T.R.
SELCUK UNIVERSITY
GRADUATE SCHOOL OF NATURAL SCIENCES

DECISION SUPPORT SYSTEM FOR A FOOTBALL TEAM MANAGEMENT BY USING MACHINE LEARNING TECHNIQUES

Mustafa Aadel Mashjal AL-ASADI

MASTER'S THESIS

Computer Engineering Department

August -2018
KONYA

THESIS ACCEPTANCE AND APPROVAL

Mustafa Aadel Mashjal AL-ASADI prepared a thesis titled "Decision Support System for a football team management by using machine learning techniques" on 01/08/2018 by the following jury with unanimous vote consent in Selcuk University Institute of Science and Technology Computer Engineering Department has been accepted as a Master's Thesis.

Jurors	**Sign**
Head of Jurors Assoc. Prof. Dr. Halife KODAZ	.....
Advisor Prof. Dr. Şakir TAŞDEMİR	.....
Juror Assist. Prof. Dr. Abdullah Erdal TÜMER	.....

My final conclusions, above.

Prof. Dr. Mustafa YILMAZ
Graduate School of Natural Sciences Director

TEZ BİLDİRİMİ

Bu tezdeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edildiğini ve tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

DECLARATION PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Sign

Mustafa Aadel Mashjal AL-ASADI

Date: 01/08/2018

ÖZET

YÜKSEK LİSANS TEZİ

MAKİNE ÖĞRENMESİ TEKNİKLERİ İLE BİR FUTBOL TAKIMI YÖNETİMİ İÇİN KARAR DESTEK SİSTEMİ

Mustafa Aadel Mashjal AL-ASADI

Selçuk Üniversitesi Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Prof. Dr. Şakir TAŞDEMİR

2018, 110 Sayfa

Jüri
Prof. Dr. Şakir TAŞDEMİR
Doç. Dr. Halife KODAZ
Dr. Öğr. Üyesi Abdullah Erdal TÜMER

Futbol oyuncu ve izleyici sayısı bakımından dünyadaki en popular spordur. Futbolun popülerliği son yıllarda artmıştır ve küresel ekonominin önemli bir parçası olmuştur. 2017 yılı içerisinde sadece Avrupa kulüplerinin geliri 27 milyar dolardır. Bu spordaki başlıca zorluklardan birisi, belli bir takım formasyonu için her pozisyona uygun oyuncunun yerleştirilmesidir. Bu zorluğun sebebi takımdaki her oyuncunun uygun olduğu pozisyonu verecek bilimsel bir formül veya denklemin olmayışıdır. Futbolcuların uygun pozisyonlarının belirlenmesi takım koçları tarafından gözlemlere ve tecrübeye dayalı olarak yapılmaktadır ve bu durum kişisel yargılara sebep olmaktadır. Bu zorlukların üstesinden gelebilmek için bir karar destek sistemi oluşturulmuştur.

Bu tez çalışmasında futbol takımı yönetimi için makine öğrenmesi yöntemlerinden faydalanan yeni bir zeki karar destek sistemi önerilmiştir. Bu karar destek sisteminin başlıca hedefi takımdaki her oyuncu için kişisel yeteneklerini temel olarak en uygun pozisyonu belirlemek ve istenen formasyona göre en iyi takımı oluşturmaktır. Son olarak, sistem her oyuncunun top sürme yeteneğini belirleme yeteneğine sahiptir. Oyuncunun top sürme becerisinin gözlenmesi yöneticilerin oyucu alım, satım ve sözleşme yenileme işlemlerinde daha uygun kararlar vermesine yardımcı olur.

Bu tez çalışmasında bir sezon için 17359 oyuncu içeren FIFA futbol oyunu verileri kullanılmaktadır. Oyuncu verilerini analiz ederken, sınıflandırma ve regresyon problemleri için makine öğrenmesi teknikleri kullanılmıştır (linear and logistic regression, random forest, neural network and k nearest neighbor). Ayrıca, veri boyutunu düşürmek için principal component analysis ve recursive feature elimination algoritmalarından yararlanılmıştır. Bu algoritmalar ile 29 nitelik içerisinden 17 tanesini kullanarak her oyuncunun uygun pozisyonunun belirlenebileceği görülmüştür.

Önceki çalışmalardan farklı olarak bu tezde, her oyuncunun uygun pozisyonunu bulmak için rastgele orman algoritması kullanılmıştır. Bu algoritma ikili sınıflandırma için % 88.60 ve çoklu sınıflandırma için % 58.53 doğruluk değerleri ile diğer algoritmalardan daha verimli değerler vermiştir. Bu algoritmaların performanslarının değerlendirilmesi için üç teknik kullanılmıştır (Hold-out, Cross Validation and Repeated Random Hold-out). Her oyuncunun uygun pozisyonunu belirledikten sonra istenilen formasyona göre en iyi takım her oyuncunun derecesi dikkate alarak oluşturulmaktadır.

Son olarak top sürme becerisini belirlemek için dört farklı algoritma kullanılmıştır (linear regression, logistic regression, random forest and neural network). En iyi sonucu 17 performans niteliği kullanılarak % 99.90 doğruluk değeriyle rastgele orman algoritması vermiştir.

Keywords— Karar Destek Sistemleri (KDS), Makine Öğrenmesi, Futbol, Takım Yönetimi, Oyuncu Seçimi, Takım Seçimi, Bireysel Yetenekler, Top Sürme, FIFA Futbol Video Oyunu

ABSTRACT

MS THESIS

DECISION SUPPORT SYSTEM FOR A FOOTBALL TEAM MANAGEMENT BY USING MACHINE LEARNING TECHNIQUES

Mustafa Aadel Mashjal AL-ASADI

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCE OF SELÇUK UNIVERSITY
THE DEGREE OF MASTER IN COMPUTER ENGINEERING

Advisor: Prof. Dr. Şakir TAŞDEMİR

2018, 110 Pages

Jury
Prof. Dr. Şakir TAŞDEMİR
Assoc. Prof. Dr. Halife KODAZ
Assist. Prof. Dr. Abdullah Erdal TÜMER

Football considered as most popular sport in the world in both number of spectators and players. Popularity of football has increased in the last years and it became an important contributor to the global economy. Where, the revenue for European football clubs alone for 2017 rated at $27bn. In football, team management consider as one main challenges in this sport especially those related to choosing the suitable player for the suitable position in a specific formation because there is no formula or scientific equations used to identify the preferred available position for player in team. where, the assignment generally is done by the coaches by use their experiences and observations about the players, these making selecting of players subject to many biases. Therefore, we need to build intelligent decision support systems to face these challenges. This thesis proposes a new intelligent decision support system for a football team management by using algorithms of machine learning. The main purpose of this decision support system is to find intelligent solutions based on skills of players (technical, physical and mental) to find preferred available position for player in team and find the best available squad according to formations of play. finally, the system has ability to predict dribbling skill for each player in the team to monitor the growth and performance of players because predicting player's skill (like dribbling) will help managers to make suitable decisions like sell, buy and contract renewal.

In this thesis we use dataset of FIFA Soccer video game, that contains data for 17359 players for one season. When analyzing players data, we have used machine learning techniques (linear and logistic regression, random forest, neural network and k nearest neighbor) for classification and regression problems. Further, we use recursive feature elimination algorithm (RFE) and principal component analysis (PCA) algorithm for reducing data dimension. where, we found 17 performance attributes using which we can predict the preferred available position for each player in team out of 29 attributes.

Differently from the previous studies, in this thesis we use random forest algorithm to find preferred available position for each player in team, and it has proved to be more efficient in classification of players position's than other algorithms, where we obtained a predictive accuracy of 88.6% for binary classification (2 position) and predictive accuracy of 58.5% for multi classification (14 position). Where,

the performance of all these algorithms are evaluated using three common techniques (Hold-out (train and test split), Cross Validation (CV) and Repeated Random Hold-out) and comparison the result among them. After assigning each player to the position we determine the best team squad according to formation plays (like 4-3-3 or 3-5-2) based on rating of player.

Finally, to predict the skill of dribbling, we used four algorithms (linear regression, logistic regression, random forest and neural network). We got best result by using random forest, where predictive accuracy was 99.9% by using 17 performance attributes.

Keywords— Decision Support Systems (DSS), Machine learning, Football, Team management, Player selection, Team selection, Individual Skills, Dribbling, FIFA video game series system.

ACKNOWLEDGMENTS

First of all, I would like to thank ALLAH almighty to enabling me to complete this thesis, his continuous mercy was with me through my life and ever more during the tenure of my study.

Now, I would like to express my deep and sincere gratitude to my supervisor Prof. Dr. Sakir Tasdemir for his continuous support, great guidance, endless help and huge confidence that gave for me to grow as a research scientist. Further, I would thank all professors in computer engineering department, whom I have worked with over the last two years.

I am also very grateful to Research Assistant Burak Tezcan for the advice and support who has shown a large interest in my work.

I would like to thank Assist. Prof. Dr. Ilker Ali Ozkan and Lecturer Ali Yasar for their ideas and advices which have been absolutely invaluable.

Also, I wish to thank Mr. Suheyb Tumer to help us connect with specialists in the Sports Science Faculty at the Selcuk University.

Further, I would like to express my deep gratitude to my friend Ahmed Meften to help me in gathering data set from the web and all other great services.

Dearest friend Mouaz Al-habal, I thank his for his constant support and encouragement throughout my graduate career and for his valuable information about football that helped me a lot to achieve this study.

Nearest friend Mustafa Hussein, thank you for being in my life.

Finally, I am extremely grateful to my parents, brothers and amazing sister for their love, prayers and sacrifices for educating and preparing me for my future.

Mustafa Aadel Mashjal AL-ASADI
KONYA-2018

TABLE OF CONTENTS

SYMBOLS AND ABBREVIATIONS

ANNs	Artificial Neural Networks
Bagging	Bootstrapped Aggregation
Blending	Stacked Generalization
CNN	Convolutional Neural Network
CV	Cross Validation
DCNNs	Deep Convolutional Neural Networks
DSS	Decision Support System
EM	Expectation Maximization
EPL	English Premier League
FC	Football Club
FS	Feature selection
FDA	Flexible Discriminant Analysis
FIFA	Fédération International de Football Association
FIFA Soccer	Video Game
FIS	Fuzzy Inference System
FM	Football Manager
GBM	Gradient Boosting Machines
GBRT	Gradient Boosted Regression Trees
GPS	Global Positioning System
KBS	Knowledge based systems
ID3	Iterative Dichotomiser 3
IDSS	Intelligent Decision Support System
K-NN	k-Nearest Neighbor
LDA	Linear Discriminant Analysis
LR	Logistic regression
LED	Light Emitting Diode
LOESS	Locally Estimated Scatterplot Smoothing
LVQ	Learning Vector Quantization
LWL	Locally Weighted Learning
MARS	Multivariate Adaptive Regression Splines
MCFC	Manchester City Football Club
MDA	Mixture Discriminant Analysis

MDS	Multidimensional Scaling
ML	Machine Learning
MLP	Multi-layer perceptron
OLSR	Ordinary Least Squares Regression
OOM	Object Oriented Methodology
OOB	Out-Of-Bag
PCA	Principal component analysis
RFE	Recursive Feature Elimination
RSS	Residual sum of squares
ReLU	Rectified Linear Unit
RF	Random Forest
TSS	Total Sum of Squares
W	Weight coefficients
Δw	Change of the weight
TP	Number of true positives.
FP	Number of false positives.
TN	Number of true negative.
FN	Number of false negatives.
X	Input data
y	Output data
R^2	Coefficient of determination
r_{xy}	Correlation coefficient
β_0	Intercept from the linear regression equation
β_1	Regression coefficient
f	Activation function
t	Target value
η	Learning rate
θ	Threshold

TABLES

FIGURES

1. INTRODUCTION

Football considered as most popular sport in the world in both number of spectators and players (Cotta ve ark., 2016). Popularity of football has increased in the last years and have become a new emerging industry. The sports industry has seen a lucrative rise in stature and has now become an important contributor to the global economy (Asif ve ark., 2016). where the revenue for European football clubs for 2017 rated at $ 27 billion (Davis, 2017). Therefore, the European clubs have become trade organizations (Moor, 2007).

1.1. General

Technological advances led to increase the generation of football data (for players and matches), In this regard, a number of commercial companies have emerged to provide data analysis and collection. For example, in the UK, three data companies have been provided since the late 1990s: Opta, Amisco and Prozone. All these data systems provide detailed data on players and matches, as well as the clubs' reliance on these systems to provide information after the matches (Schulenkorf, 2016). The FIFA video game considered as another source of football data, where it offering detailed data about players (Markovits ve Green, 2017). This volume of data, combined with the development of sports technology led to facilitated the build new decision support systems in sports management.

In football, each player in game is allocate to one of 11 specific locations in the playing field. All these positions in the law of game represent the main role of the player and his operating area on the pitch. The problem of identifying football players by their position on the pitch is complicated because of the fluid nature for the game. The means of fluid nature of game refer to the positions in football game are not constant for each player in team like American football or rugby (Kennelly, 2010).

Sport teams, usually require the coach to set the team formation and choose the best available players for all positions in this formation, where the success of any football team lies in the players skills who make up the team. The selecting of players and forming a team is a complex problem, where final success for coaches is determined by gathering players to form a strong and effective team (Tavana ve ark., 2013). Generally,

coaches do not disclose the criteria they use to classify players, therefore the selection of players can be prone to biases by coaches, where the process is done by the coaches by use their experiences and observations about the players.

Based on above, obviously people could not check whether a certain selection of players is fair. Coaches usually are using game statistics in the evaluation, such as: total shots of player on target, total number of passed, goals scored, shot efficiency, steals, turnovers and others. that is mean, there is a lack of criteria that enable coaches to evaluate players accurately (Hraste ve ark., 2008). therefore, one of the crucial criteria's that must be considered when assigning the position for each member is taking into account his individual attributes (physical, mental and technical) (Abidin ve ark., 2016a). It is necessary to define accurately those criteria and to determine the degree of importance of each and every criterion in relation to playing positions because the lack of suitable criteria that can be used to judge a player, in addition to the current analyses and researches on systematic analysis of the team and Organizing players through information technology and artificial intelligent have not been widely used and very insufficient (Papić ve ark., 2009).

In the same context, talent identification considered as one of important tasks in football. Therefore, all clubs (especially the European) seek to determine the football talent of players at an early age (Roderick, 2006). Predicting of player's skill is one of the most important ways of talent identification. as well as predicting may help managers to make suitable decisions like sell, buy and contract renewal. Therefor coaches are in great need for a modern technological solution for team management (Silva ve ark., 2009).

1.2. Define of Problem and Aim of Study

Machine learning algorithms are widely used in many different disciplines. One of them is a sports field. Research in football analytics with Machine Learning techniques is limited and is mostly employed on result prediction, as well as there are very few researches about algorithms or decision support systems have been made that could be used by decision makers to aid in the process of forming a strong team.

One of main challenge in team management is concern on how to choose preferred available position for player in team. there is no formula or scientific equations used to

identify the preferred available position for player in team. where, the assignment generally is done by the coaches by use their experiences and observations about the players. In the same context, one of the other challenges for football managers is knowing how the skill of players changed over time because predicting player's skill will help managers to make suitable decisions like sell, buy and contract renewal.

For all these reasons, there is a need for research and studies to determine how machine learning applications can yield results in soccer analytics. To achieve this purpose, we will seek to build a decision support system (DSS) based on machine learning techniques has the ability to:

i. Predict preferred available position for each player in team such as striker, wing backs, right and left back, etc.

ii. Find the best available squad according to formations of play such as 4-4-2, 3-5-2, etc.

iii. Predict player skills (especially dribbling skill) for each player (where previous studies indicated the most important technical skill and most discriminating variable among players skills is dribbling (Soto-Valero, 2017)).

1.3. Overview of This Thesis

This thesis is organized as follows. Introduction and our contributions are outlined in Chapter 1. Chapter 2 provides a literature review about using machine learning in football, and decision support system models that built for this purpose.

Chapter 3 introduces the general concepts and information about football game. In addition to explain players' individual qualities (physical, mental, and technical).

In Chapter 4, introduces the general information about traditional decision support systems (DSS) and intelligent decision support systems (IDSS), addition to explain main algorithms in machine learning according to their learning style, similarity and main methods in evaluate the performance of algorithms. Chapter 5 provides a material and methods.

In Chapter 6, introduces the general information about dataset and analyzes

In Chapter 7, introduces decision support system model to predict dribbling skill based on filter and wrapper strategy.

In Chapter 8, introduces decision support system model to predict preferred available position for each player in team. Further, the first section in this chapter explained player positions in football team and required skills for determining player position in team.

In Chapter 9, introduces decision support system model to find the best available squad according to formations of play such as 4-4-2, 3-5-2, etc.

Chapter 10 provides a recommendation and conclusions. Further, references.

2. LITERATUARE REVIEW

Machine learning have been successfully applied in sport. and there are many researches focused on developing DSS to be used to help in sport management (Abidin ve ark., 2016b). As a result of literature review, machine learning has been used to assists coaches and managers in five topics in football which are:

- Result prediction
- Player injury prediction
- Evaluation players & Select best players for formation
- Predicting of player skill's, wages and value
- Football Analytics

2.1. Result Prediction

The selection of important variables in football and the prediction of match results has made many efforts. Prediction is very important in football to help club managers and coaches to make the right decision to win in tournaments and matches. As well as businesses and gamblers have been trying to prediction game results in football for both tournaments and single matches. Organized football gambling on the other hand has developed into a growing industry and is now worth billions. As a result, there is literature on match prediction models. Studies below are the most important studies regarding to match result prediction using machine learning techniques:

Hijmans et al., (2017) proposed a learning algorithm through multiple data mining are analyzed and prediction outcomes are compared to come to a right model for predicting matches of the Dutch football team. Based on the prediction results of Naïve Bayes model, a random tree model, and a k-nearest neighbor model one single model is selection and results are looked at more in-depth. From the random tree model, which was the most predictive power (Hijmans ve Bhulai, 2017).

Razali et al., (2017) proposed a learning algorithm by using Bayesian Networks to predict the outcome of matches in win of term of home (H) or away (A) win and draw (D). The English League are selected for three seasons of 2011, 2012, 2013 and reviewed. K-fold has been used for testing the accuracy of prediction. Bayesian Networks done predictive accuracy at 75.09% (Razali ve ark., 2017).

Kınalıoğlu et al., (2017) predicted result of 15 elimination rounds including 8 2nd round, 4 quarter, semi-final and final will be played in 2017 UEFA Champions using ANN, SVM and K-nn algorithm methods. The statistical data of 7 seasons played between 2010-2016 which is obtained from "whoscored.com". This host regularly publishes soccer statistics are compiled and used as training data. In the last part of the study, the successes of the prediction methods were compared (Kınalıoğlu ve ark., 2017).

Velcich (2017) using machine learning techniques researcher attempted to predict the results of European league fixtures. Using match statistics from the Football-Data.co.uk, researcher calculated parameters to use in several different machine learning algorithms: polynomial regression, quadratic discriminant analysis (QDA), SVM and RF classifier (Velcich, 2017).

Prasetio (2016) proposed a logistic regression model is construct to predict outcome of English League for 2015/2016 season for home or away win. they are also used data from video game FIFA. the prediction accuracy of built model was 69.5% (Prasetio, 2016).

Bo Shen (2016) proposed a learning algorithm by using a neural network to predict the results of soccer games Depending on the many factors like players skills, coach abilities, home ground away ground effect, team tactics, etc. Their data source was a computer game called Football manager (FM). in the result they proved the neural network is the appropriate model for the problem and their able to predict soccer football results with test error about 25% (Shen, 2016).

Wang et al., (2015) proposed a learning algorithm by using ANN to predict the results for soccer matches depending on the many factors like players skills, coach abilities, home ground away ground effect, team tactics, etc. Their data source was a computer game called Football manager (FM) (Wang ve ark., 2015).

Tax et al., (2015) proposed a public data-based match prediction system for the Dutch Eredivisie. Model training was done on a self-made dataset from public sources, consist of thirteen seasons of Dutch Eredivisie match data. Several combinations of dimensionality reduction techniques and classification algorithms have been tested on the public data training set in a structured way. The highest prediction accuracy on the public data feature set was achieved by using a combination of PCA (with 15% variance) with a Multilayer Perceptron classifier or a Naive Bayes (Tax ve Joustra, 2015).

Igiri (2015) seek to investigate the ability of the Support Vector Machine to predict match results, in this model was used Gaussian combination kernel to generate 79 support vectors at 100000 iterations. 16 example football match results were trained to predict 15 matches. The result showed 53.3% prediction accuracy, which is comparatively low. an SVM-based system (as devised here) is not good enough in this application domain (Igiri, 2015).

Gomes et al., (2015) proposed decision support system (DSS) to support bookmaker's users to increase their profits on bets related to football matches. The aim of the project is to support betting users to increase their profits in bets that related with football (away win, home win or draw) (Gomes ve ark., 2015).

Shin et al., (2015) proposed using data from virtual games like FIFA to predict the match result. several features of the players were combined and compared that with the real-time prediction by applied Logistic Regression and Linear support vector machines. Accuracy predictor at 75% and virtual predictor at 80% (Shin ve Gasparyan, 2014).

Arabzad et al., (2014) proposed a machine learning algorithms and neural networks to predict the result of one week in the Iranian football league for the 2013-2014 season Based on previous games in the last seven leagues, the results have proved the ability of neural networks to predict match results (Arabzad ve ark., 2014).

Yezus (2014) proposed using data set from two sources to predict the football match result. in order to achieve the highest accuracy. Classifiers used are nearest neighbor and Random forest. The accuracy of these two models was at 55.8% and 63.4% (Yezus, 2014).

Moroney (2014) proposed analyses football scores from football-data.co.uk and check if match facts, such as goals, fouls, shots on target etc. can predict match outcomes. The aim is to further develop the skills obtained during the course, such as databases, programming, statistics, Business Analysis and Data Mining. The analysis was conducted through the construction of an SQL database, statistical analysis in R and machine learning in WEKA. The study proved that there are relationships between fouls, shots on goal etc. and that the outcome of the game can be predicted by match facts (Moroney, 2014).

Igiri et al., (2014) proposed analyses a complex set of data, they predict the match winner with assist tool called rapid miner in addition to using another process called Knowledge Discovery in Database. Classifiers used are logistic regression and artificial

neural network. The accuracy of at 93% is obtained in predicting the match winner (Igiri ve Nwachukwu, 2014).

Ulmer et al., (2013) proposed a machine learning algorithms (Naïve Bayes, Linear from stochastic gradient descent, Random forest and Support Vector Machine, hidden Markov model) to predict the football match results in English Premier League. The accuracy of each model was calculated to find the best approach. After comparing all the previous methods, they found that SVM had the best approach, where the accuracy at 55%-69% was showed in the prediction (Ulmer ve ark., 2013)

Owramipur et al., (2013) proposed using BN to predict the results of football matches for Barcelona Football club. The period under study was the 2008-2009 season in Spanish football league. they found BN can uses this to predict football results in future matches and they saw the final result in predictions was correct in 92%(Owramipur ve ark., 2013).

Constantinou et al., (2012) proposed using a Bayesian network model for prediction Football result according to knowledge and data, to predict English Premier League (EPL) matches before they start, and demonstrated profitability against all of market odds, and compared with another published football prediction models, pi-football it proved exceptionally accurate in prediction (Constantinou ve ark., 2012).

Hucaljuk et al., (2011) proposed using machine learning model are developed to solve the problem of Predicting football result. During the development of the model, several of tests have been made in order to determine the optimal attributes and classifications. The results of this model show a good ability of prediction (Hucaljuk ve Rakipović, 2011).

Huang et al., (2010) proposed a prediction model based on using multi-layer perceptron with back propagation learning rule. Based on the MLP prediction way, the prediction accuracy can achieve 76.9% if the draw games are excluded. prediction system is based on the Multilayer perceptron (MLB) with back propagation neural network learning, the prediction accuracy of the model was 76.9% (Huang ve Chang, 2010).

Buursma (2010) proposed a system for predicting the results of football matches that beats the bookmakers' odds is presented. The predictions for the matches are based on previous results of the teams involve (Buursma, 2010).

Van Gemert et al., (2010) proposed a statistical model to fulltime scores of Premier League football matches. the statistical model accounts for dependence between the

number of goals scored by the home and away team. For the marginal distributions of the number of home and away goals, the censored zero inflated Poisson distribution and the censored Negative Binomial distribution are compared. Also, the profitability of these models against the bookmakers is investigated (Van Gemert ve van Ophem, 2010).

Joseph et al., (2006) proposed a machine learning algorithms and Bayesian network to predicting the matches outcome (win, lose and draw) for Tottenham hotspur football club, machine learning techniques are Naive Bayesian learner, Data Driven Bayesian, MC4, K-nearest neighbor learner and a decision tree learner. The results showed that Bayesian network outperforms other techniques in predictive accuracy (Joseph ve ark., 2006).

Rotshtein et al., (2005) proposed using a fuzzy knowledge base and based on the outcome of previous matches. They conclude, it is possible to predict the outcome of the match based on previous outcomes (Rotshtein ve ark., 2005).

It is clear from the study of literature in this region that most of the machine learning algorithms were used to predict the results of the matches but were limited to predicting (win, lose and draw).

2.2. Player Injury Prediction

Injuries are a big problem in football and It is considered as the one main factor that prevents football players from not being able to participate in Matches and training, as well as costs of rehabilitation for players. As a result, there is literature on injury prediction in football players. Studies below are the most important studies regarding to prediction player injuries using machine learning techniques:

Rossi et al., (2017) proposed a multidimensional approach to injury prediction in professional football which is based on machine learning and GPS measurements. By using GPS technology, they collect data describing the training workload of players in a professional football club during a season. their show that their injury predictors are both accurate and interpretable by providing a set of case studies of interest to football practitioners (Rossi ve ark., 2017)

Carey et al., (2016) proposed a learning algorithm to predict athlete ratings of perceived exertion (RPE) was studied for Australian football players. The data used was

collected from the global positioning system such as accelerometers and heart rate from 45 players across a full season. The study has proved by using a machine learning approach that RPE can be predicted in Australian football players. Regression modelling outperformed classification approaches and linear approaches (Carey ve ark., 2016)

Kampakis (2016) proposed a learning algorithms to investigate the predictability of football injuries. This work was completed in cooperation with Tottenham Hotspur FC, three investigate were conducted, which are predicting injuries of players, Predicting the recovery time of injuries and predicting an intrinsic injury, for predicting injuries They used Gaussian process model, for Predicting the recovery time of injuries They used negative binomial and ordinal regression as well as Poisson. finally, the third problem of predicting intrinsic injury was solved by using a different type of algorithm which are (supervised PCA, naïve Bayes, random forests, SVM, ANN, Ridge Logistic Regression (RLR) and k-nn) (Kampakis, 2016).

Ehrmann et al., (2016) examines the relationship between variables measured by GPS in gameplay and training, 19 football players competing in the Australian League were monitored for 1 full season using (GPS) units in training and preseason games. Noncontact soft tissue injuries were documented during the season and results proved indicating a raise in training and gameplay intensity leading to injuries (Ehrmann ve ark., 2016)

Kampakis (2011) attempt to detect the possibility of predicting the recovery time of the injured player based on information at the moment of injury, also he used three methods of machine learning (neural networks, support vector machines and gothic processes). The tests were making on data from the Tottenham Hotspur FC. The results of the study show that this task can be done with amount of accuracy (Kampakis, 2011).

Venturelli et al., (2011) examines the factors that increase the risk of muscle pull by using a multivariate survival model (Specifically, Cox regression) for youth players. The study has shown that the previous injuries are the most serious factor. further, proved that an elevated stature increased the probability of muscle pull (Venturelli ve ark., 2011).

Brink et al., (2010) seek to investigate how measures to monitor stress and recovery, and its analysis, provide useful information for the prevention of injuries and sicknesses in elite young football players. The study involved 53 elite footballers aged between 15 and 18. To identify physical stress, football players recorded training, duration of the game and evaluation of the course of stress for two competitive periods through daily

training logs. Using FIFA's standard recording system, injury and sickness data were collected, OR and 95% CIs were calculated for injuries and illnesses using MRA. MR demonstrated that Injuries are related to physical stress (Brink ve ark., 2010).

From the literature study in this region, machine learning algorithms were used to predict the occurrence of injury in the players, especially those related to the heart and muscles and the times of recovery from injury according to the medical analysis of the players.

2.3. Evaluation Players & Select Best Players for Formation

The goal of selecting players and team formation is a complex problem where the final success is specified by how the collection of players forms an effective team. Highly structured models have been developed to support trainers in this domain. Studies below are the most important studies regarding to evaluation players & select best players for formation using machine learning techniques:

Sathe et al., (2017) proposed a machine learning algorithms such as support vector machine, random forest and naïve bayes for English premier league football for making features selection (Sathe ve ark., 2017) .

Vroonen et al., (2017) proposed a projection system for football players called APROPOS which is inspired from the CARMELO system. APROPOS predicts the player potential's via searching in a historical dataset (Vroonen ve ark., 2017).

Soto-Valeroet et al., (2017) proposed using (PCA) in related with a model based Gaussian clustering method in order to describe football players. this model is tested using 40 features from FIFA video for 7705 players. The players were classified according to these roles. They found the dribbling skill is the most distinct variable between different combinations of mixed players (Soto-Valero, 2017).

Asif et al., (2016) presented a unique situation where by a rating system for quantitatively measuring a player’s performance was desired. This would eventually enable predictions derivations on various factors, such as player performance or match outcomes. Data for player rating had to be gathered from different sources; however, this Case Study outlines the solutions that were used to gather such data (Asif ve ark., 2016).

Klaiber (2016) proposed design a statistic based performance rating system which is called the Player Performance Index (PPI) for the Bundesliga (Klaiber, 2016).

Abidin et al., (2016) proposed appropriate research procedure that can be referred to while conducting a Decision Support System (DSS) study, especially when the development activity of system artifacts becomes one of the research objectives. The design of the research procedure was based on the completion of a football DSS development that can help in determining the position of a player and the best team formation to be used during a game. After studying the relevant literature for this research, researchers found that it is necessary to combine the conventional rainfall System Development Life Cycle (SDLC) approach with Case Study approach to help in structuring the research task and phases, which can contribute to the fulfillment of the research aim and objectives (Abidin ve ark., 2016b).

Cotta et al., (2016) proposed using data from FIFA video game as dataset. they justify its use and discuss probable implementations by analyzing two recent widely discussed subjects (Cotta ve ark., 2016).

Uzochukwu et al., (2015) proposed a model that groups the attributes needed for player selection into four major categories which include the player's technique, the player's speed, the player's physical status and the player's resistance using neural network technique to determine these major attributes for each player. The result has shown that Neural Network is a good tool for selecting players in a football team (Uzochukwu ve Enyindah, 2015).

Sarda et al., (2015) proposed a solution for problem of team Selection by using of genetic algorithm to find the best solution for these problem and formation of team. In this paper they created a model which collect the commonly used quantitative approach with some new extensions such as features related personal and team performances along with the collaborative performance of a player in the presence of other players in the team (Sarda ve ark., 2015).

Enefiok et al., (2015) proposed an improved system was developed using fuzzy logic and ANN to help managers in the operation of team selection. The result shows that the new system for decision support has an improved accuracy in determining the player selection decision (Enefiok ve ark., 2015).

Tavana et al., (2013) proposed a model for selecting the best football team formation through two phases, the first to choose the players and the second to choose the best formation. The first phase evaluates the players with a fuzzy ranking and selects

the maximum performers for inclusion in the team. The second phase evaluates the alternative combinations of the selected players with a Fuzzy Inference System and selects the better combinations for team formation. this approach assists the coaches in decision making problems and improves the quality of their decisions. The coaches' judgments are essential in evaluating players; therefore, the efficiency of the model depend on the cognitive abilities of the coaches (Tavana ve ark., 2013).

Kumar (2013) attempt to find a way to classify football players according to the most important attributes of player's performance to find the hidden knowledge which the experts use to assign ratings to players. Researcher performed three classifications experiments and different algorithms from Machine Learning. The better results for predicting ratings using performance metrics had mean absolute error of 0.17 (Kumar, 2013).

Bazmara et al., (2013) proposed K-nn learning algorithm use to evaluate football talents for proper positions considering player skills. The selection of players done by using the proposed method is done using real data, further the results show this method are very efficiency (Bazmara ve Jafari, 2013).

Febianto (2010) proposed AHP decision support system (DSS) to support the ideal placement of a player using multiple criteria to select an appropriate player. DSS would help the trainer make the right decision and use AHP as a model for multiple weighing in the selection process. In the method of data collection techniques, literature, observation and interviews are used for related problems. In addition, techniques and data analysis models using an organized method in which the flow of tools used are a data flow diagram (DFD) and an entity relationship diagram (ERD) (Febianto, 2010).

It is seeming from the literature study in this region that a few models were developed to predict the preferred position of the player in the team, where it was limited to three positions (attack, defense and the midfielder). In addition to, there are a few algorithms used for this purpose.

2.4. Predicting of Player Skill's, Wages and Value

Predicting of player skill's, wages and value may help managers to make suitable decisions like sell, buy and contract renewal. As well as predicting of player's skill like passing, dribbling and ball control is one of the most important ways of talent

identification. Where these skills are the basic technical skills of the player (Reilly ve Holmes, 1983) Especially Dribbling skill is considered critical to the outcome of the match (Huijgen ve ark., 2010) in addition to a previous study (Soto-Valero, 2017) indicated the most discriminating variable among player skills is dribbling. Studies below are the most important studies regarding to predicting of player skill's, wages and value using machine learning techniques.

Dey (2017) proposed a multilayer perceptron neural network to predict the price of a football (soccer) player using data on more than 15,000 players from the football simulation video game FIFA 2017. The network was optimized by experimenting with different activation functions, neurons and layers, learning rate and its decay, Nesterov momentum based stochastic gradient descent, L2 regularization, and early stopping. Simultaneous exploration of various aspects of neural network training is performed and their trade-offs are investigated. final model achieves a top-5 accuracy of 87.2% among 119 pricing categories and places any footballer within 6.32% of his actual price on average (Dey, 2017).

Yaldo et al., (2017) proposed an objective quantitative method for determining the wages of football players based on their skills. By using data for 6082 players, the experimental results that the Pearson correlation is ~0.77 ($p < 001$) between the actual and expected salary of the players (Yaldo ve Shamir, 2017).

He et al., (2015) showed how the market value of players and their performance of La Liga players can be designed by using extensive data sources using machine learning techniques (He ve ark., 2015).

From the study of literature in this region. We noted that a number of models have been developed to classify players and national teams according to their performance, and a model has been developed to predict the market value of the player according to his skills and performance.

2.5. Football Analytics

Sports analysis is the use of quantitative data analysis of performance data to support training decisions. Sports analysis is not only analysis of performance data, but also analysis with a clear practical purpose. Sports analysis it also useful for making training programs, building strategies and development of game, and player recruitment

(Schulenkorf ve Frawley, 2016). Recently with technological advances, a number of commercial companies have emerged to provide data collection and analysis for the sports elite. The provision of tracking data from matches had led to an explosion of interest in the area of football analytics but research in football analytics with Machine Learning techniques is limited and involves on analyzing football game play like formation identification (Vroonen ve ark., 2017). In fact, there are many Sports Analytics Companies which are providing vast volumes of data in statistical packages and data visualizations and the most important of these companies are Prozone and Opta Sports. researches below are the most important studies related to football analyzes:

Wagenaar et al., (2017) explored how to use the machine learning to predict the opportunities for achieving goals in the football of the position data. they propose the use of deep learning convolutional neural networks for this problem. The results show that the Google Net architecture better than all another method with an accuracy of 67.1% (Wagenaar ve ark., 2017).

Brooks et al., (2016) proposed design a player ranking system called (novel) according to the value of passes completed. This value based on the relation between pass locations in a possession and shot opportunities generated. The data used to build the model was taken from La Liga for 2012-2013 season (Brooks ve ark., 2016).

Sgro et al., (2016) analyses the differences amongst the technical performance profiles of the teams involved in the 2016 European Football Championship. A k-means cluster analysis was preliminarily performed to identify the close matches of that tournament. Then, the team-match statistics gathered from the official website of the Union of European Football Championship (UEFA) (Sgro ve Lipoma, 2016).

Horton et al., (2014) proposed constructed a Framework for classifying passes made during a football match according to the quality of the pass and rates each pass as Good, OK or Bad. where it takes player trajectories and a list of passes made. The chosen approach is to use supervised machine learning algorithms in order learn the classification function. The experiments were conducted on five classifiers. First, they used multinomial logistic regression with three different regularized cost functions. Second, they used classifiers RUSBoost and Support Vector Machine algorithms. in general, they produced a classifier with 86% accuracy on the pass labelling mission (Horton ve ark., 2014).

Lasek et al., (2013) provided an overview of the predictive ability of various rating systems of football teams. The main benchmark was the FIFA ranking. Their experiences have shown that this system can outperform FIFA ranking (Lasek ve ark., 2013).

Gedikli et al., (2007) proposed system called ASPOGAMO. ASPOGAMO is a vision system have ability to estimating motion trajectories of football players taped on video. The system achieves a high level of robustness through the use of model-based vision algorithms for camera estimation and player estimation (Gedikli ve ark., 2007).

From the study of literature in this region, we noted that there are a few studies have been conducted for analysis sports (especially in football) due to lack of data. Some sports analytics data has been conducted by using video games such as FIFA Soccer, PES and Football Manager.

3. GENERAL CONCEPTS AND INFORMATION ABOUT FOOTBALL

3.1. Definition of Football and its Importance

Football is sport that played between two teams, each team have 11 players with a spherical ball. The aim of the game is to score the goals by kicking the ball. Football played in over 200 countries, So it is considered the most famous sport in the world in both number of spectators and players (Dunning, 1999).

Recently, Football have become a new emerging industry, where the revenue for European football clubs for 2017 rated at $ 27 billion (Davis, 2017). Therefore, the European clubs have become trade organizations (Moor, 2007).

3.2. Rules and Facts of Game

There are 17 laws in football game, (Association, 1995) which are:

3.2.1. Play field

The game played on natural surfaces, the surface should be green and have rectangular shape. The long side of the rectangle called side lines and ranges between 100 and 110 meters and have ranges between 64 and 75 meters. The field is split in half by the center line Figure 1.3.

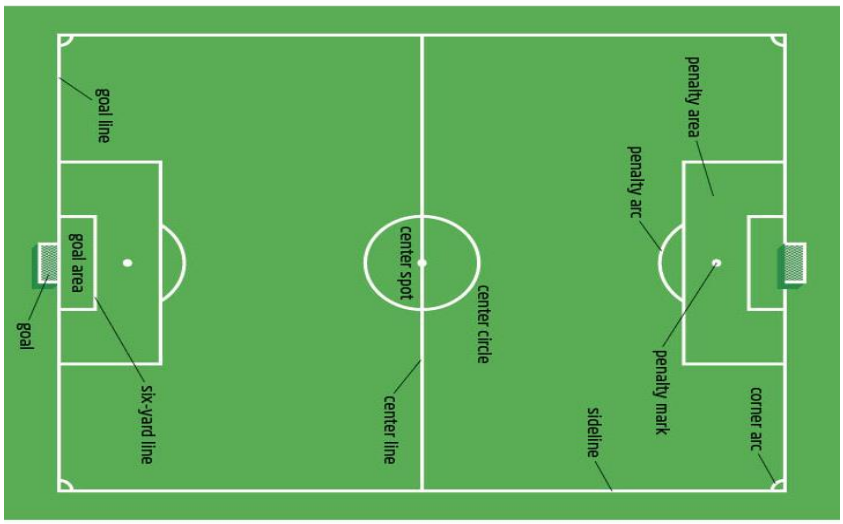


Figure 3.1. Field of Play in Football Game

3.2.2. Ball

Ball is should spherical, made from leather and Its perimeter shall not exceed 70 and not less than 68 in specific pressure.

3.2.3. Players number

Football matches consist of two teams, each with 11 players, and each team is allowed to switch 3 players during the match. Players are classified into three groups: Defender, midfielder and Forward player.

3.2.4. Equipment

In the match Players should wear jersey shirt, shorts, high knee socks, guards for shin and footwear.

3.2.5. Referee

He is the supervisor of the game, has the power to make decisions, apply laws and declare the outcome, from a neutral point of view.

3.2.6. Assistant referees

In the football game, the assistant referee is one of the officials who help the referee manage the game. Two of them are called assistant referees, standing on the line of contact, while the fourth referee assists the referee in managing the match and all related matters as directed by the referee.

3.2.7. Duration of the match

Football game is played in 90 minutes Divided into two halves and the duration of each half is 45 minutes. The duration between two halves is 15 minutes. According to referee estimates, it can be compensated more time for any interruption during play.

3.2.8. Start and restart of play

The starting kick is a way to start or resume a game and is executed at the start of the game or after scoring a goal. When the starting kick is made, all players must be in their own half, and the opponents must be at least 9.15 m from the ball.

3.2.9. Ball in and out of play

When the ball is outside of play a goal have scored. otherwise the ball is in play at any times.

3.2.10. Scoring methods

The goal is calculated when the entire soccer passes over the goal line and under the crossbar provided that the team that scored the goal has not committed a violation of the rules of the game beforehand.

3.2.11. Offside

Offside occurs when the player "any part of the head, body or feet" is closer to the line of his opponent than the ball, the moment the ball passes him, not at the moment of receiving him.

3.2.12. Fouls/Misconduct

Fouls are many and varied and generally occur as a result of the use of excessive force in playing in a deliberate or unintended way. The referee may offer a yellow card to warn players and red card to exclude players from game.

3.2.13. Free kicks

Given due to players' mistakes. A free kick may be either be direct or indirect. from a direct free kick, goal can be scored.

3.2.14. Penalty kicks

Are given when a player doing any Fouls or Misconduct in his own Penalty box. the ball is kicked from the penalty region.

3.2.15. Throw in

Is a way of restarting play in a match when the ball has exited the side of Play Field?

3.2.16. Goal kick

Is a way of restarting play after a goal?

3.2.17. Corner kick

Is a way of restarting play in a match and Are given when the ball goes outside of border along the end line and was last touched from the defending team?

3.3. Player Attributes

The player's attributes represent his skills and are the most important factor in determining his performance. player's attributes are divided into three categories are technical, mental and physical.

3.3.1. Mental Attributes

A player's mental attributes indicate the player's sound and stable in matches and is when performing training. Generally, players with high mental toughness will be more consistent even when suffering from bad Morale. mental toughness consider very important in any environment that requires performance setting, adversities and challenges (Miçoogullari ve ark., 2017). The main mental attributes are:

i. Aggression

Indicates the player's desire to participate in the game and how aggressive it will be in tackling.

ii. Composure

Indicates the ability of the player to be calm and professional regardless of the situation of the game.

iii. Interceptions

Indicates the ability of player to read the game and intercept passes during any particular moment in a match.

iv. Marking

Marking is the ability to mark, defend and track an opposing player. it is also player’s ability to stay close to an opposing attacker to stop him from a pass or cross from a teammate.

v. Positioning

Refers to the player's ability to judge the play properly and move to a strategic place when he does not control the ball on the defense.

vi. Vision

It refers to a player’s mental awareness about position of his teammates for passing the ball to them.

3.3.2. Physical Attributes

A physical like speed, height, balance, strength and agility are all very important in the football game, we notice the most growth of the Physical Attributes during youth of players And it will develop naturally during this period, where studies have shown

increased physical activity in children and young people at an early age (ŞİMŞEK ve ark., 2014). The main Physical attributes are:

i. Acceleration

Indicates the player's fast to reach his highest running speed.

ii. Agility

Indicates the ability of the player to change directions quickly or stop, especially during dribbling.

iii. Reactions

Reactions measure a quickly of a player to responds for a situation happening around him.

iv. Sprint Speed

Sprint speed measures the speed rate of a player’s sprinting.

v. Stamina

It determines the average at which a player will tire during a match.

vi. Strength:

It the player physical strength. The higher the strength, the more probable the player will win a physical challenge.

vii. Balance

Indicates the player's ability to maintain balance after challenged by a tackle, or any physical challenge.

3.3.3. Technical Attributes

A Technical like passing, shooting, dribbling and Finishing, all these skills are learned and practiced from player. Often, these skills combined and used in unusual ways (Giacomini, 2009). The main technical attributes are:

I. Curve

it is used to measures ability of player to curve the ball when shooting or passing.

II. Ball Control

It is the ability of a player to control in the ball when he receives it.

III. Finishing

Indicates the player's power and accuracy of any given shot using foot.

IV. Crossing

Indicates to the accuracy of a player's ability when performing a cross pass during normal running or free kick.

V. Dribbling

it is used to measures ability of player to get around defenders.

VI. Free Kick

Indicates the ability of player to kick a free kick.

VII. Heading

Indicates the ability of player to accurately head the ball.

VIII. Passing

Indicates the accuracy of all the passes of the player.

IX. Penalties

Indicates the ability and accuracy of the player to shots penalty.

X. Tackling

Indicates the ability and accuracy of the player to tackles.

4. DECISION SUPPORT SYSTEM AND MACHINE LEARNING

The current review contributes to a comprehensive review of decision support systems and the most important machine learning algorithms and their integration into sport.

4.1. Decision Support System

Decision Support Systems (DSS) refer to the role of computers in the decision-making process. For some writers, DSS mean are management-level information systems that link data, complex analytical models, and data analysis tools to support

decision making. whilst others consideration it as an extension for management techniques where it serves the management level of the organization and help managers make their own unique and fast-changing decisions that are not easy to identify in advance (Keen, 1980).

According to (Simchi-Levi ve ark., 2007) DSS is an analytical tool to help operations in addition to production planning. The DSS may range from simple tool to intelligent systems (Intelligent Decision Support System).

According to (Power, 2002) DSSs consists of three parts (as in Figure 4.1) are, Database, Software system & user interface.

A- DSS Database: Database it contains information from different sources, including inner information from the organization and the external information extracted from the Internet, and so on. DSS Database can be a big data warehouse or a small database.

B- DSS Software System: It consist of different mathematical and analytical models that are utilized to analyze the data, in this way producing the important data. A model predicts the result based on various inputs or various conditions to produce the wanted output.

The main commonly used statistical and mathematical DSS models are:

Statistical Models: They contain an extensive variety of statistical functions, for example, mean, median, mode, standard deviations and so on. These models are utilized to build up connections between the events and different variables identified with that events.

Optimization Analysis Models: They are utilized to discover ideal value for an aim variable under given conditions. They are generally utilized to making decisions associated with ideal use of resources in an association.

Forecasting Models: They utilize different forecasting instruments and strategies, such as regression models, time series analysis etc., to make prediction for something in advance. They give data that aides in analyze the business conditions.

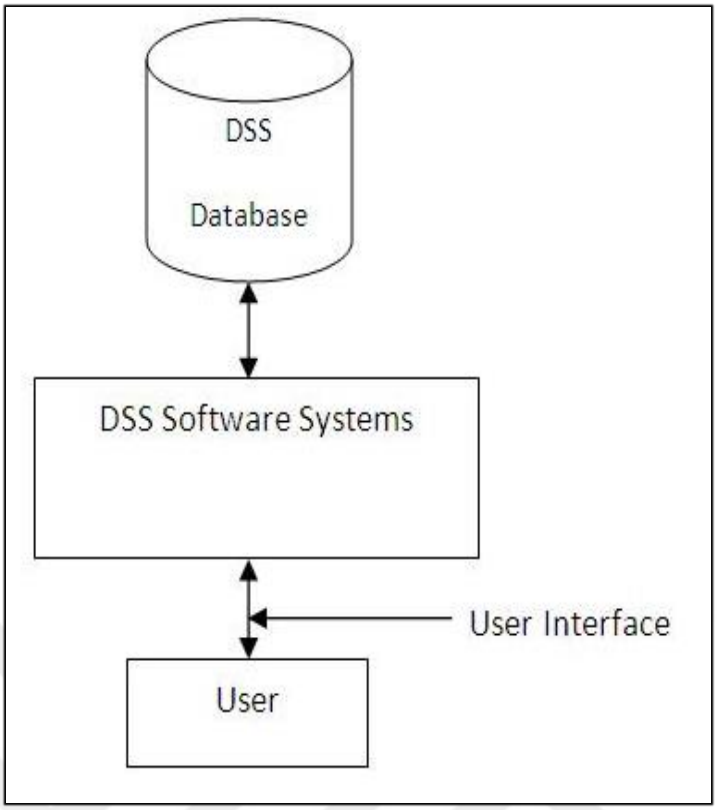


Figure 4.1. Block diagram of DSS

C- DSS User Interface: It is an intelligent graphical interface which makes the interaction less demanding between the DSS and clients. It shows the outcomes of the analysis in different forms.

4.2. Intelligent Decision Support System (IDSS)

According to (Sarma, 1994) Intelligent Decision Support System (IDSS) is a decision support system (DSS) that can use artificial intelligence (AI) techniques in one or all of its components. From a long history, artificial intelligence techniques are used in management information systems, where it called Knowledge based systems (KBS), but the term of (intelligent systems) is thought to originated with Holsapple in 1978 (Holsapple, 1978).

Currently IDSS provides decision support through different techniques like (data mining, data analytics, machine learning, etc.).

The most common of IDSSs are:

- Text analytics and data mining based on DSS.

- Internet of things and ambient intelligence based on DSS.
- Remote sensing and integration with DSS.
- Biometrics based on DSS.
- Recommender and expert systems,
- Computer vision (CV) based on DSS.
- Fuzzy set DSS
- Sensor DSS
- Robotic DSS.
- Adaptive DSS.

Although the importance of large scale data analysis for design, there are only a few studies have tried to analyze scale data in the context of data analytics (Ma ve ark., 2014). Data analytics, recently developed in different areas of human activity, but also they are rarely integrated with DSS (Kaklauskas, 2015). DSSs based on data analysis and have been used for many sectors like tourism, aviation, medicine, education and sports (like rugby, tennis, basketball & football).

In football, many of DSS have been built. Some of these focused-on team selections, scheduling league and identifying talent in football players. But there are very few researches about algorithms or decision support systems have been made that could be used by decision makers to aid in the process of forming a strong team (Abidin ve ark., 2016b).

4.3. Machine Learning Methods for Intelligent Decision Support

Machine learning is one branches of artificial intelligence which are provides systems have the capability to automatic learn to find structured patterns from data sets. We need machine learning in matters where we unable write a computer program to fix a given problem, as well as we need for examples data or experience for use machine learning Tanique's (Alpaydin, 2014).

Machine learning usually divided into two groups; supervised and unsupervised further there are type fall somewhere in between supervised and unsupervised learning called (semi-supervised).

Decision support systems have a complex software with a rich user interface and database. The intelligence of the decision support system is programmed by a set of machine learning algorithms which are a set of selected mathematical models. The aim

of the algorithms is to extract the information existent in the data and present it to the user in an understandable shape. Model structures can vary a lot. Typical models used are decision trees, artificial neural networks, random forest ... etc. (Rasku ve ark., 2014).

4.4. Classification of Machine Learning Algorithms

Algorithms are most often categorized in two group:

- Algorithms classified by learning style
- Algorithms classified by similarity

4.4.1. Algorithms classified by learning style

The general target of machine learning is to produce intelligent programs, or models, through a process of learning, Therefore, the type of learning adopt by the algorithm must be considered first. Two types of machine learning algorithms are commonly used today, "supervised" and "unsupervised" further there are type fall somewhere in between supervised and unsupervised learning called "semi-supervised". In a supervised learning mode, what have been learned in the past is used to analyze new data, whilst unsupervised algorithms are able to inferring from new datasets.

4.4.1.1. Supervised learning

In machine learning and the artificial intelligence, Training data in Supervised Learning includes both the input and the desired results. Input and output data are classification to use in learning in new data processing, these methods are fast and accurate (Donalek, 2011).

4.4.1.2. Unsupervised learning

In Unsupervised Learning Algorithms is not classified Input data and not provided with the correct results during the training. A model is designed by conclude structures existent in the input data. This perhaps to extractor general rules. It perhaps during an

arithmetical process to systematically minimize redundancy, or it perhaps to organize data by likeness (Donalek, 2011).

4.4.1.3. Semi-Supervised learning

Semi-supervised Learning Algorithms, is halfway between supervised and unsupervised learning, where the Input data is a mixture of classified and unclassified examples. There is a required prediction problem but the model should learn the structures to organize the data further make predictions (Chapelle ve ark., 2009).

4.4.2. Algorithms classified by similarity

Algorithms that have been classified according to similarity, are classified according to their works. For example, neural network inspired methods and tree-based methods Figure 4.2.

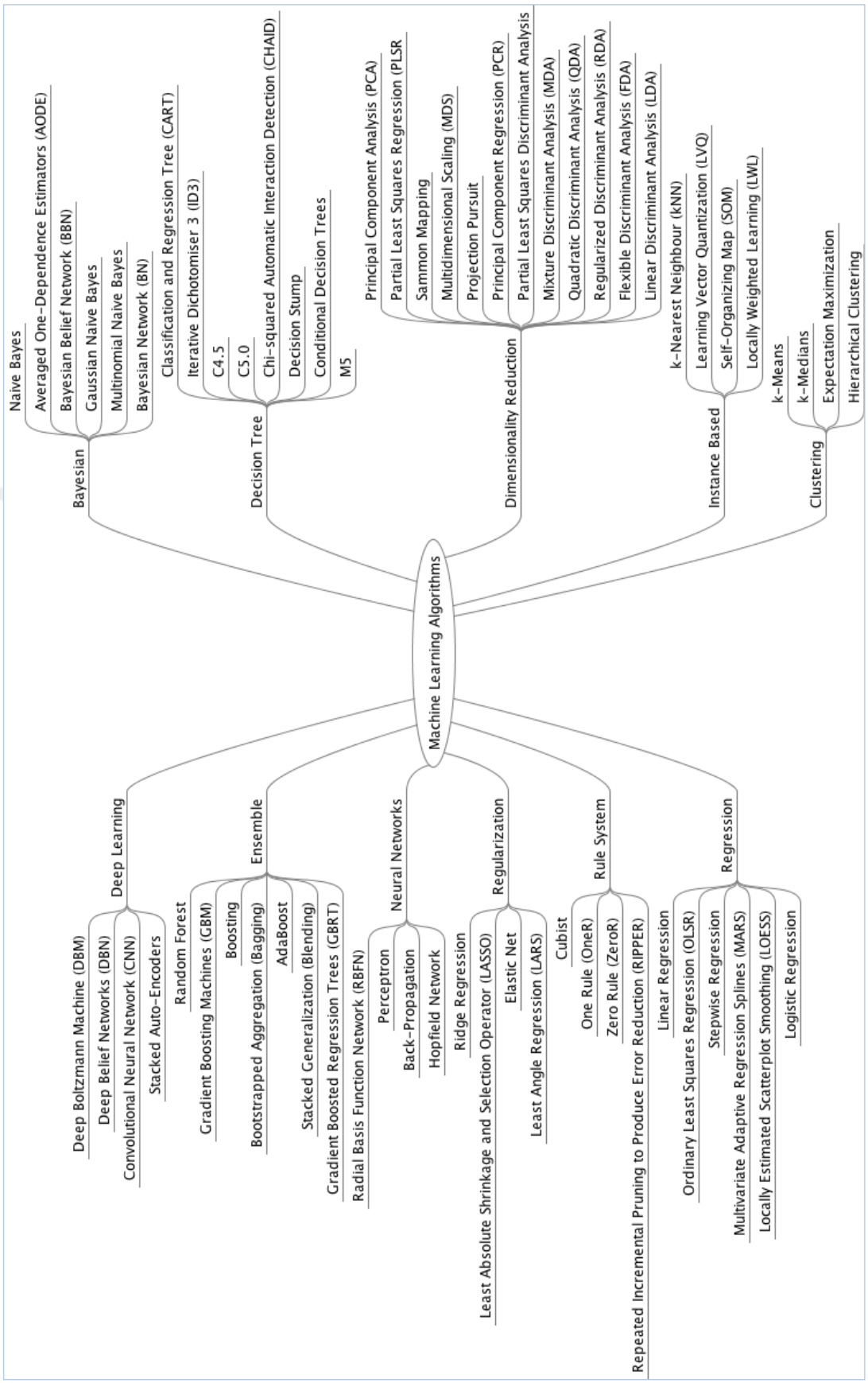


Figure 4.2 The Main Types of Machine Learning Algorithms (Brownlee, 2016)

4.4.2.1. Regression algorithms

Regression involves modeling the relationship among variables that are frequently repeated using a scale of error in the predictions made by the machine learning model. Regression methods are the basis of statistics and have been adopted in statistical machine learning (Ibrahim, 2015).

The most popular regression algorithms are:

- Linear, Logistic and Stepwise Regression
- Multivariate adaptive regression splines
- Ordinary least squares multiple regression

4.4.2.1.1. Linear regression

linear regression is a way to modelling the relationship between dependent variable (y) and one or more independent variables (x). The case of one independent variable is called simple linear regression (SLR). For more than one independent variable, the model is called MLR (multiple linear regression) (Seltman, 2017).

In a simple linear regression (As Figure 4.3) is a straight line passing through a set of points in such a way that the sum of the remaining square of the model is as low as possible. This indicates the fact that regression is one of the simplest methods used in the field of statistics where the slope of the line is the relationship between y and x corrected by the standard deviations of these variables.

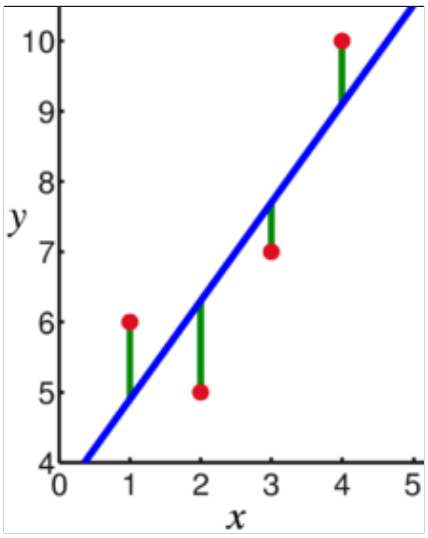


Figure 4.3. Relation between the independent variable (x) dependent variable (y)

It is common to assume that the ordinary least squares (OLS) must be used to minimize the residuals (vertical distances among fitted line & the points of the data). according to these hypotheses, line accuracy through the sample points is calculated by the sum of squared residuals (SSL) and the aim is to make this sum is small as possible.

4.4.2.1.1.1. Fitting the regression line

Suppose that you have n number of points (Xi, Yi), i = 1,2, ... n}, the function that describes Y and X is: $Yi = \alpha + \beta Xi + \varepsilon_i$ the goal is to find the line equation $y = \alpha + \beta x$ that gives the best representation of points. Here the best is known as the small squares method: the line that reduces the sum of the remaining squares of the linear regression model. In other words, α (point of intersection with y axis) and β (slop) are involved in solving the following reduction problem (Equation 4.1):

$$\text{Find } \min_{\alpha,\beta} Q(\alpha,\beta), \qquad \text{for } Q(\alpha,\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \quad \ldots\ldots (4.1)$$

Using the calculation of the internal space geometry of the model to obtain a second-order equation in α and β, it is possible to find the values of β and α that reduce the function Q as Equation 4.2 (Kenney, 1962)

$$\begin{aligned}
\hat{\alpha} &= \bar{y} - \hat{\beta}\,\bar{x}, \\
\hat{\beta} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{\operatorname{Cov}(x,y)}{\operatorname{Var}(x)} \\
&= r_{xy}\frac{s_y}{s_x}. \qquad \ldots\ldots (4.2)
\end{aligned}$$

Where:

r_{xy} is the correlation coefficient between x and y

s_x and s_y are the uncorrected sample standard deviations of x and y

Var and Cov are the variance and sample covariance.

Substituting the above we have:

$$f = \hat{\alpha} + \hat{\beta}x, \quad \ldots\ldots (4.3)$$

4.4.2.1.1.2. Scoring linear regression model (R^2)

A common method of measuring the accuracy of regression models is to use the coefficient of determination (R^2) statistic.

The R2 statistic is defined as Equation 4.4

R2 = 1- RSS/TSS ……...…… (4.4)

$$ESS = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 \quad \text{……...……...……} \quad (4.5)$$

$$RSS = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \quad \text{……...……} \quad (4.6)$$

- The RSS (Residual sum of squares) measures the variability left unexplained after execute the regression
- The TSS (Total sum of squares) measures the total variance in Y, where it summation of ESS & RSS (Figure 4.4).
- Therefore, the R^2 statistic measures proportion of variability in Y that is explained by X (Features) Figure 4.4.

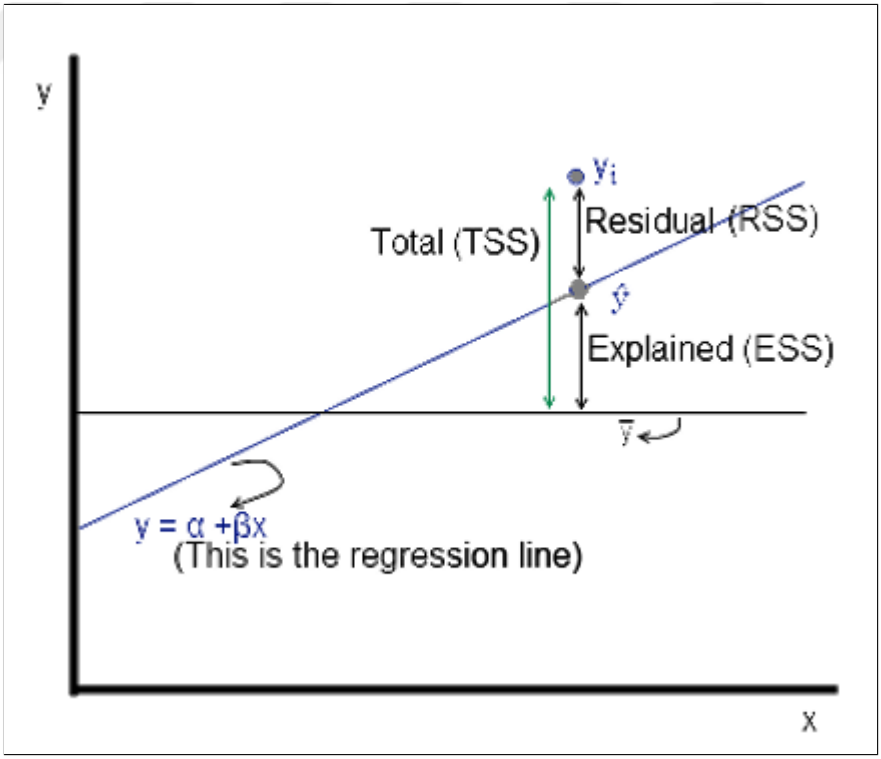


Figure 4.4. Relationship between TSS, RSS and ESS

4.4.2.1.2. Logistic regression

In probability, logistic regression (LR) is a model used to predict the probability of an event by matching the data on a logistic curve. Logistic regression uses several expected variables that can be numeric or factional. For example, a person may have a heart attack over a certain period of time that can be predicted by information about the patient's age, sex, and body mass index. Logistics regression is widely used in medicine and social sciences, and is used in marketing to calculate the consumer's inclination to buy a product or refrain from purchasing (Seltman, 2017).

The logistic regression was developed by EMI David Cox in 1958. The binary logistic regression model is used to estimation the probability of a binary response according to one or more independent variables. The logistic regression model is used for probability, although it can be used for classification. for example by choosing a cutoff value and classifying input variables with probability greater than the cutoff as one class (Cox, 1958).

The definition of logistic regression begins by defining the logistic function (Equation 4.7) and is like probability theory taking values between 0 and 1 as in Figure 4.5.

$$f(z) = \frac{e^z}{e^z + 1} = \frac{1}{1 + e^{-z}} \quad \text{.....} \quad (4.7)$$

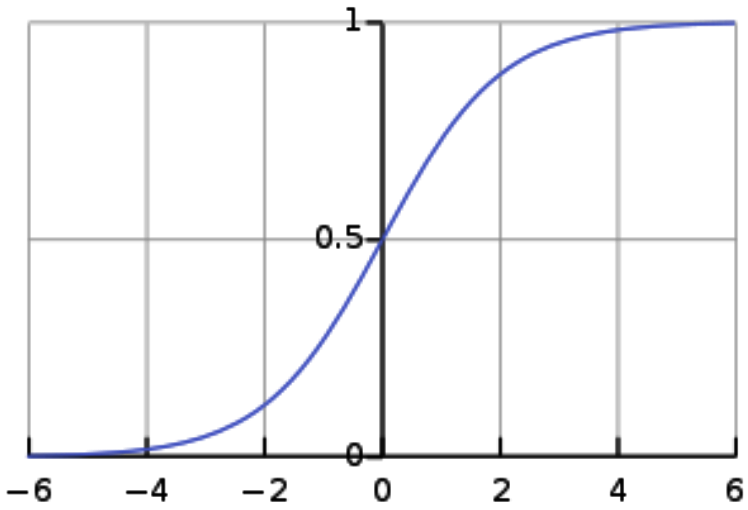


Figure 4.5. The standard logistic function

The logistic function is important because it takes input from positive infinity to negative infinity, but outputs are always between zero and one. The variable z represents the independent variables where $f(z)$ is probability of a given output of a set of independent variables, and z is the sum of the contribution of all independent variables used in this model, the variable z is defined as Equation 4.8

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k \quad \text{.....} \quad (4.8)$$

β_0 Is the intercept from the equation of linear regression

β_1 ،β_2 ،β_3 Are regression coefficient

Logistic regression is a useful way to clarify the relation between independent variables (age, gender, etc.) and dependent, which takes two different values. An example of a person with cancer is that the values for the response variable are either "cancer" or "without cancer".

4.4.2.2. Instance-based Algorithms

Instance-based model is a decision matter with examples or instances of training data that are consider important to the model. Such these methods usually build up a dataset of instances and compare new data to the dataset using a similarity likeness in order to find the better results.

4.4.2.2.1. k-nearest neighbors algorithm

k-nearest neighbors algorithm (KNN) define as a non-parametric method used for regression and classification problem (Altman, 1992). In regression and classification, the input consists of the k close training examples in the feature space. The KNN algorithm is consider the simplest algorithms of machine learning.

k-nearest neighbors measure the distance **dE (Xi, Xj)** among query points **Xi** and a set of training **Xj** to classify a new object set according to majority of K-nn category of Y attributes of training samples (Equation 4.10)

Query point **Xi** = x1, x2, x3 …. xn

Training Sample **Xj**= xl, x2, x3… xn

$$Dist(c_1, c_2) = \sqrt{\sum_{i=1}^{N} \left(attr_i(c_1) - attr_i(c_2) \right)^2}$$

$$k - NearestNeighbors = \left\{ k - MIN(Dist(c_i, c_{test})) \right\}$$

$$prediction_{test} = \frac{1}{k} \sum_{i=1}^{k} class_i \ (or \frac{1}{k} \sum_{i=1}^{k} value_i) \quad \text{..... (4.10)}$$

Below explain how it work (Mulak ve Talhar, 2015) :

1. Determine number of K.
2. Calculate distance between points (Euclidean or Manhattan)
3. Determine K-nn minimum distance
4. collect category Y values of nearest neighbors
5. Use simple majority of K-nn to predict value.

4.4.2.3. Ensemble algorithms

Ensemble methods are models consisting of multiple weak models that are independently trained then whose predictions are combined to predict in general. Ensembles typically achieve superior model performance over singular methods. However, Ensembles consider a very powerful technique and very popular. Random Forest and Gradient Boosting Machines (GBM) are both common Ensemble decision tree.

4.4.2.3.1. Random forest

A Random Forest (RF) is a classification method that consists of several uncorrelated decision trees. All decision trees have grown under a certain kind of randomization during the learning process. For a classification operation, each tree in that forest may decide and the class with the high votes decide the final classification. Random Forests can also be used for regression.

The term Random Forest was placed by Leo Breiman in 1999 (Breiman, 2001). He explored various methods of randomization of decision trees, for example by means of bagging or boosting. His work was preceded by the research of Tin Kam Ho in 1995 (Ho, 1995).

4.4.2.3.1.1. Characteristics of random forest

- The classifier trains very fast: This advantage results from the short training or setup time of a single decision tree and the fact that the training time for a random forest increases linearly with the number of trees.

- The evaluation of a test example happens on each tree individually therefore it evaluates so fast.

- It is very efficient for large scale data (many classes, many training examples, many features).

4.4.2.3.1.2. How random forest work

To understand and use different options, it's helpful to learn more about how to calculate it. Most options depend on two data objects created by random forests.

When we want the training data for the tree to be drawn by sampling with the substitution, about one-third of the cases are left outside the sample. These data are used outside of the bag (oob) to obtain an unbiased estimate of the rating error when adding decision trees to the forest.

After each tree is built, all data is run under the tree, and the rounding points are calculated for each pair of cases (As in Figure 4.6), If there are two cases that take place in the same terminal node of tree, their proximity increases by one. At the end of the run, the views are rounded by dividing the number of trees.

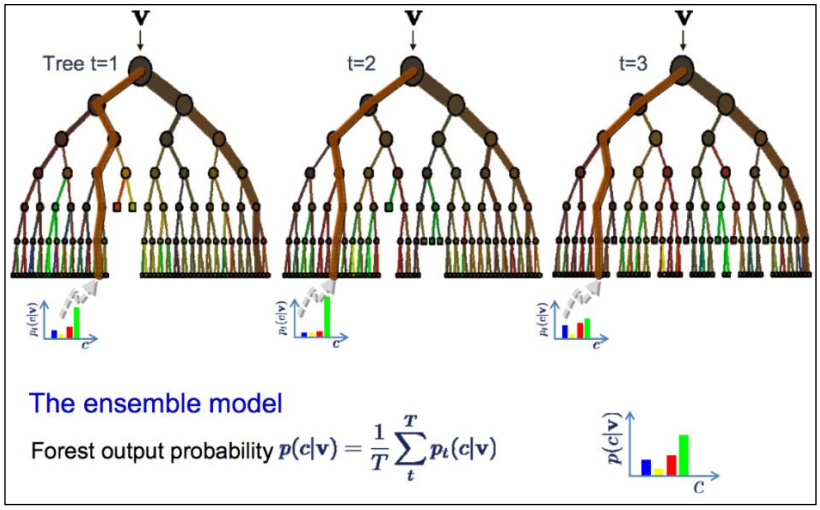


Figure 4.6 Random forest classifier (Sun ve ark., 2017)

4.4.2.4. Dimensionality Reduction Algorithms

In machine learning, reducing dimensions is the process of reducing the number of random variables under investigation by obtaining a set of basic variables. Can be divided into extract feature and feature selection (Pudil ve Novovičová, 1998). dimension reduction can be considered useful for visualize dimensional of data or to simplify data, then can be use these data for classification and regression.

4.4.2.4.1. Feature Extraction

Extracting features converts data in high-dimensional space to lower space dimensions. The data transformation may be linear, as in (PCA), but there are also many techniques to reduce non-linear dimensions.

4.4.2.4.1.1. Principal component analysis (PCA)

The main linear reduction technique, PCA, now let explain PCA Mathematics:

Suppose there is a sample with n individuals for each of which m (random) variables have been measured ***FJ***. The PCA allows you to find a number of underlying factors ***p<m*** that explain approximately the value of the m variables for each individual. The fact that there are these underlying factors can be interpreted as a reduction in the dimensionality of the data: where before we needed m values to characterize each individual, we now only need ***p*** values. Each of the ***p*** found is called the Principal component, hence the name of the method.

There are two basic ways to apply the PCA (Shlens, 2014) :

Method based on the correlation matrix, when the data are not dimensionally homogeneous or the order of magnitude of the random variables measured is not the same. consider the value of each of the m random variables *Fj*. For each of the *n* individuals, take the value of these variables and write the data set in the form of a matrix:

$$(F_j^{\beta})_{j=1,\dots,m}^{\beta=1,\dots,n}.$$

Note that each set

$$\mathcal{M}_j = \{F_j^{\beta} | \beta = 1, \ldots, n\}$$

Can be considered a random sample for the variable ***Fj***. From the (m x n) data corresponding to the m random variables, the sample correlation matrix can be constructed, which is defined by:

$$\mathbf{R} = [r_{ij}] \in M_{m\times m} \qquad \text{donde} \qquad r_{ij} = \frac{\text{cov}(F_i, F_j)}{\sqrt{\text{var}(F_i)\text{var}(F_j)}} \quad \text{………… (4.11)}$$

Due to the previous property these m eigenvalues receive the name of weights of each one of the m main components. The main factors identified mathematically are represented by the base of eigenvectors of the matrix **R**. It is clear that each of the variables can be expressed as a linear combination of eigenvectors or principal components.

Method based on the covariance matrix, which is used when the data are dimensionally homogeneous and have similar average values.

The objective is to transform a given set of data X of dimension n x m to another set of data Y of smaller dimension n x l with the least possible loss of useful information by using the covariance matrix.

It starts from a set n of samples each of which has m variables that describe them and the objective is that, each one of those samples, is described with only I variables, where l <m. In addition, the number of major components l must be less than the smallest of the dimensions of X (Equation 4.12).

$$l \leq \min\{n, m\}$$

$$\mathbf{X} = \sum_{a=1}^{l} \mathbf{t}_a \mathbf{p}_a^T + \mathbf{E} \quad \text{………………………………………………………….. (4.12)}$$

4.4.2.4.2. Feature selection

Feature Selection (FS), is a machine learning approach that uses only a subset of the available features for a learning algorithm. FS is necessary because it is sometimes technically impossible to include all features because there are differentiation issues when there are a large number of features. There are three strategies: Filter Strategy, the Wrapper Strategy, and the Embedded Strategy (Guyon ve Elisseeff, 2003).

4.4.2.4.2.1. Filter strategy

The filter feature selection ways apply a statistical measure to allocate a rating to each feature. The features are arranged by result or selected to be retained or removed from the data set.

An example of filter methods are correlation coefficient scores.

4.4.2.4.2.2. Wrapper strategy

Wrapper ways use a predictive model to record subsets of features. Each new subset is used to train a model, which is tested on a train and test split set (hold-out). The calculation of the number of errors made to hold-out (the error rate of the model) gives the rate for that subset. Because wrapper methods train a new model for each subset, they are highly computationally intensive, but typically provide the best performance feature for this particular type of model.

An example if a wrapper method is the Recursive Feature Elimination algorithm.

4.4.2.5. Artificial neural network algorithms

Artificial Neural Networks (ANN), are networks of artificial neurons. ANN models are inspired by the function and/or structure of biological neural networks (AS Figure 4.7). ANN models are commonly used for classification and regression problems. ANN consider as an enormous subfield, because it consists of hundreds of different algorithms for all kinds of problems (Staub ve ark., 2015).

4.4.2.5.1. Description

Artificial neural networks are mostly based on the cross-linking between many neurons. The topology of a network (the assignment of connections to nodes) must be well thought-out, depending on its task. After the construction of a network, come the training phase in which the network learns. Theoretically, a network can learn in a smart way to reach an acceptable degree of expected solutions. ANN creates interconnected

neurons with different weights and each neuron is responsible for one input. For example, we have a problem consisting of five inputs and one output. Therefore, when we configure the grid, we actually have created a network of 5 neurons and each neuron responsible for one input. In training, the network continuously adjusts the weights associated with each neuron so that the output is closest to reality. The training process continues on all inputs available and the weights associated with each neuron are adjusted, provided that the resulting value is as close as possible to the true output value. Therefore, the training process and the abundance of data is very important to make the network predictions as close as possible to reality.

ANN neurons consist from 3 components are:

i- a set of connecting links

ii- an adder function

iii- activation function

The connecting links identify by a weight: W1, W2, ... Wn. An adder function seeks to compute the weighted sum of the inputs. Finally, Activation function (squashing function) seek for limiting the range of the output of the neurons.

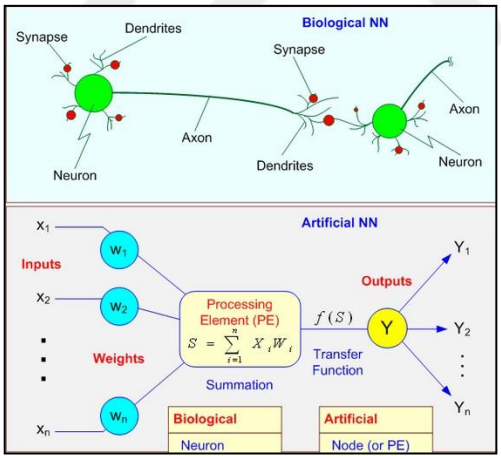


Figure 4.7. Neuron in Artificial neural network vs Biological neural networks (Staub ve ark., 2015)

4.4.2.5.2. Perceptron

Perception is a simplified artificial neural network, first introduced by Frank Rosenblatt in 1958. It consists in the basic version (simple perceptron) of a single artificial neuron with adjustable weights and a threshold. This term is understood to mean various combinations of the original model, distinguishing between single-layered

and multi-layer perceptron (MLPs). The principal operation is to convert an input vector into an output vector and thus (Reber ve Perrig, 2001).

4.4.2.5.2.1. Single-layer perceptron

In the single-layer perceptron, there is only one layer of artificial neurons, which at the same time represents the output vector. Each neuron is represented by a neuron function and receives the entire input vector as a parameter. The processing is very similar to the so-called Hebbian learning rule for natural neurons. However, the activation factor of this rule is replaced by a difference between the setpoint and the actual value. Since the Hebbian learning rule refers to the weighting of the individual input values, the learning of a perceptron takes place by adjusting the weighting of each neuron. Once the weights have been learned, a perceptron is also able to classify input vectors that differ slightly from the original learned vector.

4.4.2.5.2.2. Perceptron learning rule

There are several versions of the learning rule to deal with the perceptron. For a perceptron with binary input and output values, the learning rule is specified here. This rule only converges if the training record is linearly separable.

The following considerations are based on the learning rule of the perceptron:

i. If the output of a neuron is 1 (or 0) and should take the value 1 (or 0), then the weights will not be changed.
ii. If the output is 0, but should take the value 1, then the weights are incremented.
iii. If the output is 1, but should take the value 0, then the weights are decremented.

Mathematically, the facts are expressed as Equation 4.12 and 4.13

$$w_{ij}^{\text{neu}} = w_{ij}^{\text{alt}} + \Delta w_{ij}, \quad \text{.....} \quad (4.13)$$

$$\Delta w_{ij} = \alpha \cdot (t_j - o_j) \cdot x_i. \quad \text{.....} \quad (4.14)$$

Δw_{ij} the change of the weight ***ij*** for the connection between the input ***i*** and output ***j***.

t_j the desired output of the neuron ***j***

o_j the actual output.

x_i entering the neuron

4.4.2.5.2.3. Multi-layer perceptron

The limitation of the single-layer perceptron could later be solved with the multilayer perceptron, in which there are at least one further layer of hidden neurons in addition to the output layer (hidden layer). All neurons in a layer are fully linked to the neurons of the next layer (As in Figure 4.8). In many applications, the modules of these networks implement the sigmoid function as an activation function.

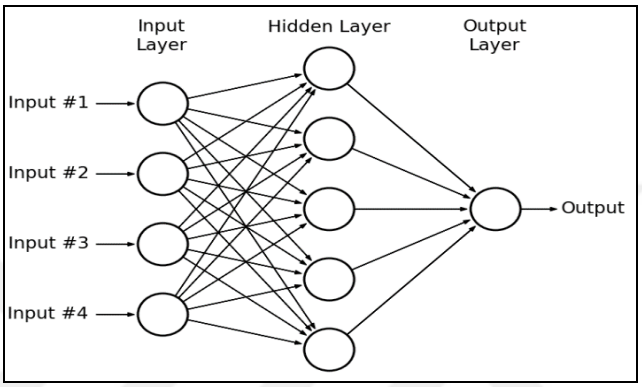


Figure 4.8. Simple multi-layer perceptron.

A multi-layer perceptron can be trained, with the error feedback (backpropagation). Here, the weights of the connections are changed so that the network can classify the desired patterns after a controlled training phase (supervised learning).

The common 3 activation functions in perceptron Figure 4.10 are:

I. ReLU (Rectified Linear Unit)
II. tanh function
III. Sigmoid function

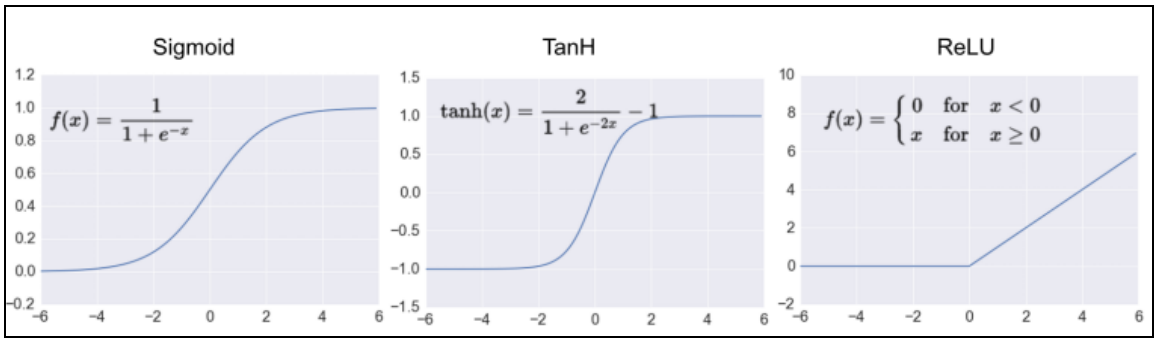


Figure 4.10. Three different activation functions for unites

4.5. Evaluate the Performance of Algorithms

Resampling methods, are the best statistical techniques to evaluate the performance of an algorithm. Where it permit to make accurate estimates of how the algorithm performs on new data (Brownlee, 2016).

The common types of resampling techniques are:

- Hold-out (Train and Test Split)
- Cross Validation (CV).
- Repeated Random Hold-out

4.5.1. Hold-out (Train and Test split)

The simplest way to evaluate the performance of algorithm is by use different sets of training and testing. In this technique original data split into two parts. the first part train the algorithm and make predictions on the second part then evaluate predictions against with the expected results. Generally, the size of the split data based on the size of dataset. The common to use 60% of the data for training and 40% for testing.

4.5.2. Cross validation (CV)

Cross Validation (CV), is a statistical method used to measure the accuracy of the model (which you programmed), and there are several types of it are:

4.5.2.1. K-fold cross validation

In the K-fold cross validation, the sample is randomly divided into equal sized subsamples, one part from the k subsamples is selected to for testing the model and the remaining part used to train the model k times, and it is repeated k times until cover all the parts, then the k results can be averaged to produce one estimation. 10-fold cross-validation is most commonly used in this method.

4.5.2.2. *2-fold* cross validation

2-fold cross validation is a special case of the previous type where k = 2, in this method we will use part of the data for training and the other for the test. where both

sets are equal size. In this validation, we train d0 and test d1, followed by training d1 and test d0.

4.5.3. Repeated Random Hold-out

In this method we seek to create a random split of the dataset as hold-out method but repeat the process of dividing (split) and evaluating the algorithm for several times, like K-fold cross-validation. The common use, splits the data into a 60% train, 40% test split and repeats the process 10 times.

4.6. Describe the Performance of Classifier

In order to display score of accuracy and describe the performance of a classifier, python provide two tools from *scikit learn* are:

- Confusion matrix
- Classification report.

4.6.1. Confusion matrix

Error matrix (or confusion matrix) is one way to describe the performance of a classifier. This square matrix (Figure 4.11) consists of rows and columns that list the number of instances as "actual" and "predicted" class.

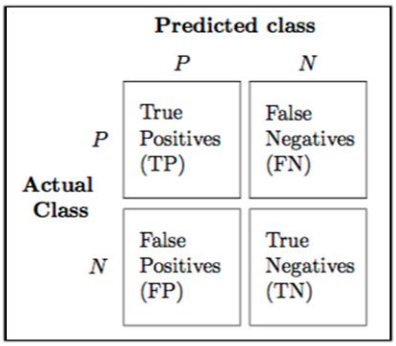


Figure 4.11. Confusion Matrix

TP = n. of true positives, FP = n. of false positives.

TN = n. of true negative, FN = n. of false negatives.

4.6.2. Classification report

The classification report displays the precision, recall, and F1 scores for the model. In order to support interpretation and problem detection.

Precision and Accuracy are general terms throughout machine learning and science. Precision is the ability of a classifier not to label an instance positive that is actually negative.

For each class it is defined as the rate of true positives to the sum of true and false positives. While Accuracy is how often is the classifier correct (Equation 4.14 and 4.15).

precision = TP / (FP+TP) ……...… (4.15)

Accuracy = (TN+TP)/ (TP+FP+ FN+TN) ……..... (4.16)

Recall, is the ability of a classifier to find all positive instances. For each class it is defined as the rate of true positives (TP) to the sum of true positives and FN (false negatives) Equation 4.16.

recall = TP / (FN + TP) ……..… (4.17)

F1 score, Is a measure of the accuracy of the test. It considers both recall and precision testing to calculate the result. F1 can be interpreted as a weighted average of the precision and recall, where F1 reaches the best value at 1 and the worst at 0 Equation 4.17.

F1 score = 2 * ((precision * recall) / (recall + precision) ……..… (4.18)

5. MATERIAL AND METHODS

Machine learning have become an essential methodology to transform the football statistics into useful information for helping teams, coaches in analyses opponents and make better decisions in real-time by using data generated by sensors. these data comprise videos from cameras to all types of physical measurements and human monitoring. But research in football analytics with Machine Learning techniques is limited and the main cause for that is the lack of a large-scale dataset for players because the collection such rich information about players might be very expensive. Therefore, may be only teams with high buying power can gather sensed data for their team. In addition, even if we went beyond the cost problem, sensed data is always susceptible to physical interference. Therefore, the use of sensed data for football analysis may be not be possible for worldwide large-scale research (Cotta ve ark., 2016).

In this thesis study, it is aimed to create a Decision Support System for a football team management by using machine learning techniques. The difficulty of obtaining large-scale reliable data, and cost problems related to this process were explained above. For these reasons, in the study, we aim to use the data of FIFA football video game which is widely used in the literature.

Recently, (Mathien, 2016) compiled, cleaned and shared a dataset of statistics of the European professional football. He used the EA Sports' FIFA video game series system for organizing as Excel database which includes a characterization of more than 10,000 players from the top football leagues in 11 European countries. This data allows finding insights about the footballers' performance onto a quantitative perspective. Further, these data were successfully used by (Soto-Valero, 2017) in their study.

In this thesis, we used the players' statistics of the Mathien's database. The total number of selected players for our analysis was for 17359 players for one season. This data includes 29 different skills of the players and these skills are rated on the scale of 0-100.

FIFA video game dataset has been used successfully to predict the results of football matches (Shin ve Gasparyan, 2014); (Prasetio, 2016), and have demonstrated to be comparable or better than other sources of football data (Shin ve Gasparyan, 2014).

After completing the data collection, we are classified each attributes of players (29 attribute) by following these steps:

I. Reviewed the literature (Ratomir ve ark., 2004), (Ostojić, 2000), (Raven ve ark., 1976), (McIntyre ve Hall, 2005), (Hughes ve ark., 2012).

II. We have consulted with a number of specialists in the Sports (coach, coach, player, teacher, referee etc.) In order to positioning players according to the most important skills required in each position, and the attributes of the players were classified as follows ‘‘Not important’’, ‘‘Not so important’’, ‘‘Normal’’, ‘‘Important’’ and ‘‘Very important’’. Where the attributes are evaluated according to the player's position to describe the effect of each attribute on players in each position. And then we represented the most important skills required in each position. Then, we have reduced the dimensions of the data and then we have performed the required operations for classification and regression problems by using machine learning techniques which are linear regression, logistic regression, random forest, neural network, k nearest neighbor.

III. Finally, we have evaluated all models produced using three techniques from statistical techniques (Hold-out “Train and Test Split”, Cross Validation (CV) and Repeated Random Hold-out) and comparison the result among them.

To implement our research, we have followed the following steps:

- First, assigning each player to the position, we calculated the average of the attribute strength "Rating" (Depending on the features of player).
- Second, selecting the required attributes for each player and assigning each player to the position we determined the best team squad according to formation plays such as (4-3-3), (3-5-2) based on rating of player.
- Third, after predict the player's preferred position and best team squad according to formation plays, we predicted dribbling skills of a players.

Below are defined the thesis requirement in terms of software and hardware to develop DSS.

- Operating System (Microsoft Windows 10 pro) integrated with laptop have following Specifications:

- Processor: Intel(R) Core (TM) i7-7500U CPU @ 2.70GHz 2.90 GHz
- Installed memory (RAM): 8.00 GB (7.80 GB usable)
- System type: 64-bit Operating System, x64-based processor

- python 3.6.5

Python provides highly efficient libraries like SciPy, NumPy, matplotlib, pandas, sklearn, yellowbrick. etc. All of these libraries allow you to easily develop the model instead of using complex coding.

- Server (Internet Information Service): It is used to run the Python environment
- Microsoft office 2010: For documentation and typing.

6. DATASET

In this section we will describe the data used in our research, the justification for its use, and the most important studies related with these data.

6.1. Dataset Collection

In the previous section, we explained the difficulty of obtaining large-scale reliable data, therefore we proposed the use of video game data from the EA FIFA series that we obtained from (Mathien, 2016).

6.2. Dataset Description

The presented dataset from FIFA video game depend on human's scouts (see Figure 6.1). Where, the company employs the scouts around the world to evaluate players' skill (like shooting, passing, ball control, dribbling, etc.) as realistically as possible.

Figure 6.1. Football scout (fieldoo, 2012)

The FIFA Soccer video game has over 500 licensed teams since the version of 2007. At the beginning of each season, an upgraded version of the game is offered to add new attributes to players. For instance, in FIFA Soccer 2007 there are 25 attributes, while in FIFA Football 2016 there are 29 attributes.

6.3. Dataset Analysis

6.3.1. Expletory Data Analysis

After loading the dataset and import main libraries of python like pandas and NumPy, we are going to take a look at and explore the data by a few different ways:

- Dimensions of Dataset
- Peek the Data
- Statistical Summary
- Missing Data

6.3.1.1 Dimensions of Dataset

We can get a quick idea of how many instances (rows and columns) in the dataset by use "*shape*" property. From Figure 6.2, we have 17359 instances and 29 attributes.

```
In [4]: #shape
        print(df.shape)

        (17359, 29)
```

Figure 6.2. Use shape command in python code

6.3.1.2. Peek at the Data

There is another important property data called "*head*". From Figure 6.3, We see the first 4 rows of the data.

```
In [6]: # head
        print(df.head(4))

                        Marking  Sliding tackle  Standing tackle  Interceptions  \
Name
Cristiano Ronaldo          22.0            23.0             31.0           29.0
L. Messi                   13.0            26.0             28.0           22.0
Neymar                     21.0            33.0             24.0           36.0
L. Suárez                  30.0            38.0             45.0           41.0

                        Finishing  Positioning  Volleys  Long shots  Aggression  \
Name
Cristiano Ronaldo            94.0         95.0     88.0        92.0        63.0
L. Messi                     95.0         93.0     85.0        88.0        48.0
Neymar                       89.0         90.0     83.0        77.0        56.0
L. Suárez                    94.0         92.0     88.0        86.0        78.0

                        Vision  ...  Ball control  Balance  Acceleration  \
Name                            ...
Cristiano Ronaldo         85.0  ...          93.0     63.0          89.0
L. Messi                  90.0  ...          95.0     95.0          92.0
Neymar                    80.0  ...          95.0     82.0          94.0
L. Suárez                 84.0  ...          91.0     60.0          88.0
```

Figure 6.3. Use head command in python code

6.3.1.3. Statistical Summary

We can take a look at a summary for each attribute by use "*describe*" property. This includes count, mean, the min and max values and percentiles. In Figure 6.4, note the all of the numerical values have score ranges between 0 and 100.

```
In [7]: # descriptions
        print(df.describe())
```

```
            Marking  Sliding tackle  Standing tackle  Interceptions  \
count  17359.000000    17359.000000     17359.000000   17359.000000
mean      44.024541       45.491445        47.339593      46.471283
std       21.609222       21.516340        21.873626      20.708161
min        4.000000        4.000000         4.000000       4.000000
25%       22.000000       23.000000        26.000000      25.000000
50%       48.000000       51.000000        54.000000      51.000000
75%       63.000000       64.000000        66.000000      64.000000
max       92.000000       91.000000        92.000000      92.000000

          Finishing   Positioning       Volleys    Long shots    Aggression  \
count  17359.000000  17359.000000  17359.000000  17359.000000  17359.000000
mean      45.106746     49.452561     43.062216     47.038539     55.674175
std       19.511232     19.496790     17.754378     19.315026     17.528163
min        2.000000      2.000000      4.000000      3.000000     11.000000
25%       29.000000     38.000000     30.000000     32.000000     43.000000
50%       48.000000     54.000000     44.000000     51.000000     58.000000
75%       61.000000     64.000000     57.000000     62.000000     69.000000
max       95.000000     95.000000     91.000000     92.000000     96.000000

             Vision      ...       Ball control       Balance  Acceleration  \
count  17359.000000      ...       17359.000000  17359.000000  17359.000000
mean      52.891872      ...          57.924304     63.721873     64.424276
std       14.412623      ...          16.904073     14.145971     14.966014
min       10.000000      ...           8.000000     11.000000     11.000000
25%       43.000000      ...          53.000000     55.000000     56.000000
50%       54.000000      ...          62.000000     65.000000     67.000000
75%       64.000000      ...          69.000000     74.000000     75.000000
max       94.000000      ...          95.000000     96.000000     96.000000

            Stamina  Sprint speed  Short passing     Reactions      Crossing  \
count  17359.000000  17359.000000   17359.000000  17359.000000  17359.000000
mean      63.061928     64.668356      58.155827     61.847226     49.622098
std       16.018156     14.682878      15.000103      9.181470     18.516431
min       12.000000     11.000000      10.000000     28.000000      5.000000
25%       56.000000     56.000000      53.000000     55.000000     37.000000
50%       66.000000     67.000000      62.000000     62.000000     54.000000
75%       74.000000     75.000000      68.000000     68.000000     64.000000
max       95.000000     96.000000      92.000000     96.000000     91.000000

       Long passing     Composure
count  17359.000000  17359.000000
mean      52.320756     57.776139
std       15.576183     12.975968
min        7.000000      5.000000
25%       42.000000     51.000000
50%       56.000000     60.000000
75%       64.000000     67.000000
max       93.000000     96.000000
```

Figure 6.4. Use describe command in python code

6.3.1.4. Missing Data

We can explore any missing values in data set (NaN value) and length of data by using "*info*" property. In Figure 6.5, It seems no missing value in dataset.

```
In [8]: df.info()

<class 'pandas.core.frame.DataFrame'>
Index: 17359 entries, Cristiano Ronaldo to L. Sackey
Data columns (total 29 columns):
Marking               17359 non-null float64
Sliding tackle        17359 non-null float64
Standing tackle       17359 non-null float64
Interceptions         17359 non-null float64
Finishing             17359 non-null float64
Positioning           17359 non-null float64
Volleys               17359 non-null float64
Long shots            17359 non-null float64
Aggression            17359 non-null float64
Vision                17359 non-null float64
Dribbling             17359 non-null float64
Penalties             17359 non-null float64
Strength              17359 non-null float64
Heading accuracy      17359 non-null float64
Shot power            17359 non-null float64
Curve                 17359 non-null float64
Agility               17359 non-null float64
Free kick accuracy    17359 non-null float64
Jumping               17359 non-null float64
Ball control          17359 non-null float64
Balance               17359 non-null float64
Acceleration          17359 non-null float64
Stamina               17359 non-null float64
Sprint speed          17359 non-null float64
Short passing         17359 non-null float64
Reactions             17359 non-null float64
Crossing              17359 non-null float64
Long passing          17359 non-null float64
Composure             17359 non-null float64
dtypes: float64(29)
memory usage: 4.0+ MB
```

Figure 6.5. Test missing data use *describe* command in python

6.3.2. Dataset Reduction Based on Filter Strategy

6.3.2.1. Heat map

As we know that correlation coefficient is main method for filter strategy to make feature selection. Heat map, consider a powerful method used to observe all correlation between features. Therefor we are going to use *seaborn* (python library) at first to discover the relationship between features that will be used to predict the skill of Dribbling (As in Figure 6.6)

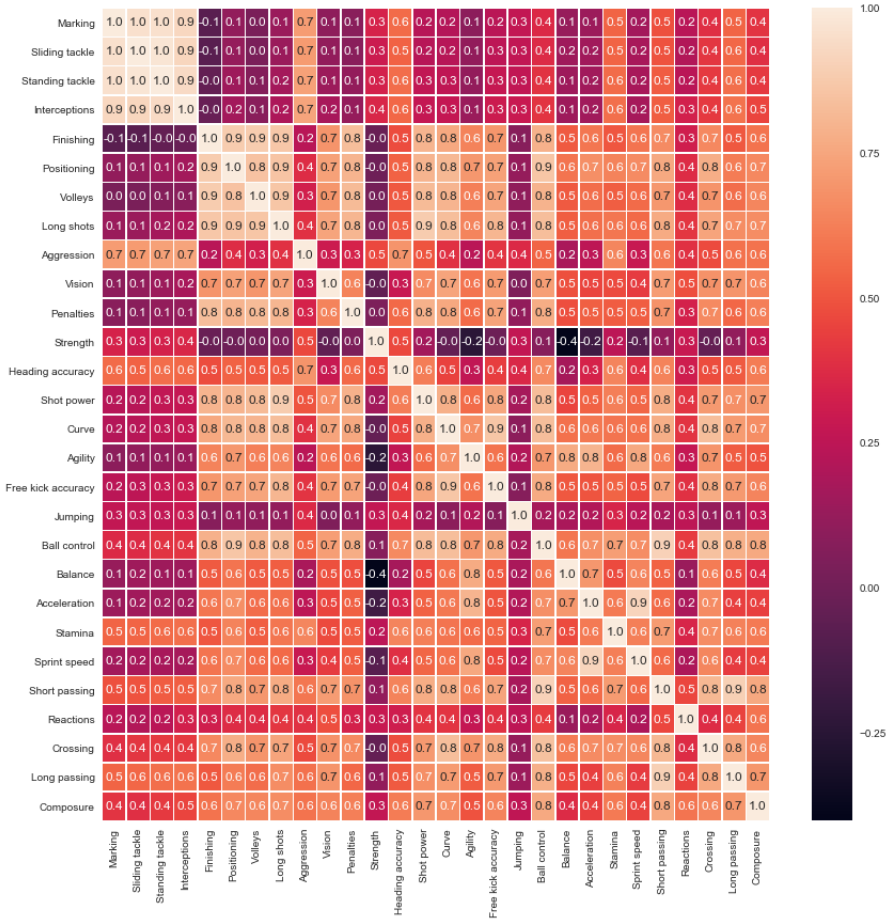


Figure 6.6. Heatmap (correlation matrix for 28 attributes)

We should look for a light patch in the heatmap because they represent a high correlation. For example, we have one for the (Long passing) and (Short passing) variables. This essentially means these two variables are highly correlated to each other, which means that we only need to keep one of them for modeling. Where, if we kept both, then our model would be prone to homoscedasticity.

6.3.2.2. Scatter plot

In order to compare two features or more in deeper shape, we use "*pair grid plot"*.

From heat map we found the correlation between features as below:

A- Sliding tackle, standing tackle, Interceptions and Marking are correlated.

Where, Pearson value is correlation value and 1 is the highest. Therefore, 0.9 is looks enough to say that they are correlated (Figure 6.7.).

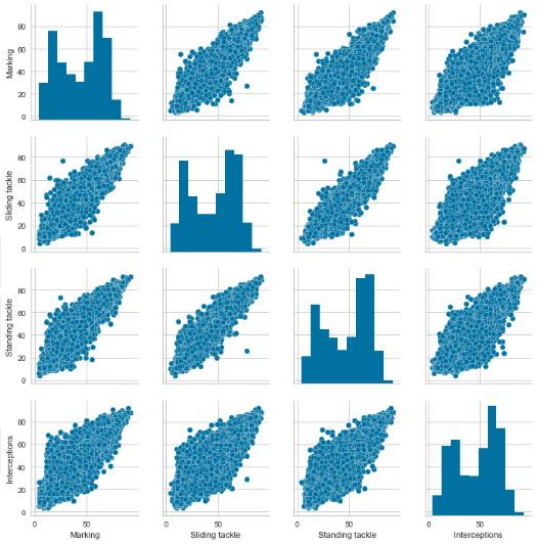


Figure 6.7. Scatter matrix for Sliding tackle, standing tackle, Interceptions and Marking

B- Positioning, Volleys, Long shots and Finishing are correlated (Figure 6.8).

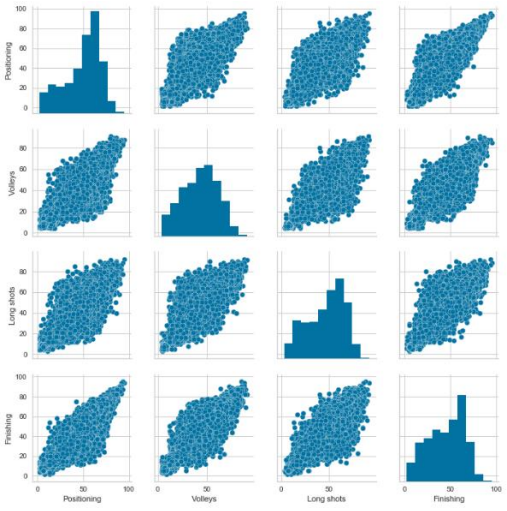


Figure 6.8. Scatter matrix for Positioning, Volleys, Long shots and Finishing

C- Long shots and Shot power are correlated (Figure 6.9).

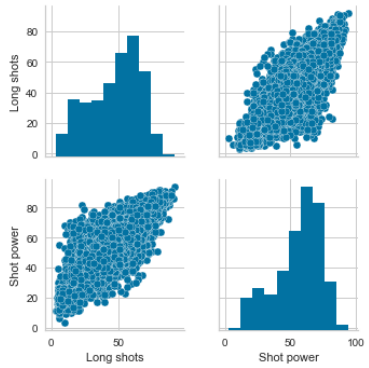


Figure 6.9. Scatter matrix for Long shots and Shot power

D- Curve and Free kick accuracy are correlated (Figure 6.10).

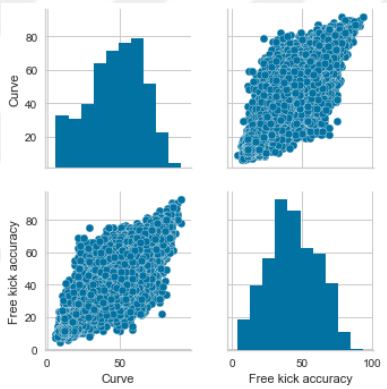


Figure 6.10. Scatter matrix for Curve and Free kick accuracy

E- Acceleration and Sprint speed are correlated (Figure 6.11)

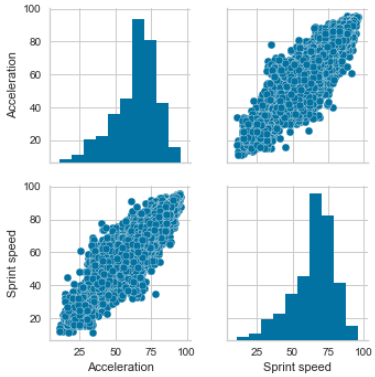


Figure 6.11. Scatter matrix for Acceleration and Sprint speed

F- Long passing, Short passing and Ball control is correlated (Figure 6.12).

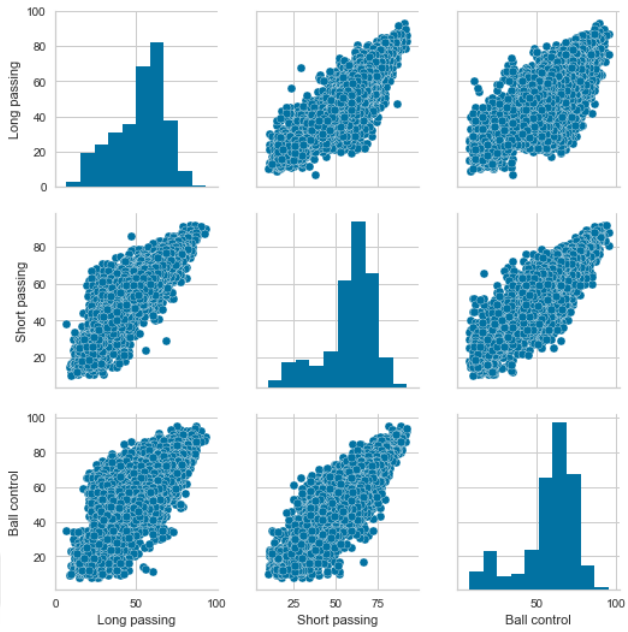


Figure 6.12. Scatter matrix for Long passing, Short passing and Ball control

6.3.2.3. Heat map and Scatter plot Analysis

As it seems in heat map and pair grid plot, some attributes are correlated with each other. Therefore, one attribute will be kept from related features and dropping the other attributes (As in Table 6.1).

Table 6.1. Dropped attributes according to correlated

n	Correlated attributes	Selected attribute	Dropped attributes
1	Sliding tackle, standing tackle, Interceptions and Marking	Marking	Sliding tackle, Standing tackle and Interceptions
2	Positioning, Volleys, Long shots and Finishing	Finishing	Positioning, Volleys and Long shots
3	Long shots and Shot power	Long shots	Shot power
4	Curve and Free kick accuracy	Curve	Free kick accuracy
5	Acceleration and Sprint speed	Acceleration	Sprint speed
6	Long passing, Short passing and Ball control	Ball control	Long passing and Short passing

After drop correlated features, as it can be seen in below correlation matrix, there are no more correlated features ‘values over 0.9’ Figure 6.13.

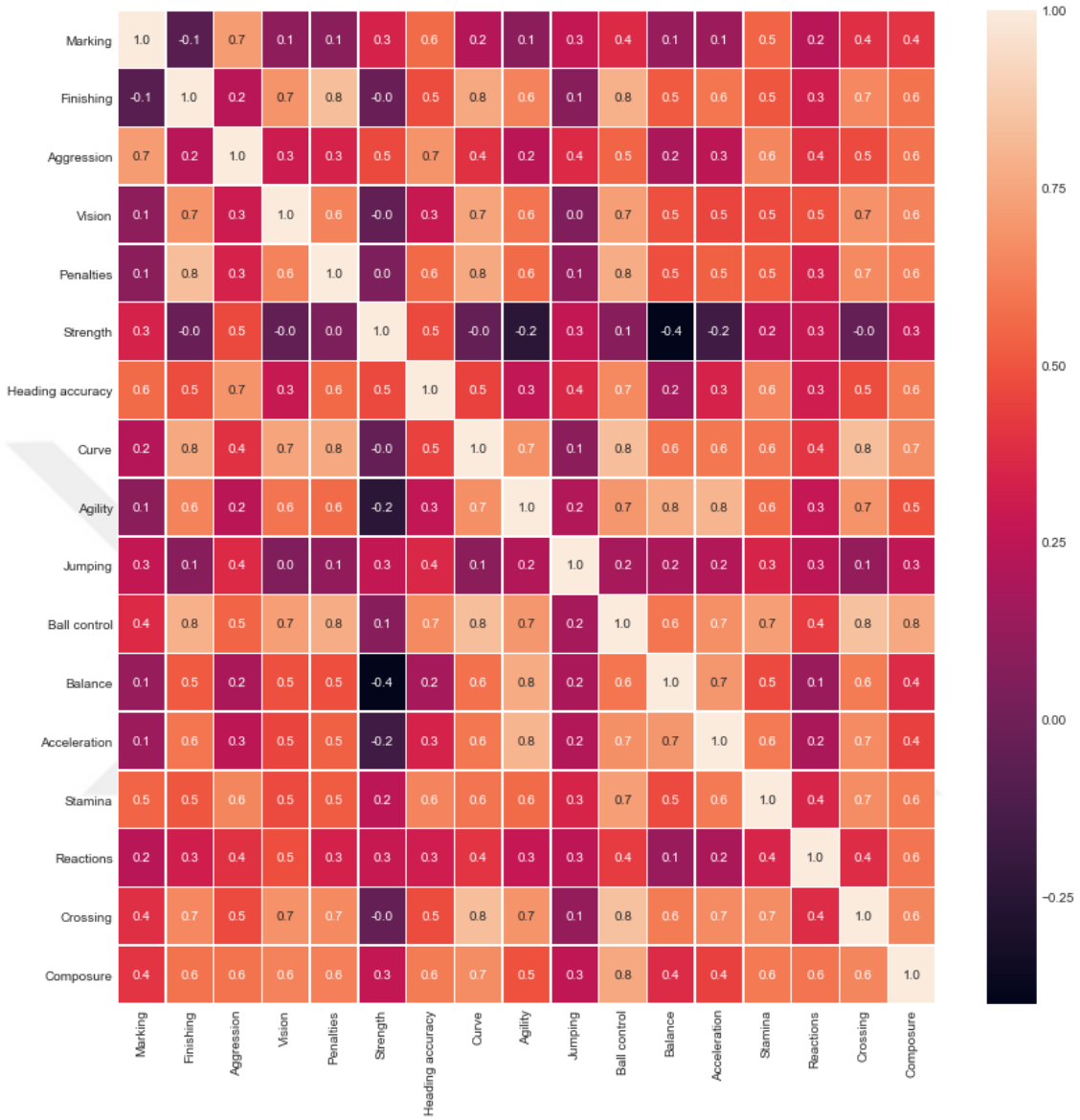


Figure 6.13. Heatmap (correlation matrix for 17 attributes)

6.3.2.4. Reduce Dimensionality through (PCA)

In the first section in our research, we seek to compute principal components of several variables in the football dataset. The football data set contains performance measures for video game data from the EA FIFA series during the 2017 season which are consist of 29 features (with dribbling skill's). We seek to reduce the number of this

features from 28 to 4 or 5 features by using PCA in order to predict Dribbling skill for player.

We will do that with following steps:

Step 1: Calculation of Explained Variance from the values

Step 2: Built PCA plot to find the reasonable number of components of PCA for exploratory data analysis on Variance (As Figure 6.14).

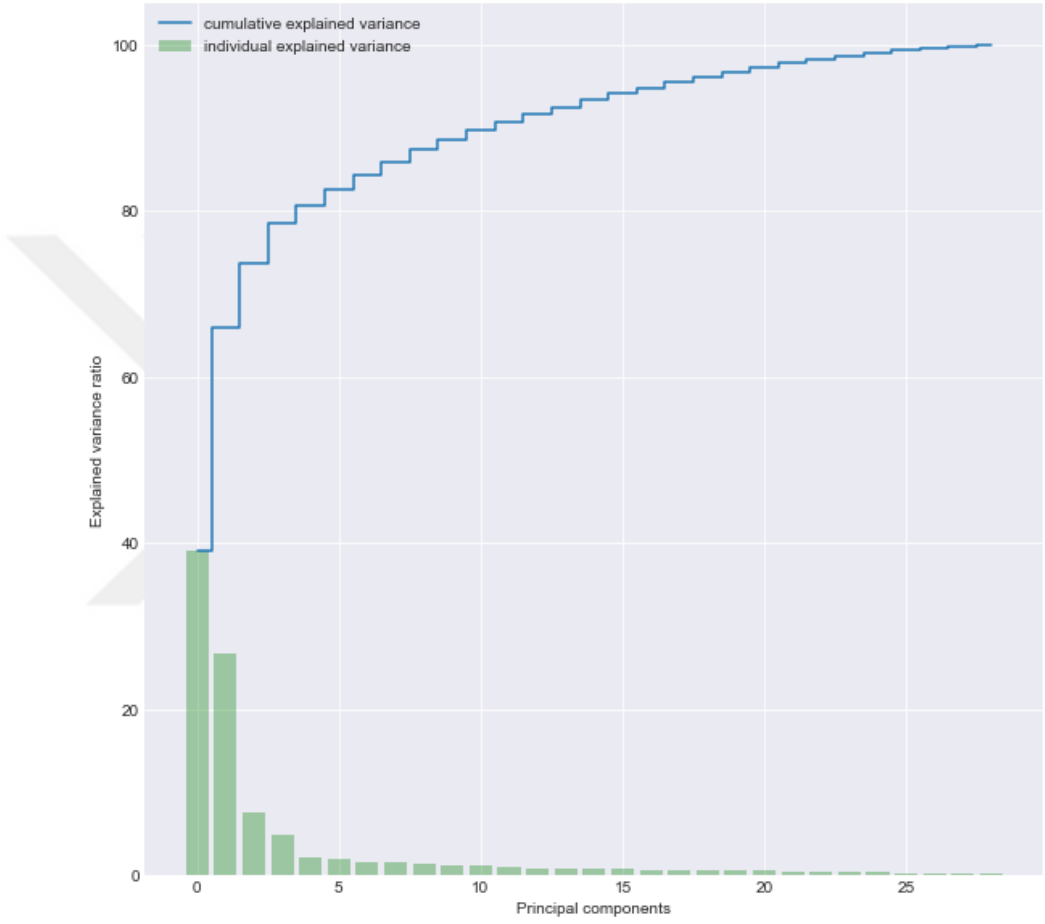


Figure 6.14. PCA plot

From Figure 6.14, there are 17 components (features) needed for 95%+ of variance level. Therefore, we can adopt the 17 features that resulted from correlation matrix. Where we will use these 17 features to build first prediction model to predict dribbling skill of players by using (linear regression, logistic regression, random forest and neural network).

7. PREDICT DRIBBLING SKILL

A dribbling is a game action, in a football sport using a ball. Dribbling used to measure ability of player to get around defenders, while avoiding that the player adversaries do not seize it. Dribbling skill, consider is one of the most important ways of talent identification. Where these skills are the basic technical skills of the player (Reilly ve Holmes, 1983) Especially Dribbling skill is considered critical to the outcome of the match (Huijgen ve ark., 2010) in addition to a previous study (Soto-Valero, 2017) indicated the most discriminating variable among player skills is dribbling. Therefore, predicting dribbling will help managers to make suitable decisions like sell, buy and contract renewal.

To predict dribbling skill, we will use two Strategies which are Filter Strategy and Wrapper Strategy.

7.1. Regression Based on Filter Strategy

In this section, we seek to build first decision support system model based on filter strategy to predict dribbling skill using *scikit-learn* in python. We will use player attributes that are selected according to n. PCA (17 features) which are:

1. 'Marking', (Mental)
2. 'Finishing', (Technical)
3. 'Aggression', (Mental)
4. 'Vision', (Mental)
5. 'Penalties', (Technical)
6. 'Strength', (Physical)
7. 'Heading accuracy', (Technical)
8. 'Curve', (Technical)
9. 'Agility', (Physical)
10. 'Jumping', (Physical)
11. 'Ball control', (Technical)
12. 'Balance', (Physical)
13. 'Acceleration', (Physical)
14. 'Stamina', (Physical)
15. 'Reactions', (Physical)

16. 'Crossing', (Technical)

17. 'Composure', (Mental)

The predictive model is built and updated by use four machine learning algorithms which are:

- Linear Regression
- Logistic Regression
- Random Forest
- Multilayer Perceptron Artificial Neural Network

These algorithms have proven to be effective in Regression of complicated data by statistics-based methodologies (resampling methods). Then, the predictive model with best performance would be selected and place into the IDSS.

Predictive models were created through these four learning algorithms. The performance of each predictive model was evaluated through a standard experience:

- Hold-out (Train and Test Split)
- K-fold cross validation
- Repeated Random Hold-out

In the Hold-out experiment, the original samples are randomly split into 60% for training and 40% for testing.

In the k-fold cross validation experiment, the original samples are randomly split into 10 subsets. One subset is kept as validation data to test the model, and the remaining 9 subsets are used as training data. This step is repeated 10 times. Finally, the average of 10 results from 10 subsets is calculated to produce a single performance estimate.

In the Repeated Random Hold-out experiment, the original samples are randomly split into 60% for training and 40% for testing (like Hold-out), then repeats the process 10 times (like cross validation).

7.1.1. Results

The performance of the four machine learning algorithms for this Strategy is evaluated and compared In Table 7.1, 7.2, and 7.3 by using three resampling methods are Hold-out, cross-validation and Repeated Random Hold-out.

Each machine learning algorithm was trained by the training set of 10415 instances (players) and evaluated its performance by the test set of 6944 instances (Mathien, 2016).

Table 7.1. Performance comparison among algorithms using Hold-out Based on Filter Strategy

Machine learning algorithms	**Accuracy**
Linear Regression	0.927
Logistic Regression	0.641
Multilayer Perceptron (MLP)	0.942
Random Forest	0.830

Table 7.2. Performance comparison among algorithms using K-fold cross-validation Based on Filter Strategy

Machine learning algorithms	**Accuracy**
Linear Regression	0.926
Logistic Regression	0.884
Multilayer Perceptron (MLP)	0.805
Random Forest	0.999

Table 7. 3. Performance comparison among algorithms using Repeated Random Based on Filter Strategy

Machine learning algorithms	**Accuracy**
Linear Regression	0.926
Logistic Regression	0.884
Multilayer Perceptron (MLP)	0.805
Random Forest	0.999

from above tables, Random Forest had the best performance in this group (As in Figure 7.1, 7.2 and 7.3), which have accurecy equal to (99.9%) in both K-fold cross-validation and Repeated Random Hold-Out Based on Filter Strategy.

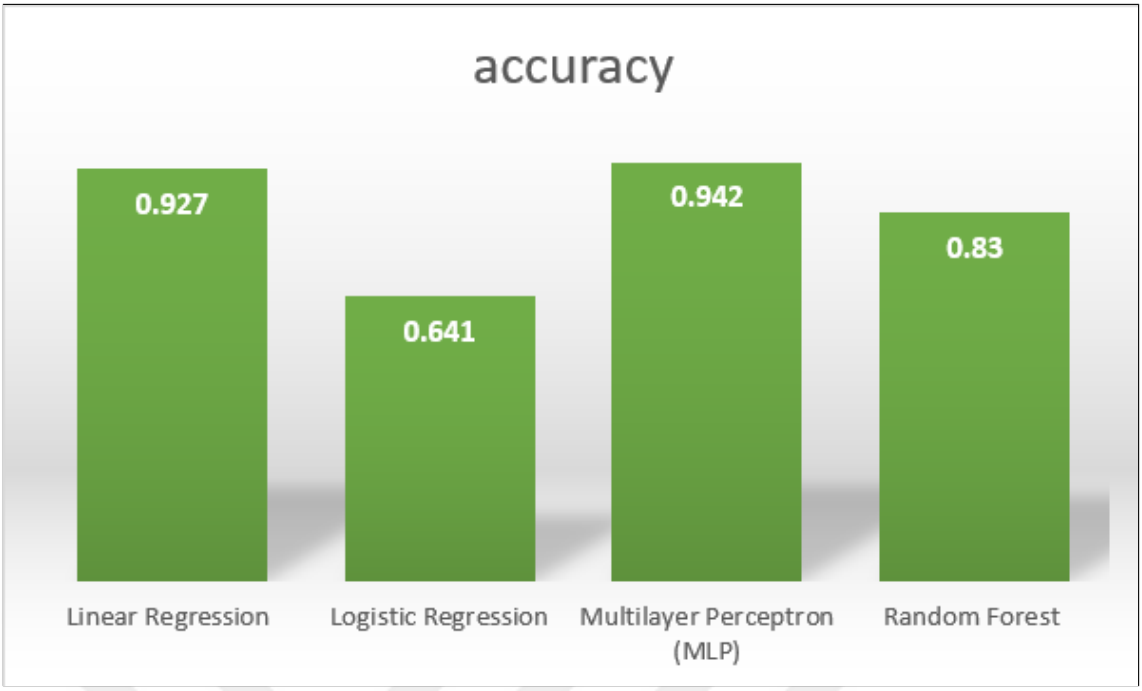


Figure 7.1. Summarize Performance comparison among algorithms using Hold-out Based on Filter Strategy

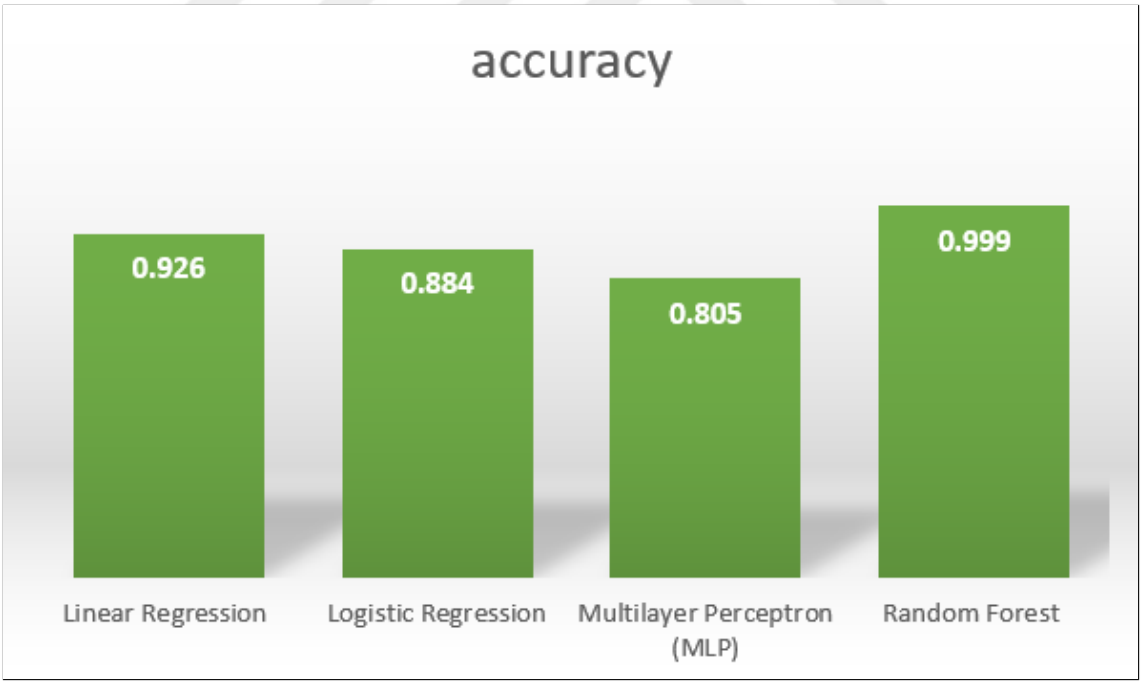


Figure 7.2. Summarize Performance comparison among algorithms using K-fold cross-validation Based on Filter Strategy

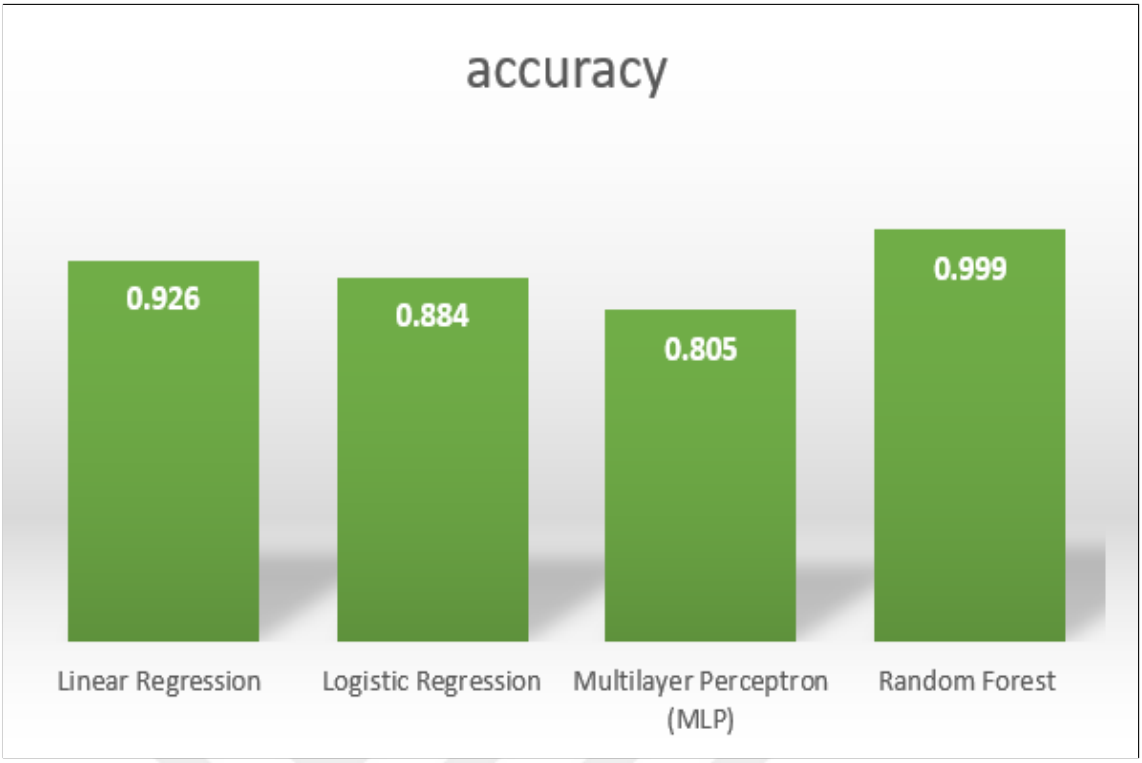


Figure 7.3. Summarize Performance comparison among algorithms using Repeated Random Based on Filter Strategy

Note that in the second experiment (predict dribbling skill using K-fold cross-validation Based on Filter Strategy) and third experiment (predict dribbling skill using Repeated Random hold-out Based on Filter Strategy), we don't have any different in results. Where, we have accurecy equal to (99.9%) in both K-fold cross-validation and Repeated Random Hold-Out Based on Filter Strategy. Therefore, we can adopt the results of any model from them. The predictive model trained by the 17 features has satisfactory performance in all models. The performance of the decision support system is expected to improve continuously if we add extra instances of players data into the decision support system.

7.2. Regression Based on Wrapper Strategy

From the previous section (Regression Based on Filter Strategy), we built four models to predict dribbling skill by using 17 features (selected based PCA), and we got the best results using linear regression. in this section we will use another technique (Regression Based on Wrapper Strategy) using 5 features to predict dribbling skill by using linear regression.

7.2.1. Recursive Feature Elimination algorithm (RFE)

As we know that Recursive Feature Elimination algorithm (RFE) is main method for wrapper strategy to done feature selection. Therefor we are going to use *sklearn* (python library) to find number of selected features by RFE algorithm that will be used to predict the skill of Dribbling.

According to an external estimator that allocates weights for features (for example, linear model coefficients), the goal of RFE is to identify features by repeating the consideration of smaller and smaller combinations of features.

First, the estimator is trained on the primary set of features and the importance of each feature is obtained either through a *(coef_ attribute)* or through a *(feature_importances_)* attribute. Then, the least important features (least coefficient) are exclude from current set of feature set. This procedure is repeated on the group until the desired number of features is reached at the end.

After executing (RFE) algorithm, we got the number of features required (5 features), which are (Finishing, Positioning, Ball control, Crossing and Acceleration).

7.2.2. All Possible Subset to Linear Model

After selected player attributes according to n. RFE (5 features), we created all possible subset to linear model and counting the number of mistakes made on that hold-out set, which are :

MODEL1 (Finishing, Positioning, Ball control, Acceleration, Crossing)

MODEL2 (Finishing)

MODEL3 (Finishing, Positioning)

MODEL4 (Finishing, Positioning, Ball control)

MODEL5 (Finishing, Positioning, Ball control, Acceleration)

MODEL6 (Positioning)

MODEL7 (Positioning, Ball control)

MODEL8 (Positioning, Ball control, Acceleration)

MODEL9 (Positioning, Ball control, Acceleration, Crossing)

MODEL10 (Ball control)

MODEL11 (Ball control, Acceleration)

MODEL12 (Ball contron, Acceleration, Crossing)

MODEL13 (Ball contron, Acceleration, Crossing, Finishing)

MODEL14 (Acceleration)

MODEL15 (Acceleration, Crossing)

MODEL16 (Acceleration, Crossing, Finishing)

MODEL17 (Acceleration, Crossing, Finishing, Positioning)

MODEL18 (Crossing)

MODEL19 (Crossing, Finishing)

MODEL20 (Crossing, Finishing, Positioning)

MODEL21 (Crossing, Finishing, Positioning, Ball control)

7.2.3. Results

Table 7.4, show all possible subset to linear model and their accurecy. Model 1, it has the best accuracy among all models (As in Figure 7.4).

Figure 7.5 show the flowchart of constructing the IDSS for predict dribbling skill.

Table 7.4: R^2 and Linear Equation for all Models

Model	**Estimation Equation (y)**	**Intercept**	**R^2**
1	-8.72 + 0.11*Finishing + 0.16*Positioning+ 0.5*Ball control+ 0.17*Acceleration+ 0.16*Crossing	-8.72	0.9207
2	18.65+ 0.80*Finishing	18.65	0.6777
3	11.53+ 0.13*Finishing+ 0.75*Positioning	11.53	0.8111
4	-3.31+ 0.07*Finishing+ 0.28*Positioning+ 0.70*Ball	-3.31	0.9002
5	-10.35+ 0.07*Finishing+ 0.21*Positioning+ 0.64*Ball control+ 0.21*Acceleration	-10.35	0.9188
6	11.64+ 0.87*Positioning	11.64	0.7995
7	-3.35+ 0.34*Positioning+ 0.71*Ball control	-3.35	0.9049
8	-10.22+ 0.27*Positioning + 0.65*Ball control + 0.21*Acceleration	-10.22	0.9193
9	-9.01+ 0.26*Positioning+ 0.55*Ball control + 0.18*Acceleration+ 0.14*Crossing	-9.01	0.9198
10	-6.16+1.05*Ball control	-6.16	0.8734
11	-14.28+0.88*Ball control + 0.27*Acceleration	-14.28	0.8932
12	-12.64+ 0.75*Ball control + 0.23*Acceleration + 0.17*Crossing	-12.64	0.9058
13	-9.84+ 0.56*Ball control + 0.19*Acceleration + 0.19*Crossing + 0.20*Finishing	-9.84	0.9192
14	-6.11+ 0.94*Acceleration	-6.11	0.5494
15	-4.24+ 0.65*Acceleration + 0.41*Crossing	-4.24	0.7866
16	-2.92+ 0.25*Acc. + 0.47*Crossing + 0.39*Finishing	-2.92	0.8652
17	-1.75 + 0.21*Acceleration + 0.37*Crossing + 0.20*Finishing + 0.30* Positioning	-1.75	0.8894
18	11.28+ 0.87*Crossing	11.28	0.7353
19	6.38+ 0.56*Crossing +0.45*Finishing	6.38	0.8535
20	5.65+ 0.43*Crossing + 0.20*Fishing +0.37Positioning	5.65	0.8733
21	-2.80+ 0.21* Crossing +0.11* Finishing + 0.20*Positioning + 0.54*Ball control	-2.80	0.9123

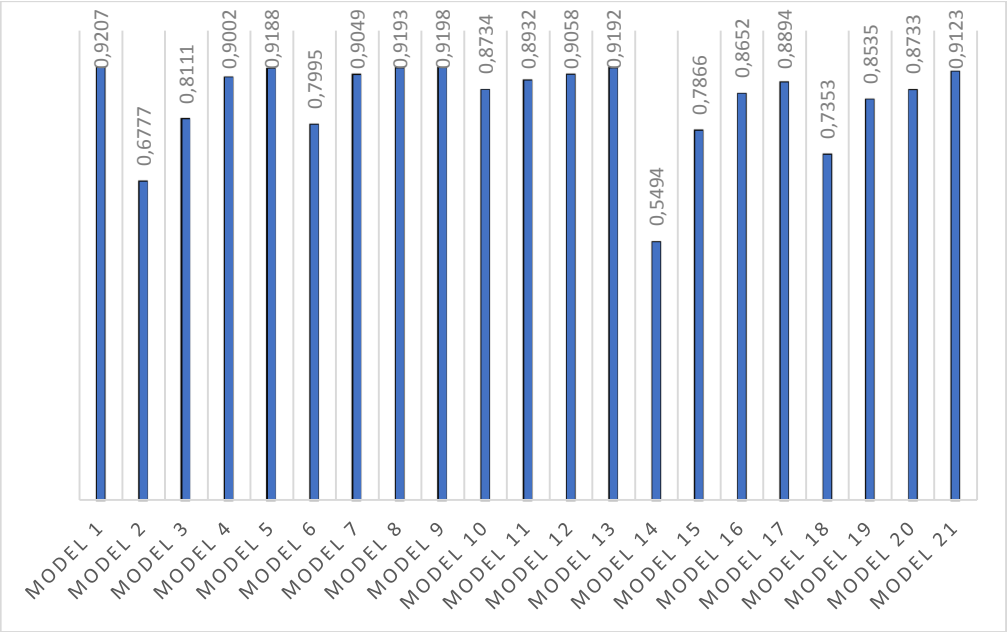


Figure 7.4. Summarize Performance comparison among all possible subset to linear models Based on Wrapper Strategy

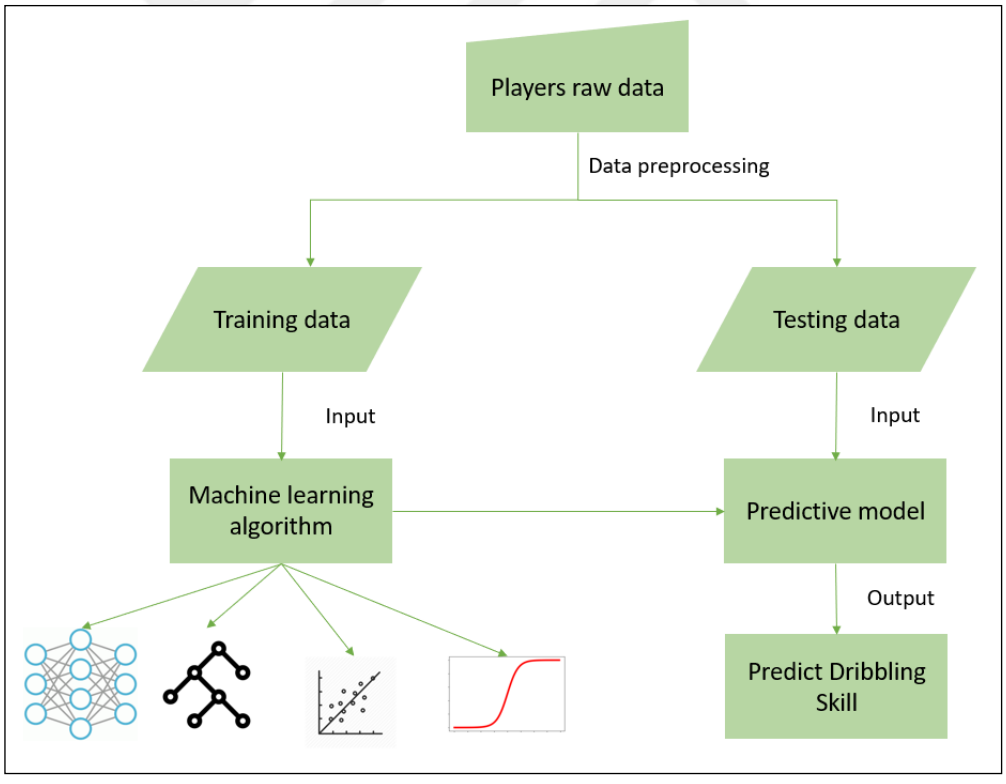


Figure 7.5. The flowchart of constructing the IDSS for predict dribbling skill

8. PREDICT PLAYER PREFERRED POSITION

In the sport of football, each of the 11 players on a team is assigned to a specific position on the play field. A team is consisting of one goalkeeper and ten players who fill various attacking, midfield and defensive positions depending on the formation of play.

8.1. Required Skills for Determining Player Position

In our research, we seek to assign players' positions based on their individual skills, which covers three parts (physical, mental, and technical skills). All these skills have been explained in Table 8.1 according to the following literature (Ratomir ve ark., 2004), (Ostojić, 2000), (Raven ve ark., 1976), (McIntyre ve Hall, 2005), (Hughes ve ark., 2012).

Table 8.1. Skills required among football players

Physical Skills	Mental Skills	Technical Skills
Acceleration	Aggression	Ball Control
Agility	Composure	Crossing
Balance	Interceptions	Curve
Jumping	Marking	Dribbling
Reactions	Positioning	Finishing
Sprint Speed	Vision	Free Kick
Stamina		Heading
Strength		Passing
		Penalties
		Short Passing
		Shot Power
		Sliding Tackle
		Standing Tackle
		Volleys

There are two steps to complete the research. Firstly, the different positions in football game should be identified. Then, the players should be classified based on the skill requirements for these positions.

8.2. Player positions in football team

In every football team, there are three main position (Defenders, Midfielders and Forwards), as well as Goalkeeper. These main three positions are divided into 14 position as follows (BUNDESLIGA, 2018):

Defenders

- Center back (CB)
- Right and Left Full backs (RB/LB)
- Right and Left-Wing backs (RWB/LWB)

Midfielders

- Center Midfielders (CM)
- Central Attacking Midfielders (CAM)
- Central Defensive Midfielders (CDM)
- Midfielders (RM/LM)
- Wingers (RW/LW)

Forwards

- Center Forward (CF)
- Striker (ST)

It is worth mentioning that goalkeeper is a special position which differs from other positions in some characteristics like "overhead exit" and "person to person battles". So, we ignore this position as a separate position.

In order to positioning players according to the most important skills required in each position we following these steps:

A) Reviewed the literature in (Ratomir ve ark., 2004), (Ostojić, 2000), (Raven ve ark., 1976), (McIntyre ve Hall, 2005), (Hughes ve ark., 2012).

B) We have consulted with a number of specialists in the Sports Science Faculty at the Selcuk university as well as consulted with a number of specialists in the Sports

(coach, coach, player, referee etc.), then attributes of the players were classified as follows:

‘‘Not important’’, ‘‘Not so important’’, ‘‘Normal’’, ‘‘Important’’ and ‘‘Very important’’. Where the attributes are evaluated according to the player's position (14 positions) to describe the effect of each attribute on players in each position (Figure 8.1).

C) Finally, we represented the most important skills required in each position after analysis data set by:

I. Find the mean of each skills in data set according to position of player (Figure 8.2)

II. Then, draw skills pattern based on attributes mean and Preferred Positions (Figure 8.3)

		attribute name		özellik adi	Pozisyon RW	(önem) - Not important (NI) - Not so important (NS) - Normal (N) - Important (I) - Very important (VI)
Attacking	1	Crossing	Hücum	Orta Açma		VI
	2	Finishing		Bitiricilik		N
	3	Heading accuracy		Kafa İsabeti		NS
	4	Short passing		Kısa Pas		I
	5	Volleys		Voleler		N
Skill	6	Dribbling	Yetenek	Dribling		VI
	7	Curve		Falso		I
	8	Free kick accuracy		S. Vuruş İsabeti		N
	9	Long passing		Uzun Paslar		VI
	10	Ball control		Top Kontrolü		I
Movement	11	Acceleration	Hareket	Hızlanma		VI
	12	Sprint speed		Sprint Hızı		VI
	13	Agility		Çeviklik		N
	14	Reactions		Reaksiyonlar		N
	15	Balance		Denge		N
Power	16	Shot power	Fizik	Şut Gücü		N
	17	Jumping		Zıplama		N
	18	Stamina		Dayanıklılık		N
	19	Strength		Güç		N
	20	Long shots		Uzaktan Şut		N
Mentality	21	Aggression	Mantalite	Saldırganlık		N
	22	Interceptions		Top Kesmeler		VI
	23	Positioning		Pozisyon Alma		N
	24	Vision		Oyun Görüşü		I
	25	Penalties		Penaltı		N
	26	Composure		Soğukkanlılık		N
Defending	27	Marking	Defans	Markaj		N
	28	Standing tackle		Ayakta Müdahale		N
	29	Sliding tackle		Kayarak Müdahale		N
Goalkeeping	30	GK diving	Kalecilik	KL Uçarak Kurtarış		NI
	31	GK handling		KL Elle Kontrol		NI
	23	GK kicking		KL Topa Vurma		NI
	33	GK positioning		KL Yer Tutma		NI
	34	GK reflexes		KL Refleks		NI

Figure 8.1: Classification of skills importance for (RW) position according to the opinion of specialists

```
In [44]: df_new_normalized_all.groupby('Preferred Positions').mean()
```

Out[44]:

Preferred Positions	Aggression	Crossing	Curve	Dribbling	Finishing	Free kick accuracy	Heading accuracy	Long shots	Penalties	Shot power	Volleys	Short passing	Long passing
0	52.590556	49.725070	53.119913	64.429015	66.445480	46.417521	61.904939	60.053122	62.712022	66.681889	58.608885	59.157192	47.05498
1	51.691221	61.456105	57.457114	68.271443	59.135217	49.956609	52.182644	57.272452	56.761857	63.942482	53.894046	63.245207	55.36024
2	52.080330	62.367662	59.478888	69.077240	59.217302	52.182286	51.733265	58.197734	56.940268	64.045314	54.269825	63.796087	55.67662
3	53.114083	62.660897	58.714792	68.431550	58.755311	52.316680	51.073958	58.406373	56.789929	63.844217	53.732887	64.623918	58.01534
4	63.055169	57.653664	56.737253	64.113402	51.931179	53.991084	53.975481	58.373920	54.570911	63.929507	50.064085	68.998050	65.54472
5	52.952904	63.386185	60.331240	68.886185	58.972527	54.181319	50.934458	59.206436	57.277080	64.432496	54.351256	64.959184	58.45879
6	53.218325	60.886015	61.779483	68.635248	59.989040	58.099518	51.095134	61.323542	59.300745	64.861464	56.177554	67.983341	62.67952
7	50.531429	58.460000	61.274286	69.737143	65.351429	55.805714	56.057143	63.188571	62.757143	67.460000	61.468571	65.225714	56.40285
8	68.351483	55.074352	52.937289	60.805858	47.227187	51.114532	58.117161	55.530229	52.768306	62.938415	46.399549	68.764551	65.39954
9	66.922314	42.771350	38.022865	45.327824	31.658127	36.223967	65.054545	36.357851	42.544353	50.901102	33.445730	57.217631	52.46005
10	63.916138	62.219405	51.950268	60.147245	39.685519	45.502682	56.479766	46.259873	46.601658	54.994637	39.795709	61.098489	55.78400
11	64.880333	60.339872	48.489456	58.956842	39.327612	41.508583	57.486513	44.074546	45.840118	53.645905	39.432565	61.087788	55.19078
12	63.111111	62.904762	52.338624	62.751323	44.481481	43.661376	55.301587	46.428571	48.269841	55.888889	41.825397	62.322751	56.50793
13	63.085106	65.452128	56.680851	63.643617	43.489362	49.595745	54.664894	49.351064	47.494681	57.021277	40.781915	62.686170	56.69148

Figure 8.2. Mean of each skill in data set according to position of player

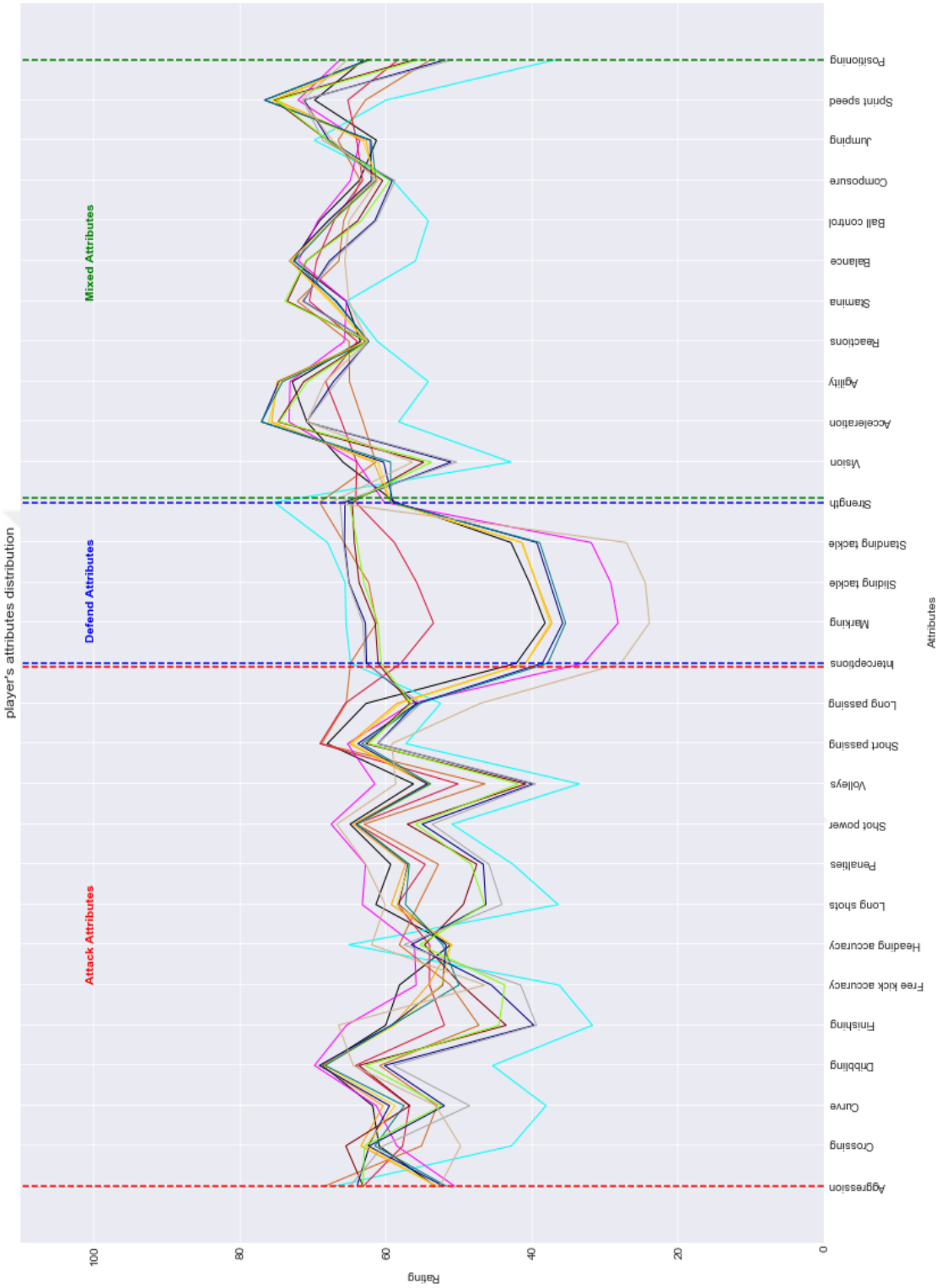


Figure 8.3. The most important skills required in each position

8.3. Principal component analysis (PCA)

In the previous section from our research (section six), we have computed PCA with 28 features. When we repeated the operation with 29 features (features required to classification player positions) we have same result (17 components are needed for 95%+ of variance level (see figure 6.16)).

8.4. Recursive Feature Elimination algorithm (RFE)

As we said in first section, the goal of (RFE) is to select features by repeated considering smaller and smaller sets of features according to an external estimator that assigns weights to features e.g., *(coef_ attribute)* of a linear model or *(feature_importances_)* of Random forest, this procedure is repeated on the group until the desired number of features is reached at the end (As Figure 8.4). The following chart shows features importance in random forest.

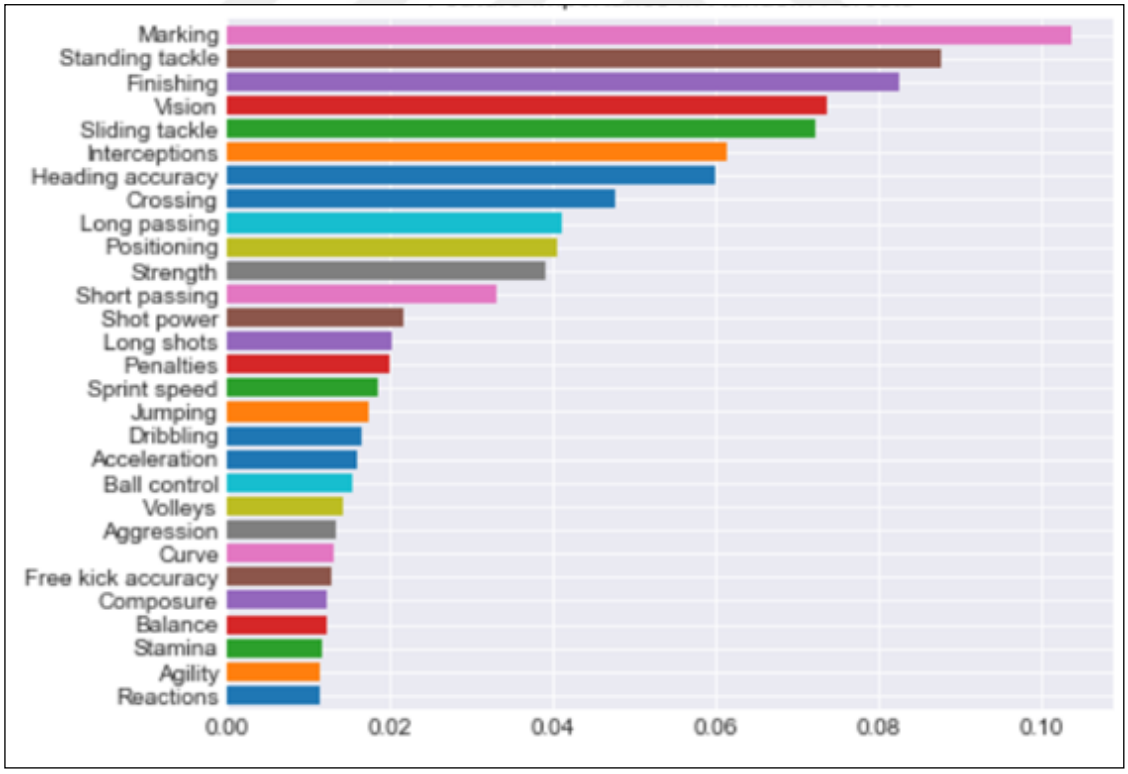


Figure 8.4. Features importance in random forest.

We will choose first 17 attributes to make prediction. Where, according to PCA, 17 components are needed for 95%+ of variance level. These attributes are:

(Note that these attributes differ from those used to predict dribbling skill)

1. 'Marking', (Mental)
2. 'Standing tackle', (Technical)
3. 'Finishing', (Technical)
4. 'Vision', (Mental)
5. 'Sliding tackle', (Technical)
6. 'Interceptions', (Mental)
7. 'Heading accuracy', (Technical)
8. 'Crossing', (Technical)
9. 'Long passing', (Technical)
10. 'Positioning', (Mental)
11. 'Strength', (Physical)
12. 'Short passing', (Technical)
13. 'Shot power', (Technical)
14. 'Long shots', (Technical)
15. 'Penalties', (Technical)
16. 'Sprint speed', (Physical)
17. 'Jumping', (Physical)

8.5. Classification algorithms

The core of IDSS is the predictive model (Figure 8.5). The predictive model is built and updated by machine learning algorithms for predict player preferred position. Study cases were classified into two groups:

- Group A: for binary classification (2 position: forward and defender)

- Group B: for multi classification (14 position: (CB), (RB), (LB), (RWB), (LWB), (CM), (CAM), (CDM), (RM), (LM), (RW), (LW), (CF), (ST)).

in this study we used supervised learning through four well known machine learning algorithms which are:

- Logistic Regression (LR)
- Random Forest (RF)
- Multilayer Perceptron Artificial Neural Network (MLP)
- K nearest neighbor (K-nn)

These algorithms have proven to be effective in classification of complicated data by statistics-based methodologies (resampling methods). Then, the predictive model with best performance would be selected and place into the IDSS.

Predictive models were created through these four learning algorithms. The performance of each forecasting model was evaluated through a standard experience:

- Hold-out (Train and Test Split)
- K-fold cross validation
- Repeated Random Hold-out

In the Hold-out experiment, the original samples are randomly split into 60% for training and 40% for testing.

In the k-fold cross validation experiment, the original samples are randomly split into 10 subsets. One subset is kept as validation data to test the form, and the remaining 9 subsets are used as training data. This step is repeated 10 times. Finally, the average of 10 results from 10 subsets is calculated to produce a single performance estimate.

In the Repeated Random Hold-out experiment, the original samples are randomly split into 60% for training and 40% for testing (like Hold-out), then repeats the process 10 times (like cross validation).

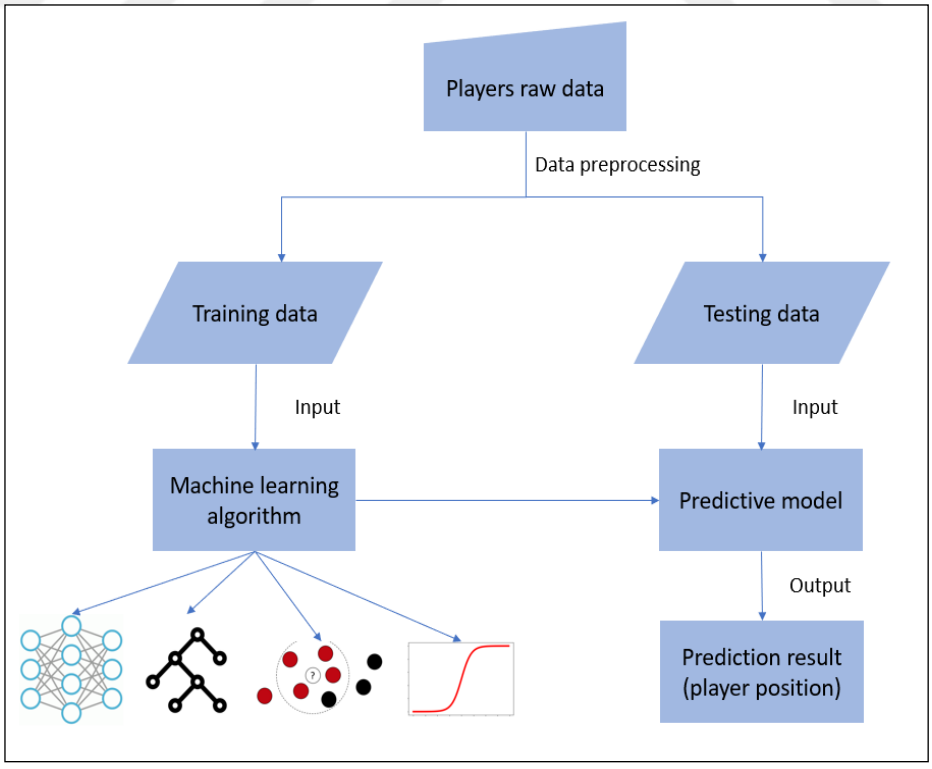


Figure 8.5. The flowchart of constructing the IDSS for predict player preferred position

8.6. Results

8.6.1. Result for binary classification

The performance of the four algorithms for this group is evaluated and compared In Table 8.2, 8.3 and 8.4 by using three resampling methods are Hold-out, cross-validation and Repeated Random Hold-out.

Each algorithm was trained by the training set of 16350 instances (players) and evaluated its performance by the test set of 10901 instances.

Table 8.2. Performance comparison among algorithms using Hold-out for Group A

Machine learning algorithms	accuracy	precision	recall	f1-score
Logistic Regression (LR)	0.856	0.856	0.857	0.856
Random Forest (RF)	0.801	0.801	0.802	0.801
Multilayer Perceptron (MLP)	0.859	0.859	0.859	0.859
K-nearest neighbor (K-nn)	0.822	0.821	0.822	0.822

Table 8.3. Performance comparison among algorithms using K-fold cross-validation for Group A

Machine learning algorithms	accuracy	precision	recall	f1-score
Logistic Regression (LR)	0.863	0.864	0.863	0.863
Random Forest (RF)	0.886	0.801	0.802	0.801
Multilayer Perceptron (MLP)	0.865	0.866	0.865	0.866
K-nearest neighbor (K-nn)	0.882	0.882	0.882	0.882

Table 8.4. Performance comparison among algorithms by using Repeated Random Hold-out for Group A

Machine learning algorithms	accuracy	precision	recall	f1-score
Logistic Regression (LR)	0.863	0.864	0.863	0.863
Random Forest (RF)	0.886	0.886	0.886	0.886
Multilayer Perceptron (MLP)	0.865	0.866	0.865	0.866
K-nearest neighbor (K-nn)	0.882	0.882	0.882	0.882

from above tables, Random Forest had the best performance in this group (In Figure 8.6, 8.7 and 8.8) which have accurecy equal to (88.6%) in both K-fold cross-validation and Repeated Random Hold-out.

Figure 8.6. Summarize the accuracy results of classification algorithms Performance comparison among algorithms by using Hold-out for Group A

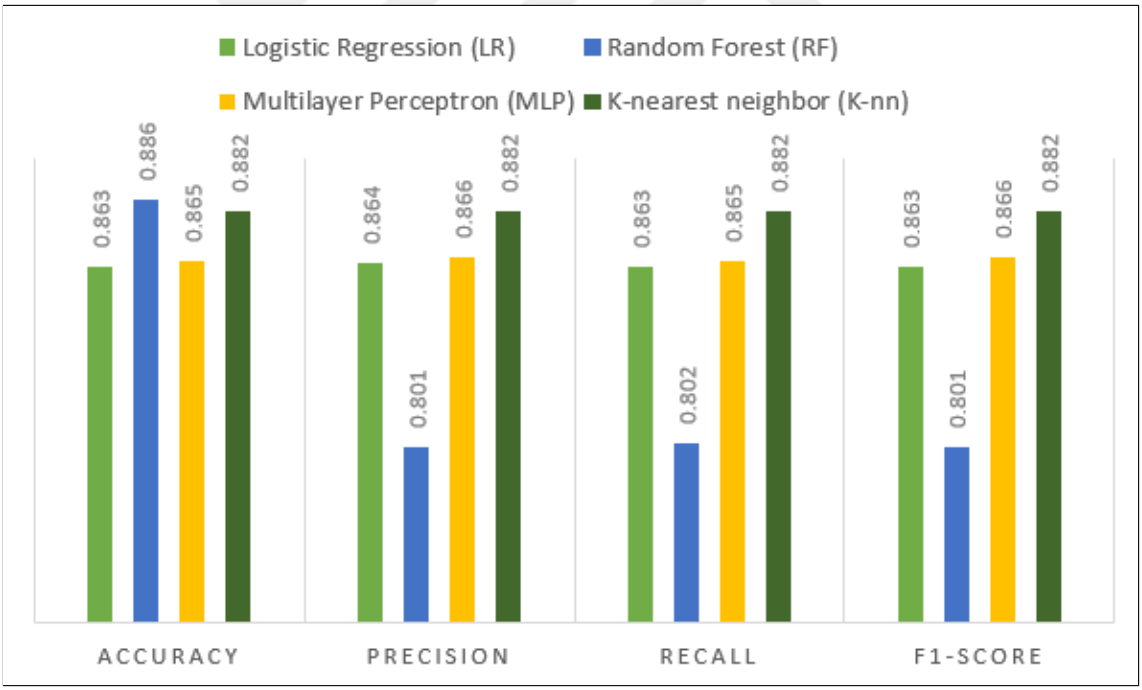


Figure 8.7. Summarize the accuracy results of classification algorithms Performance comparison among algorithms by using K-fold cross-validation for Group A

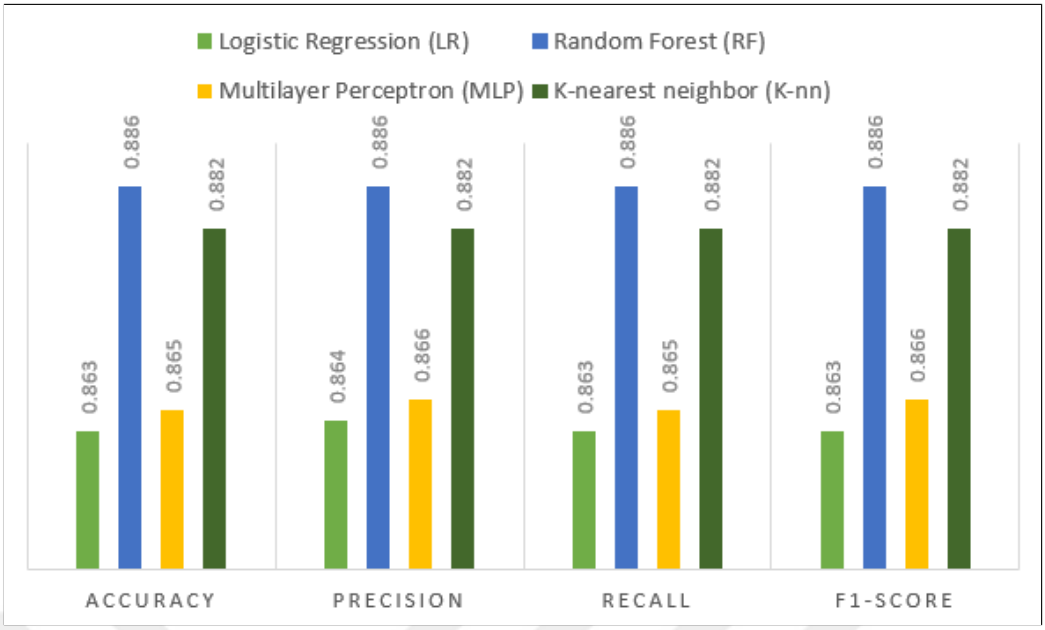


Figure 8.8. Summarize the accuracy results of classification algorithms Performance comparison among machine by using Repeated Random Hold-out for Group A

8.6.2. Result for multi classification

The performance of the four algorithms for this group is evaluated and compared In Table 8.5, 8.6 and 8.7 by using three resampling methods are Hold-out, cross-validation and Repeated Random Hold-out.

Each algorithm was trained by the training set of 16350 instances (players) and evaluated its performance by the test set of 10901 instances.

Table 8.5. Performance comparison among algorithms using Hold-out for Group B

Machine learning algorithms	accuracy	precision	recall	f1-score
Logistic Regression (LR)	0.360	0.263	0.360	0.231
Random Forest (RF)	0.277	0.266	0.277	0.271
Multilayer Perceptron (MLP)	0.434	0.364	0.423	0.373
K-nearest neighbor (K-nn)	0.300	0.279	0.300	0.280

Table 8.6. Performance comparison among algorithms by using K-fold cross-validation for Group B

Machine learning algorithms	accuracy	precision	recall	f1-score
Logistic Regression (LR)	0.377	0.319	0.377	0.266
Random Forest (RF)	0.585	0.581	0.585	0.582
Multilayer Perceptron (MLP)	0.443	0.373	0.428	0.373
K-nearest neighbor (K-nn)	0.514	0.533	0.514	0.487

Table 8.7. Performance comparison among algorithms by using Repeated Random Hold-out for Group B

Machine learning algorithms	accuracy	precision	recall	f1-score
Logistic Regression (LR)	0.377	0.319	0.377	0.266
Random Forest (RF)	0.584	0.581	0.585	0.583
Multilayer Perceptron (MLP)	0.443	0.373	0.428	0.373
K-nearest neighbor (K-nn)	0.514	0.533	0.514	0.487

from above tables, Random Forest had the best performance in this group (In Figure 8.9, 8.10 and 8.11) which have accurecy equal to (58.5%) in K-fold cross-validation.

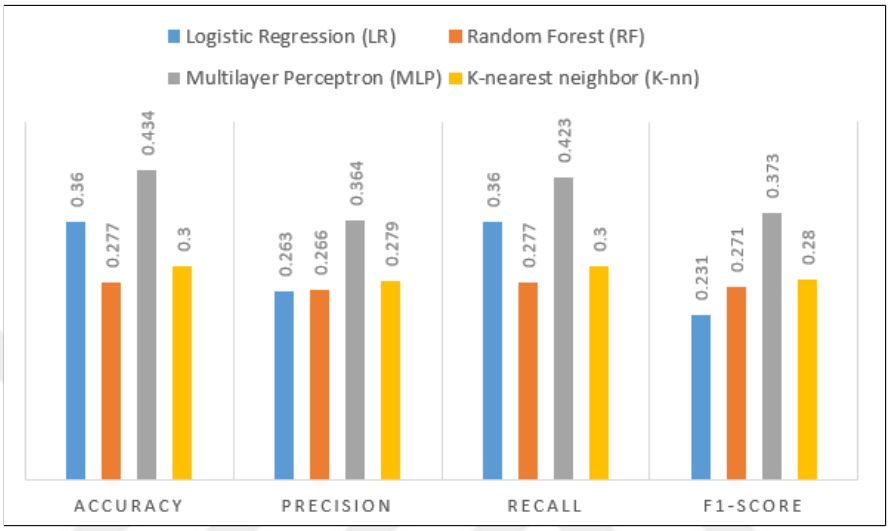


Figure 8.9. Summarize the accuracy results of classification algorithms Performance comparison among algorithms by using Hold-out for Group B

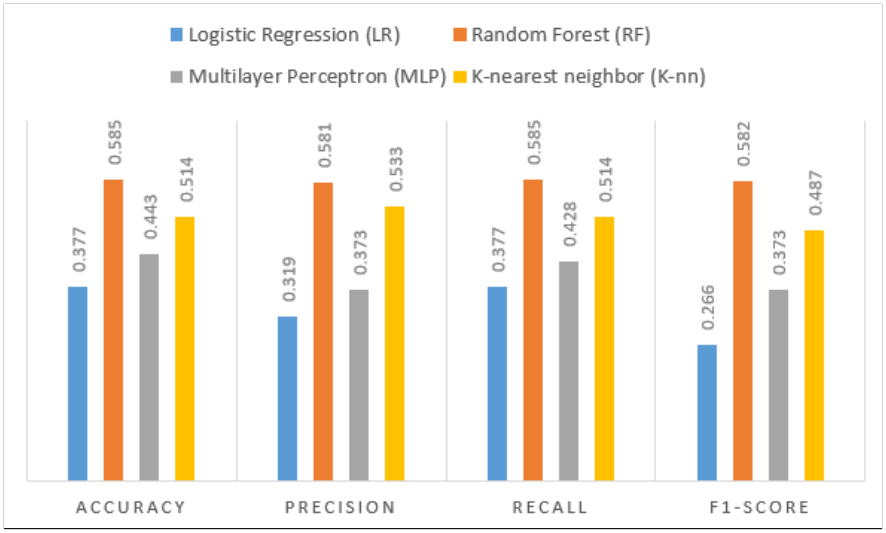


Figure 8.10. Summarize the accuracy results of classification algorithms Performance comparison among algorithms by using K-fold cross-validation for Group B

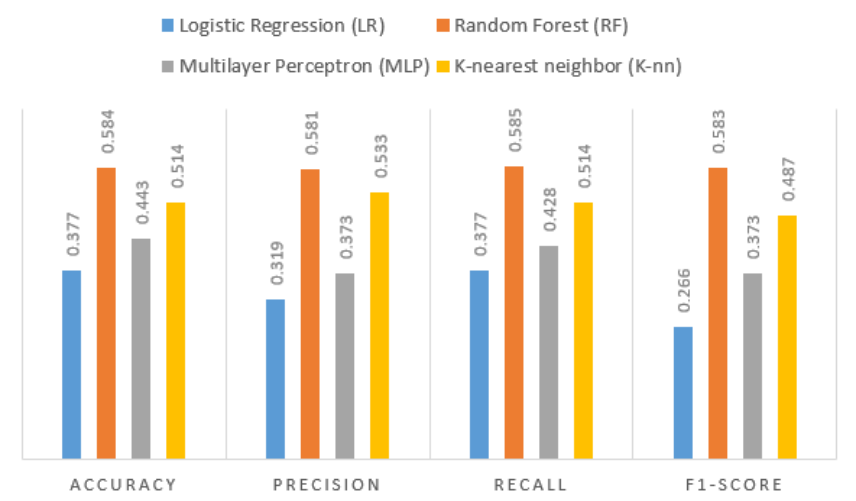


Figure 8.11. Summarize the accuracy results of classification algorithms Performance comparison among algorithms by using Repeated Random Hold-out for Group B

We note that the accuracy of the group (B) is less than from group (A), and the reason can be observed from classification report. For example, in random forest classification report (Figure 8.12), the classifier of position (12) have precision equal to (29.3%) with 189 instances. Otherwise, the classifier of position (9) have precision equal to (78.1%) with 3630 instances. This indicates to that the decrease in the number of instances for a classifier lead to decrease the accuracy.

```
print ('Random forest clasification report\n')
print(classification_report(y, y_pred3, digits=3))
```

```
Random forest clasification report

             precision    recall  f1-score   support

          0      0.744     0.769     0.757      3219
          1      0.443     0.423     0.433       991
          2      0.431     0.408     0.419       971
          3      0.485     0.463     0.474      2542
          4      0.551     0.579     0.565      3589
          5      0.468     0.458     0.463      2548
          6      0.484     0.493     0.488      2281
          7      0.426     0.400     0.412       350
          8      0.544     0.519     0.531      2663
          9      0.781     0.802     0.791      3630
         10      0.641     0.655     0.648      2051
         11      0.613     0.593     0.603      2039
         12      0.293     0.259     0.275       189
         13      0.310     0.335     0.322       188

avg / total      0.582     0.585     0.583     27251
```

Figure 8.12. Random forest classification report by using K-fold cross-validation for Group B

9. FIND THE BEST AVAILABLE SQUAD ACCORDING TO FORMATIONS OF PLAY

In football, the formation of play describes how the players in a team generally position themselves on the pitch. the position of player in a formation defines by coach. The coach decides whether a player has a mostly defensive or attacking role, and whether they tend to play towards one side of the pitch or centrally.

Formations are typically described by three numbers, which refer to how many players are in each row of the formation from the most defensive to the most forward. For example, "4-3-3" formation has four defenders, three midfielders, and three forward (Figure 9.1). Different formations can be used depending on whether a team wishes to play more attacking or defensive football.

The choice of formation is made by a coach. Where the coach seeks to select the best available players for each formation.

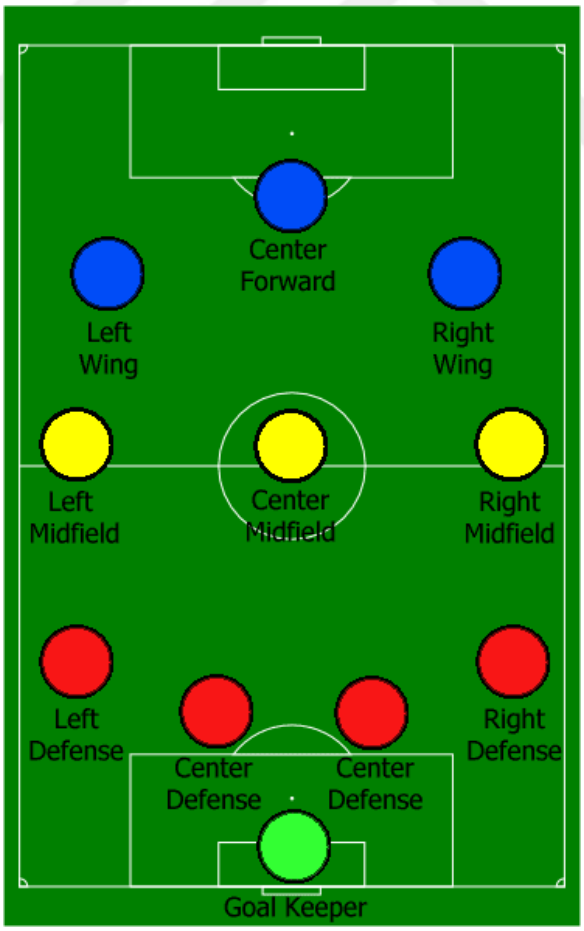


Figure 9.1. The 4–2–4 formation

9.1. Procedures to find best available squad according to formations of play

After predict preferred position for each player in team using machine learning techniques, we got best result using random forest. in this section we seek to find the best available squad according to formations of play such as 4-4-2, 3-5-2 based on position of player which is specified by random forest with the following steps:

1- Select the player's preferred position using random forest.

2- Calculate the rating of player (overall) from this equation:

Overall = Total values of player attributes / Number of attributes.

3- Choose players who have the highest rating per selected position from highest to lowest sequentially (As in figure 9.2).

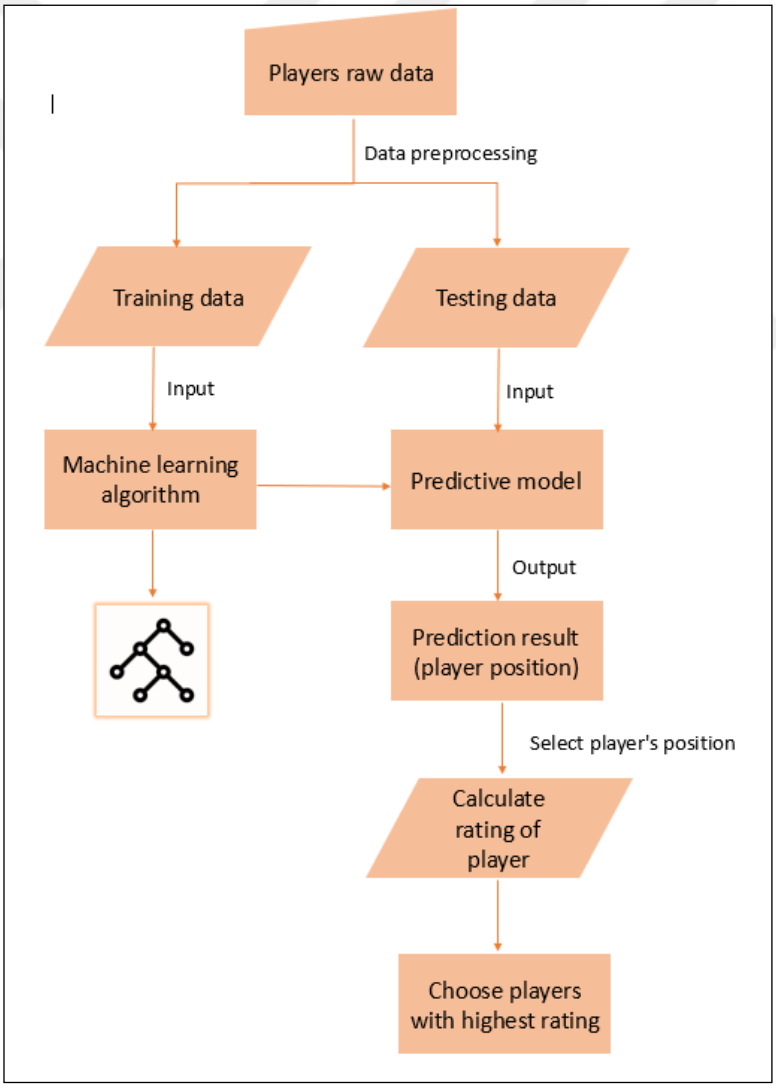


Figure 9.2. The flowchart of constructing the IDSS for find the best available squad according to formations of play

9.2. Result

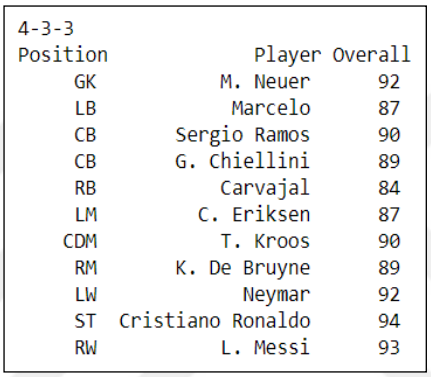
After finding a rating (overall) of players for the dataset which are consist of 17359 players, we will choose players who have the highest rating per selected position from highest to lowest sequentially.

We done the code that performs this process, and we implemented the test in two groups. The results are as follows:

Groupe 1:

We will select best available squad according to this formation (Figure 9.3):

Squad_433 = [GK, LB, CB, CB, RB, LM, CDM, RM, LW, ST, RW]



4-3-3

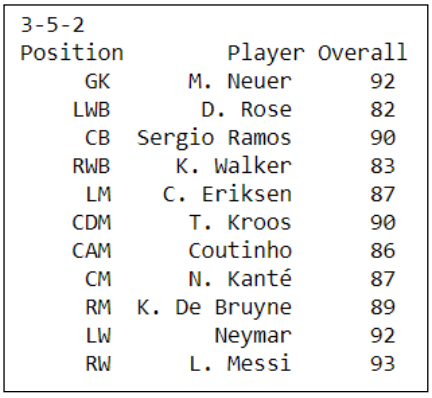
Position	Player	Overall
GK	M. Neuer	92
LB	Marcelo	87
CB	Sergio Ramos	90
CB	G. Chiellini	89
RB	Carvajal	84
LM	C. Eriksen	87
CDM	T. Kroos	90
RM	K. De Bruyne	89
LW	Neymar	92
ST	Cristiano Ronaldo	94
RW	L. Messi	93

Figure 9.3. Result of group 1 to find best available squad

Groupe 2:

We will select best available squad according to this formation (Figure 9.4):

squad_352 = [GK, LWB, CB, RWB, LM, CDM, CAM, CM, RM, LW, RW]:



3-5-2

Position	Player	Overall
GK	M. Neuer	92
LWB	D. Rose	82
CB	Sergio Ramos	90
RWB	K. Walker	83
LM	C. Eriksen	87
CDM	T. Kroos	90
CAM	Coutinho	86
CM	N. Kanté	87
RM	K. De Bruyne	89
LW	Neymar	92
RW	L. Messi	93

Figure 9.4. Result of group 2 to find best available squad

10. RECOMMENDATION AND CONCLUSION

In management a football team, the coach selects the players in addition to formation of play based on his personal experience. There is no formula or scientific equations for the coach to evaluate and compare between diff erent players. where, the assignment generally is done by the coaches by use their experiences and observations about the players, these making selecting of players subject to many biases.

This thesis is set to introduce a new intelligent decision support system (IDSS) for a football team management based on individual skills of players (technical, physical and mental). Players skills are used in this system to find preferred available position for player in team and find the best available squad according to formations of play, which will promote consistency in decision-making through elimination of personal bias in football team selection. Further, the system has ability to predict dribbling skill for each player in the team to monitor the growth and performance of players because predicting player's skill (like dribbling) will help managers to make suitable decisions like sell, buy and contract renewal.

The experiments of classification (that be used to find player position) achieved the highest predictive accuracy of 88.6% for binary classification (2 position) and predictive accuracy of 58.5% for multi classification (14 position) by using random forest. The reason for the low accuracy is the lack of sufficient instances of position classification in the dataset which is can be observed from classification report. therefore, we suggest increasing the number of instances in the dataset to increase accuracy. As well as, the experiments of regression achieved the highest predictive accuracy of 99.9% to predict dribbling skill by using random forest.

From the previous experiences we note that random forest has the highest accuracy, and it has proved to be more efficient for classification and regression than other algorithms.

The study has shown that machine learning has a great importance in the football sport, through their role to transform the football statistics into useful information for helping teams, coaches and athletes in analyses opponents and make better decisions in real-time. Further, the techniques of machine learning used in this field have varied.

The most important conclusions for thesis:

a. Differently from the previous studies (Table 10.1) in this thesis we use random forest algorithm to find preferred available position for each player in team, and it has proved to be more efficient in classification of players position's than other algorithms. Further, in this thesis and differently from the previous studies we analyzed the players' attributes using machine learning techniques before it uses in predict best position of player in team. As well as, we predicted further player positions (14 positions)

Table 10.1. Compare the results of previous research with the current research

Study name	Number of prediction locations	Algorithm name	Accuracy
Bazmara et al., (2013)	3	K- nearest neighbor	88%
Tavana et al., (2013)	3	Fuzzy logic	60%
Abidin et al., (2016)	Conceptual frame work (theory study)		
Razali et al., (2017)	10	Bayesian Networks	99%
		Decision Tree	98%
		Nearest Neighbor	97%

b. Our study has focused on use machine learning for team management, so we advise developers to search on other topics such as predict football players injuries or predict values or wages of players.

c. Our study has shown that a data collected from video games (FIFA 2017) could improve prediction quality, so we advise developers to use these games as a data source. In our study, data analysis is integrated with the decision support system and this consider as one of the rare examples in computer science. Where, this thesis has provided trustworthy decision support systems that can be used as a basis for developing an intelligent decision support system in other sports like Rugby or American football (gridiron).

d. The results of the classification experiments show the need to add additional data for player position for proper training of DSS before it uses in classification.

e. The results of the classification experiments also show the possibility of using another approach and algorithms like deep learning, fuzzy logic, support vector machine, etc.

11. REFERENCES

Abidin, M. Z. Z., Nawawi, M. K. M. ve Kasim, M. M., 2016a, Conceptual Framework of Decision Support System for Team Sports, *Journal of Engineering and Applied Sciences*, 100 (8), 1788-1791.

Abidin, M. Z. Z., Nawawi, M. K. M. ve Kasim, M. M., 2016b, Research design of decision support system for team sport, *AIP Conference Proceedings*, 040001.

Alpaydin, E., 2014, Introduction to machine learning, MIT press, p.

Altman, N. S., 1992, An introduction to kernel and nearest-neighbor nonparametric regression, *The American Statistician*, 46 (3), 175-185.

Arabzad, S. M., Tayebi Araghi, M., Sadi-Nezhad, S. ve Ghofrani, N., 2014, Football match results prediction using artificial neural networks; the case of Iran Pro League, *Journal of Applied Research on Industrial Engineering*, 1 (3), 159-179.

Asif, R., Zaheer, M. T., Haque, S. I. ve Hasan, M. A., 2016, Football (Soccer) Analytics: A Case Study on the Availability and Limitations of Data for Football Analytics Research, *International Journal of Computer Science and Information Security*, 14 (11), 516.

Association, F. I. d. F., 1995, Laws of the game, FIFA., p.

Bazmara, M. ve Jafari, S., 2013, K Nearest Neighbor Algorithm for Finding Soccer Talent, *Journal of Basic and Applied Scientific Research*, 3 (4), 981-986.

Breiman, L., 2001, Random forests, *Machine learning*, 45 (1), 5-32.

Brink, M. S., Visscher, C., Arends, S., Zwerver, J., Post, W. J. ve Lemmink, K. A., 2010, Monitoring stress and recovery: new insights for the prevention of injuries and illnesses in elite youth soccer players, *British journal of sports medicine*, 44 (11), 809-815.

Brooks, J., Kerr, M. ve Guttag, J., 2016, Developing a data-driven player ranking in soccer using predictive model weights, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 49-55.

Brownlee, J., 2016, How To Create an Algorithm Test Harness From Scratch With Python, https://machinelearningmastery.com/create-algorithm-test-harness-scratch-python/: [6/22/2018].

BUNDESLIGA, 2018, Soccer positions explained: names, numbers and what they do, https://www.bundesliga.com/en/news/Bundesliga/soccer-positions-explained-names-numbers-what-they-do-507060.jsp: [6/22/2018].

Buursma, D., 2010, Predicting sports events from past results, *14th Twente Student Conference on IT*.

Carey, D. L., Ong, K., Morris, M. E., Crow, J. ve Crossley, K. M., 2016, Predicting ratings of perceived exertion in Australian football players: methods for live estimation, *International Journal of Computer Science in Sport*, 15 (2), 64-77.

Chapelle, O., Scholkopf, B. ve Zien, A., 2009, Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews], *IEEE Transactions on Neural Networks*, 20 (3), 542-542.

Constantinou, A. C., Fenton, N. E. ve Neil, M., 2012, pi-football: A Bayesian network model for forecasting Association Football match outcomes, *Knowledge-Based Systems*, 36, 322-339.

Cotta, L., de Melo, P., Benevenuto, F. ve Loureiro, A. A., 2016, Using fifa soccer video game data for soccer analytics, *Workshop on Large Scale Sports Analytics*.

Cox, D. R., 1958, The regression analysis of binary sequences, *Journal of the Royal Statistical Society. Series B (Methodological)*, 215-242.

Dey, S., 2017, Pricing Football Players using Neural Networks, *arXiv preprint arXiv:1711.05865*.

Donalek, C., 2011, Supervised and Unsupervised learning, *Astronomy Colloquia. USA*.

Dunning, E., 1999, The development of soccer as a world game, *Sport matters: sociological studies of sport, violence and civilisation.*, 80-105.

Ehrmann, F. E., Duncan, C. S., Sindhusake, D., Franzsen, W. N. ve Greene, D. A., 2016, GPS and injury prevention in professional soccer, *The Journal of Strength & Conditioning Research*, 30 (2), 360-367.

Enefiok, E., Nwachukwu, E. ve Williams, E. E., 2015, An Improved Decision Support System for a Football Team Manager, *International Journal of Engineering Research & Technology (IJERT)*, 4 (5).

Febianto, I., 2010, Decision support system for ideal placement of players position strategy in football formation, Jurusan Teknik Informatika, *Fakultas Teknik dan Ilmu Komputer, Universitas Komputer Indonesia*.

fieldoo, 2012, Who is a football scout?, http://blog.fieldoo.com/2012/12/who-is-a-football-scout/,

Gedikli, S., Bandouch, J., von Hoyningen-Huene, N., Kirchlechner, B. ve Beetz, M., 2007, An adaptive vision system for tracking soccer players from variable

camera settings, *Proceedings of the 5th International Conference on Computer Vision Systems (ICVS)*.

Giacomini, M., 2009, The technical guide for the schools of football, *Italy 2009*.

Gomes, J., Portela, F. ve Santos, M. F., 2015, Decision Support System for predicting Football Game result, *Computers-19th International Conference on Circuits, Systems, Communications and Computers-Intelligent Systems and Applications Special Sessions. Series*, 348-353.

Guyon, I. ve Elisseeff, A., 2003, An introduction to variable and feature selection, *Journal of machine learning research*, 3 (Mar), 1157-1182.

He, M., Cachucho, R. ve Knobbe, A., 2015, Football Player's Performance and Market Value, *Proceedings of the 2nd workshop of sports analytics, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), Porto, Portugal*.

Hijmans, A. ve Bhulai, S., 2017, Dutch football prediction using machine learning classifiers.

Ho, T. K., 1995, Random decision forests, *Document analysis and recognition, 1995., proceedings of the third international conference on*, 278-282.

Holsapple, C. W., 1978, Framework for a generalized intelligent decision support system.

Horton, M., Gudmundsson, J., Chawla, S. ve Estephan, J., 2014, Classification of passes in football matches using spatiotemporal data, *arXiv preprint arXiv:1407.5093*.

Hraste, M., Dizdar, D. ve Trninić, V., 2008, Experts opinion about system of the performance evaluation criteria weighted per positons in the water polo game, *Collegium antropologicum*, 32 (3), 851-861.

Huang, K.-Y. ve Chang, W.-L., 2010, A neural network method for prediction of 2006 world cup football game, *Neural Networks (IJCNN), The 2010 International Joint Conference on*, 1-8.

Hucaljuk, J. ve Rakipović, A., 2011, Predicting football scores using machine learning techniques, *MIPRO, 2011 Proceedings of the 34th International Convention*, 1623-1627.

Hughes, M., Caudrelier, T., James, N., Redwood-Brown, A., Donnelly, I., Kirkbride, A. ve Duschesne, C., 2012, Moneyball and soccer-an analysis of the key performance indicators of elite male soccer players by position, *Journal of Human Sport and Exercise*, 7 (2).

Huijgen, B. C., Elferink-Gemser, M. T., Post, W. ve Visscher, C., 2010, Development of dribbling in talented youth soccer players aged 12–19 years: A longitudinal study, *Journal of Sports Sciences*, 28 (7), 689-698.

Ibrahim, S. E., 2015, Predicate Project Outcomes Using Machine Learning, *International Journal of Science and Research (IJSR)*.

Igiri, C. P. ve Nwachukwu, E. O., 2014, An improved prediction system for football a match result, *IOSR Journal of Engineering (IOSRJEN)*, 4, 12-20.

Igiri, C. P., 2015, Support Vector Machine—Based Prediction System for a Football Match Result, *IOSR Journal of Computer Engineering (IOSR-JCE)*, 17 (3), 21-26.

Joseph, A., Fenton, N. E. ve Neil, M., 2006, Predicting football results using Bayesian nets and other machine learning techniques, *Knowledge-Based Systems*, 19 (7), 544-553.

Kaklauskas, A., 2015, Biometric and Intelligent Decision Making Support, Springer, p.

Kampakis, S., 2011, Comparison of machine learning methods for predicting the recovery time of professional football players after an undiagnosed injury.

Kampakis, S., 2016, Predictive modelling of football injuries, *arXiv preprint arXiv:1609.07480*.

Keen, P. G., 1980, Decision support systems: a research perspective, *Decision Support Systems: Issues and Challenges: Proceedings of an International Task Force Meeting*, 23-44.

Kenney, J., 1962, Moving averages, *Mathematics of Statistics*, 1, 221-223.

Kınalıoğlu, İ. H., Kuş, C. ve Kınacı, İ., 2017, Comparison of Artificial Neural Network and Data Mining Techniques for Prediction of 2017 Uefa Champions League, *3 rd International Researchers, statisticians and young statisticians congress 24-26 May 2017 Selçuk University.*

Klaiber, J. D., 2016, Soccer Player Performance Rating Systems for the German Bundesliga, *Ghent University*.

Kumar, G., 2013, Machine Learning for Soccer Analytics, Cambridge University Press, MSc thesis, KU Leuven.

Lasek, J., Szlávik, Z. ve Bhulai, S., 2013, The predictive power of ranking systems in association football, *International Journal of Applied Pattern Recognition*, 1 (1), 27-46.

Ma, J., Kwak, M. ve Kim, H. M., 2014, Demand trend mining for predictive life cycle design, *Journal of Cleaner Production*, 68, 189-199.

Markovits, A. S. ve Green, A. I., 2017, FIFA, the video game: a major vehicle for soccer's popularization in the United States, *Sport in Society*, 20 (5-6), 716-734.

Mathien, H., 2016, Collecting football data, https://github.com/hugomathien/football-data-collection:

McIntyre, M. ve Hall, M., 2005, Physiological profile in relation to playing position of elite college Gaelic footballers, *British Journal of Sports Medicine*, 39 (5), 264-266.

Miçoogullari, B. O., Gümüsdag, H., Ödek, U. ve Beyaz, Ö., 2017, Comparative Study of Sport Mental Toughness between Soccer Officials, *Universal Journal of Educational Research*, 5 (11), 1970-1976.

Moor, L., 2007, Sport and commodification: A reflection on key concepts, *Journal of sport and social issues*, 31 (2), 128-142.

Moroney, K., 2014, Predicting match outcomes through game events" *national college of ireland*.

Mulak, P. ve Talhar, N., 2015, Analysis of distance measures using K nearest neighbour algorithm on KDD dataset, *International Journal of Science and Research*, 4 (7), 2101-2104.

Ostojić, S. M., 2000, Physical and physiological characteristics of elite Serbian soccer players, *Facta universitatis-series: Physical Education and Sport*, 1 (7), 23-29.

Owramipur, F., Eskandarian, P. ve Mozneb, F. S., 2013, Football result prediction with Bayesian network in Spanish League-Barcelona team, *International Journal of Computer Theory and Engineering*, 5 (5), 812.

Papić, V., Rogulj, N. ve Pleština, V., 2009, Identification of sport talents using a web-oriented expert system with a fuzzy module, *Expert Systems with Applications*, 36 (5), 8830-8838.

Power, D. J., 2002, Decision support systems: concepts and resources for managers, Greenwood Publishing Group, p.

Prasetio, D., 2016, Predicting football match results with logistic regression, *Advanced Informatics: Concepts, Theory And Application (ICAICTA), 2016 International Conference On*, 1-5.

Pudil, P. ve Novovičová, J., 1998, Novel methods for feature subset selection with respect to problem knowledge, In: Feature extraction, construction and selection, Eds: Springer, p. 101-116.

Rasku, J., Kuusipalo, P. ve Joutsijoki, H., 2014, Decision Support Systems Review, *university of tampere*.

Ratomir, D., Aleksandar, J. ve Stanimir, J., 2004, The Most important Attributes of soccer players, *Physical Education and Sport*, 2 (1), 13-24.

Raven, P., Gettman, L., Pollock, M. ve Cooper, K., 1976, A physiological evaluation of professional soccer players, *British Journal of Sports Medicine*, 10 (4), 209-216.

Razali, N., Mustapha, A., Yatim, F. A. ve Ab Aziz, R., 2017, Predicting Football Matches Results using Bayesian Networks for English Premier League (EPL), *IOP Conference Series: Materials Science and Engineering*, 012099.

Reber, R. ve Perrig, W., 2001, Perception without awareness, Psychology of, *Psychological Science*, 2, 119-122.

Reilly, T. ve Holmes, M., 1983, A preliminary analysis of selected soccer skills, *Physical Education Review*, 6 (1), 64-71.

Roderick, M., 2006, The work of professional football: a labour of love?, Routledge, p.

Rossi, A., Pappalardo, L., Cintia, P., Iaia, M., Fernandez, J. ve Medina, D., 2017, Effective injury prediction in professional soccer with GPS data and machine learning, *arXiv preprint arXiv:1705.08079*.

Rotshtein, A. P., Posner, M. ve Rakityanskaya, A., 2005, Football predictions based on a fuzzy model with genetic and neural tuning, *Cybernetics and Systems Analysis*, 41 (4), 619-630.

Sarda, V., Sakaria, P. ve Deulkar, K., 2015, Football team selection using genetic algorithm, *International Journal of Engineering and Technical Research*, 3 (2), 153-156.

Sarma, V., 1994, Decision making in complex systems, *Systems practice*, 7 (4), 399-407.

Sathe, S., Kasat, D., Kulkarni, N. ve Prof. Satao , R., 2017, Predictive Analysis of Premier League Using Machine Learning, *International Journal of Innovative Research in Computer and Communication Engineering*, 5 (3).

Schulenkorf, N. ve Frawley, S., 2016, Critical issues in global sport management, Taylor & Francis, p.

Seltman, H., 2017, Experimental design and analysis. 2015, *Mixed models. A flexible approach to correlated data*, 357-377.

Sgro, F. ve Lipoma, M., 2016, Technical performance profiles in the European Football Championship 2016, *Journal of Physical Education and Sport*, 16 (4), 1304.

Shen, B., 2016, Result Prediction for Soccer Games, *Stanford University*

Shin, J. ve Gasparyan, R., 2014, A novel way to Soccer Match Prediction, *Stanford University: Department of Computer Science*.

Shlens, J., 2014, A tutorial on principal component analysis, *arXiv preprint arXiv:1404.1100*.

Silva, H., Uthuranga, S., Shiyamala, B., Kumarasiri, W., Walisundara, H. ve Karunarathne, G., 2009, A trainer system for air rifle/pistol shooting, *Machine Vision, 2009. ICMV'09. Second International Conference on*, 236-241.

Simchi-Levi, D., Kaminsky, P. ve Simchi-Levi, E., 2007, Designing and Managing the Supply Chain 3e with Student CD, McGraw-Hill/Irwin, July.

Şimşek, e., Aktuğ, Z. b., Çelenk, Ç., Yilmaz, T., Elif, T. ve Ersan, K., 2014, The evaluation of the physical characteristics of football players at the age of 9-15 in accordance with age variables, *International Journal of Science Culture and Sport (IntJSCS)*, 2 (5), 460-468.

Soto-Valero, C., 2017, A Gaussian mixture clustering model for characterizing football players using the EA Sports' FIFA video game system, *RICYDE. Revista Internacional de Ciencias del Deporte*, 13 (49).

Staub, S., Karaman, E., Kaya, S., Karapınar, H. ve Güven, E., 2015, Artificial Neural Network and Agility, *Procedia-Social and Behavioral Sciences*, 195, 1477-1485.

Sun, X., Lin, X., Shen, S. ve Hu, Z., 2017, High-resolution remote sensing data classification over urban areas using random forest ensemble and fully connected conditional random field, *ISPRS International Journal of Geo-Information*, 6 (8), 245.

Tavana, M., Azizi, F., Azizi, F. ve Behzadian, M., 2013, A fuzzy inference system with application to player selection and team formation in multi-player sports, *Sport Management Review*, 16 (1), 97-110.

Tax, N. ve 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.

Ulmer, B., Fernandez, M. ve Peterson, M., 2013, Predicting Soccer Match Results in the English Premier League, *Ph. D. dissertation*.

Uzochukwu, O. ve Enyindah, P., 2015, A Machine Learning Application for Football Players' Selection, *International Journal of Engineering Research & Technology (IJERT)*, 4 (10).

Van Gemert, D. ve van Ophem, J., 2010, Modelling the Scores of Premier League Football Matches, *Econometrics*, 18, 67.

Velcich, K., 2017, Predicting Soccer Match Results.

Venturelli, M., Schena, F., Zanolla, L. ve Bishop, D., 2011, Injury risk factors in young soccer players detected by a multivariate survival model, *Journal of science and medicine in sport*, 14 (4), 293-298.

Vroonen, R., Decroos, T., Van Haaren, J. ve Davis, J., 2017, Predicting the potential of professional soccer players, *Machine Learning and Data Mining for Sports Analytics ECML/PKDD 2017 workshop*.

Wagenaar, M., Okafor, E., Frencken, W. ve Wiering, M. A., 2017, Using Deep Convolutional Neural Networks to Predict Goal-scoring Opportunities in Soccer, *ICPRAM*, 448-455.

Wang, S., Liu, X., Jin, Y. ve Qu, K., 2015, Wind Power Short Term Forecasting based on Back Propagation Neural Network, *International Journal of Smart Home*, 9 (7), 231-240.

Yaldo, L. ve Shamir, L., 2017, Computational Estimation of Football Player Wages, *International Journal of Computer Science in Sport*, 16 (1), 18-38.

Yezus, A., 2014, Predicting outcome of soccer matches using machine learning, *Saint-Petersburg University*.

CURRICULUM VITAE

PERSONAL INFORMATION

Name and surname : Mustafa Aadel Mashjal AL-ASADI
Nationality : Iraqi
Date and place of birth : 16 March 1986 – Baghdad – Iraq
Telephone : 05378624294
E-mail: engyouth@yahoo.com

EDUCATION

Degree		Name, County, Province	Finish Year
Secondary	:	Al-Taamem Secondary School, Iraq, Baghdad	2004
University	:	Middle Technical University, College of Electrical and Electronic Techniques, Computer Engineering Department, Iraq, Baghdad	2008
Master	:	Selcuk University, Computer Engineering Department Turkey, Konya	2018

WORK EXPERIENCES

Year	Place of work	Responsibility
2011 Till Now	Ministry of youth and sport - Iraq	Engineer

PROFESSION

Machine learning, Intelligent Decision Support System, Computer vision & hardware.

FOREIGN LANGUAGES

Arabic	Mother Tongue
English	Good
Turkish	Good

PUBLICATIONS

Mustafa Aadel Al-Asadi, Şakir Taşdemir, Burak Tezcan, An Online Information System For Football Club Management, I.Internatıonal Congress Of Physical Education, Sport, Recreation and Dance, Page 204, Istanbul, April 26th- 28th, 2018.

Şakir Taşdemir, Mustafa Aadel Al-Asadi, Design An Intelligent Decision Support System For A Football Team Management, I.Internatıonal Congress Of Physical Education, Sport, Recreation and Dance, Page 207, Istanbul, April 26th-28th, 2018.