. (2002). *Pro basketball prospectus*. Dulles: Brassey's Inc.

James, B. (1980). *The Bill James baseball abstract*. New York: Ballantine Books

James, B. (1982). *The Bill James baseball abstract*. New York: Ballantine Books.

James, B. (1985). *The Bill James historical baseball abstract*. New York: Villard.

James, B., & Henzler, J. (2002). *Win Shares*. Morton Grove: STATS Publishing.

Kubatko, J., Oliver, D., Pelton, K., & Rosenbaum, D. T. (2007). A starting point for analyzing basketball. *Journal of Quantitative Analysis in Sports*, *3*(3), 1-24.

Kuypers, T. (2000). Information and efficiency: An empirical study of a fixed odds betting market. *Applied Economics*, *32*, 1353-1363.

Lee, J. (2015). Measuring the accuracy of the Pythagorean theorem in Korean pro-baseball. *Journal of the Korean Data and Information Science Society*, *26*, 653-659.

Lemmer, H. H., Bhattacharjee, D., & Saikia, H. (2014). A consistency adjusted measure for the success of prediction methods in cricket. *International Journal of Sports Science & Coaching*, *9*, 497–512.

Leung, C. K., & Joseph, K. W. (2014). Sports data mining: Predicting results for the college football games. *Procedia Computer Science*, *35*, 710-719.

Lewis, M. (2003). *Moneyball*. New York: W.W. Norton & Company.

Li, Y., & Zhang, Y. (2012). Application of data mining techniques in sports training. *5th International Conference on BioMedical Engineering and Informatics*, 954-958. Chongqing, China: IEEE.

Linsey, G. R. (1963). An investigation of strategies in baseball. *Operations Research*, *11*, 477-501.

Mackolik. (2018). *League tables.* Retrieved November 5, 2018, from http://arsiv.mackolik.com

Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, *36*, 109-118.

Malhotra, A., & Krishna, S. (2017). *A statistical analysis of bowling performance in cricket*. Retrieved May 12, 2018, from https://arxiv.org/abs/1701.04438

Maszczyk, A., Gołaś, A., Pietraszewski, P., Roczniok, R., Zając, A., & Stanula, A. (2014). Application of neural and regression models in sports results prediction. *Procedia - Social and Behavioral Sciences, 117*, 482 – 487.

Maszczyk, A., Rocznıok, R., Waśkıewıcz, Z., Czuba, M., Mıkołajec, K., Zając, A., et al. (2012). Application of regression and neural models to predict competitive swimming performance. *Perceptual and Motor Skills, 114*, 610-626.

Memmert, D., Lemmink, K. A., & Sampaio, J. (2016). Current approaches to tactical performance analyses in soccer using position data. *Springer International Publishing*, 47(1), 1-10.

Miljković, D., Gajić, L., Kovačević, A., & Konjović, Z. (2010). The use of data mining for basketball matches outcomes prediction. *International Symposium on Intelligent Systems and Informatics,* 309-312. Subotica, Serbia: IEEE.

Miller, S. J. (2007). A Derivation of the Pythagorean Won-Loss Formula in Baseball. *Chance*, *20*(1), 40-48.

Oliver, D. (2004). *Basketball on paper : rules and tools for performance analysis* . Washington, D.C: Brassey's Book.

Rosenbaum, D. T. (2005). *Defense is all about keeping the other team from scoring*. Retrieved October 11, 2018, from http://www.82games.com/rosenbaum3.htm

Rue, H., & Salvesen, O. (2001). Prediction and retrospective analysis of soccer matches in a league. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *49*(3), 399-418.

Scarf, P., & Akhtar, S. (2011). An analysis of strategy in the first three innings in. *Journal of the Operational Research Society*, *62*, 1931-1940.

Schatz, A. (2003). *Pythagoras on the gridiron*. Retrieved August 08, 2018, from https://www.footballoutsiders.com

Schatz, A. (2006). *Pro football prospectus 2006: Statistics, analysis, and insight for the information age*. New York: Workman Publishing Company.

Schumaker, R. P., Solieman, O. K., & Chen, H. (2010). *Sports data mining*. Boston: Springer.

Society for American Baseball Research (n.d.). *A guide to sabermetric research*. Retrieved September 24, 2019, from https://sabr.org/sabermetrics

Tax, N., & Joustra, Y. (2015). Predicting the Dutch football competition using public data: A machine learning approach. *Transactions on Knowledge and Data Engineering*, *10*(10), 1-13.

Thorn, J. (2013). *Chadwick's choice: The origin of the Batting Average*. Retrieved March 12, 2018, from https://ourgame.mlblogs.com/chadwicks-choice-the-origin-of-the-batting-average-e8e9e9402d53

Thorn, J., & Palmer, P. (1984). *The hidden game of baseball*. New York: Doubleday & Company.

Thorn, J., & Palmer, P. (1993). *Total baseball*. New York: Harper Collins.

Tung, D. J. (2010). *Confidence intervals for the Pythagorean formula in baseball*. Retrieved September 2018, from https://www.semanticscholar.org/paper/Confidence-Intervals-for-the-Pythagorean-Formula-in-Tung/b99773a55009258dbd5cbb3dd17cc3a989a68679

Valero, S. C. (2016). Predicting win-loss outcomes in MLB regular season games – a comparative study using data mining methods. *International Journal of Computer Science in Sport*, *15*(2), 91-112.

Vine, A. J. (2016). Using Pythagorean Expectation to determine luck in the KFC Big Bash League. *Economic Papers*, *35*, 269-281.

Zminda, D., & Dewan, J. (1993). *Stats basketball scoreboard, 1993-94*. Skokie: Harpercollins Publishers.