ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES

**MSc THESIS**

**Özge BOZKURT**

# DEVELOPMENT OF PHYSICAL FITNESS PREDICTION MODELS FOR TURKISH SECONDARY SCHOOL STUDENTS USING MACHINE LEARNING METHODS

**DEPARTMENT OF COMPUTER ENGINEERING**

**ADANA, 2019**

**ÇUKUROVA UNIVERSITY**
**INSTITUTE OF NATURAL AND APPLIED SCIENCES**

# DEVELOPMENT OF PHYSICAL FITNESS PREDICTION MODELS FOR TURKISH SECONDARY SCHOOL STUDENTS USING MACHINE-LEARNING METHODS

**Özge BOZKURT**

**MSc THESIS**

**DEPARTMENT OF COMPUTER ENGINEERING**

We certify that the thesis titled above was reviewed and approved for the award of degree of the Master of Science by the board of jury on

| …………………………..….. | ……………………………… | ……………………..….….……. |
|---|---|---|
| Assoc. Prof. Dr. Fatih AKAY | Asst. Prof. Dr. Onur ÜLGEN | Asst. Prof. Dr. Melis ÖZYILDIRIM |
| SUPERVISOR | MEMBER | MEMBER |

This MSc Thesis is written at the Department of Institute of Natural And Applied Sciences of Çukurova University.
**Registration Number**:

**Prof. Dr. Mustafa GÖK**
**Director**
**Institute of Natural and Applied Sciences**

**This study was supported by Research Projects Unit Ç.Ü. Project Number: FYL-2018-10430.**

**DEDICATION**

***… to my mom and dad***

# ABSTRACT

## MASTER THESIS

**DEVELOPMENT OF PHYSICAL FITNESS PREDICTION MODELS FOR TURKISH SECONDARY SCHOOL STUDENTS USING MACHINE LEARNING METHODS**



**Özge BOZKURT**

**ÇUKUROVA UNIVERSITY**
**INSTITUTE OF NATURAL AND APPLIED SCIENCES**
**DEPARTMENT OF COMPUTER ENGINEERING**

Supervisor : Assoc. Prof. Dr. M. Fatih AKAY
Year: 2019, Pages 63
Jury : Assoc. Prof. Dr. M. Fatih AKAY
: Asst. Prof. Dr. Onur ÜLGEN
: Asst. Prof. Dr. B. Melis ÖZYILDIRIM



Physical fitness is a necessary component for daily activities and has significant effect on our bodies. Because of this importance of physical fitness, maintenance of physical fitness is essential for health and well-being. Physical fitness is a set of attributes that are either health or skill-related, which can be measured with specific tests, and maintaining is essential for the quality of life. However, there are certain difficulties associated with the direct measurement of physical fitness such as the high cost of equipment, availability of experienced staff and the long assessment time. Since the measurement of physical activity has a key role in performing physical fitness rate, researchers need different ways to determine physical fitness. The aim of this thesis is to develop new prediction models for Turkish secondary school students by using machine learning methods including Support Vector Machines (SVM), Radial Basis Function Neural Network (RBFNN), General Regression Neural Network (GRNN), and Single Decision Tree (SDT). The performance of the SVM-based, GRNN-based, RBFNN-based and SDT-based models have been evaluated by using 10-fold cross-validation and the root means square errors (RMSEs) have been used to compute the errors of prediction. On the scope of the results, this thesis has showed the efficiency of machine-learning methods to indicate high accuracy of the physical fitness prediction.



**Key words:** Machine-learning, prediction, physical fitness, validation.

## ÖZ

### YÜKSEK LİSANS TEZİ

### MAKİNE ÖĞRENME YÖNTEMLERİ KULLANILARAK TÜRK ORTAOKUL ÖĞRENCİLERİ İÇİN FİZİKSEL UYGUNLUK TAHMİN MODELLERİ GELİŞTİRİLMESİ

**Özge BOZKURT**

**ÇUKUROVA ÜNİVERSİTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ**
**BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI**

Danışman : Doç. Dr. M. Fatih AKAY
Yıl: 2019, Sayfa: 63
Jüri : Doç. Dr. M. Fatih AKAY
: Dr. Öğr. Üyesi Onur ÜLGEN
: Dr. Öğr. Üyesi B. Melis ÖZYILDIRIM

Fiziksel uygunluk, vücudunuzu ve zihninizi sağlıklı tutan günlük aktivitelerin gerekli bir bileşenidir. Fitness yalnızca fiziksel olarak uygun olmakla kalmayıp aynı zamanda kişinin zihinsel durumunu da ifade eder. Fiziksel uygunluk, belirli testlerle ölçülebilen sağlık veya beceri ile ilişkili bir takım özelliklerdir ve fiziksel uygunluğun sağlanması, hayat kalitesi için gereklidir. Bununla birlikte, fiziksel uygunluğun ölçülmesi, gelişmiş profesyonel ekipman, deneyimli personel ve çok fazla zaman gerektirdiğinden, araştırmacıların fiziksel uygunluğu belirlemek için farklı yollar bulması gerekir. Bu çalışmanın amacı, Destek Vektör Makineleri (Support Vector Machine - SVM), Radyal Tabanlı Fonksiyon Sinir Ağı (Radial Basis Function Neural Network – RBFNN), Genel Regresyon Sinir Ağı (General Regression Neural Network - GRNN) ve Tek Karar Ağacı (Single Decision Tree - SDT) gibi makine öğrenme yöntemlerini kullanarak Türk orta öğretim öğrencilerinin fiziksel uygunluğunu tahmin etmek için yeni tahmin modelleri geliştirmektir. SVM, GRNN, RBFNN ve SDT temelli tahmin modellerinin performansı, 10 katlı çapraz doğrulama (10-fold cross validation) kullanılarak kök ortalama hata karesi (Root Mean Square Error-RMSE) hesaplanmıştır.

Elde edilen sonuçların ışığında, başta SVM olmak üzere makine öğrenme metodlarının kabul edilebilir ve doğru sonuçlar vermesi nedeniyle fiziksel uygunluk tahmininde kullanılabileceği kanıtlanmıştır.

**Anahtar Kelimeler:** Makine öğrenmesi, tahmin, fiziksel uygunluk, doğrulama

## EXTENDED ABSTRACT

Physical fitness is a set of features related to health or skill that can be measured by specific tests. Physical activity and exercise can have immediate or long-term benefits since maintenance of physical fitness is essential for health and wellbeing. Also has an important role in our lives by reducing the risks of several diseases like type 2 diabetes, cancer and cardiovascular disease etc.  Physical fitness involves the performance of the heart, lungs and body muscles (Bailey and Martin, 1994).

Recently, assessment of physical fitness for children (aged between 9 and 14 years) has become a popular research area, so in this concept, our study also focused on middle-aged children to assess and track their fitness level. Fitness training and student fitness assessments provide students with the opportunity to monitor and improve their fitness levels. Physical tests show the link between health and physical activity.

The evaluation of physical activity is necessary for the determination of the physical fitness rate, but it leads to difficulties in practice because of the disadvantages such as the difficulty of conducting the directly applied measurements with the crowded test groups and the possibility of allowing only one measurement simultaneously. However, in order to obtain accurate results with the help of direct measurement techniques professionally trained personnel, professional equipment and long time periods are required. Due to these disadvantages of direct measurements of physical fitness, researchers focused on machine-learning methods in order to evaluate physical fitness rates. Therefore, in literature we see a lot of studies that concentrate on predicting physical fitness rather than direct measurement and physical fitness assessment using test protocols. In order to extend these studies and enhance more accurate results, we focused on developing new prediction models with wide coverage for Turkish secondary school student by using various machine-learning methods and many physical fitness assessment protocols.

The need for detailed studies has been observed since the studies in this area contain restriction on the number of machine learning method and these limitations prevent the detailed understanding of the performance of the estimation variables used in the estimation models. For this reason, in order to comprehensively understand the effects of the estimation variables used in the measurement of physical fitness to the outcome, there is a need to carry out studies that produce more and more estimation models.

The aim of the study is develop new prediction models for Turkish secondary school students by using machine learning methods including Support Vector Machines (SVM), Radial Basis Function Neural Network (RBFNN), General Regression Neural Network (GRNN) and Single Decision Trees (SDT). The main purpose of this study is develop new prediction models for Turkish secondary school students by using machine-learning methods with model validation features in order to have comprehensive results.

The SVM method is classified as supervised machine-learning algorithm and state-of-the-art regression method, which is widely utilized in many application areas due to its high accuracy.

The GRNN method is a variation of the radial basis neural networks that does not require an iterative training procedure as back propagation networks. The principal advantages of the GRNN are being a fast learning method and available for application on large training sets.

The SDT based mechanism can be illustrated as a graphical approach of the solutions to make a possible decision according to precise conditions. The SDT model performs well even with large enough samples and also, due to decision tree mechanism has less complexity, it’s easier to comprehend results with high accuracy.

The RBFNN is a special type of artificial neural network and becoming a popular neural network model with diverse patterns. The RBFNN has the feed-forward structure and includes several layers and the first layer is called the input layer, the second is the hidden (kernel) layer with a non-linear RBFNN activation function, and the last layer is called a linear output layer.

In this thesis, various comprehensive models for physical fitness prediction were proposed and prediction models based on machine-learning methods have been developed for nine different exercise types which have wide coverage in concepts of fitness assessments. Participants were assigned to perform the following core stabilization assessments: 30 meter (m) Speed, 20m Shuttle Run, Standing Long Jump, Sit & Reach, Hand-Grip (right/left), Flexed Arm Hang, Balance and Agility tests to predict their physical fitness. Before the physical exercise tests, a consent participant form was signed by all subjects participating in this study.

The subjects' data have been collected by the College of Physical Education and Sports at Gazi University. Physical exercise tests were applied to the subjects to acquire their physical fitness. In total, the dataset comprises of the data of 300-350 healthy secondary school student subjects according to the target variables. The predictor variables which are (gender, age, height, weight, body fat, number of curl-ups, and number of push-ups in 30 seconds) were used to develop eight different physical fitness prediction models.

The general purpose of this study is to extend the previous works on the physical fitness prediction topic and utilize different machine-learning methods with comprehensive results. To the best of our knowledge, despite of existence of several studies which predict physical fitness with the help of statistical as well as machine-learning prediction methods, there is no comprehensive study that compares the performance of different machine-learning methods and various prediction models by using Turkish subject.

In this study, the predictor variables were analysed by data validation techniques according to physical fitness assessment tests, which led to an improved understanding of activity types and features. Model validation features called V-fold cross validation has been used to performing physical fitness prediction models. The performance of the prediction models has been evaluated by using the Root Mean Square Error (RMSE) in order to assess the performance of the prediction models.

In order to carry out this study, four well-known machine-learning methods, respectively SVM, GRNN, RBFNN and SDT, have been used and aimed to develop effective physical prediction models for school-aged children. In terms of the performances of machine-learning methods from the lowest RMSE to the highest ones can be listed as SVM, GRNN, RBFNN and SDT. According to the demonstrated results of the prediction models in most of the cases, SVM and RBFNN were selected the best and the worst models respectively, among the others in terms of accuracy rates.

Prediction models have been split into various categories based on predictor variables quantity. Gender, age, weight, height, body fat, number of curl-ups and number of push-ups of the samples are considered as the predictor variables used in this study. Eight different physical fitness prediction models have been developed by using machine-learning methods. As far as the results are concerned, the variables Curl-up, Push-up and Body Fat can be seen as the effective contributors. Moreover, based on analysis that we made for the model groups with all physical tests, prediction model with all predictor variables (Model 8) shows the highest performance in terms of prediction accuracy for all samples.

Regarding physical fitness test types, the RMSEs of the 30m speed test yields the best assessment results due to its lowest error rates (0.49 in average), and its independency on the machine-learning method or the prediction model was utilized. On the other hand, the RMSEs of the Standing Long Jump test yields the worst results with its highest error rates (20.60 in average). According to the findings from the developed models, the ones that include the predictor variables curl-up, push-up and body fat lead to the lowest RMSEs.

The machine-learning approach for the physical fitness prediction is relatively new and mainly focused on the educational area as we aimed in this study as well to obtained results are clearly promising. This thesis has showed the efficiency of machine-learning methods to indicate high accuracy of the physical fitness prediction for specified subject groups. On the other hand, based on our observations, different machine-learning methods and additional datasets that include various predictor variables can improve the accuracy of results for physical fitness prediction.

## GENİŞLETİLMİŞ ÖZET

Fiziksel uygunluk, belirli testlerle ölçülebilen sağlık veya beceri ile ilgili bir dizi özelliktir ve fiziksel uygunluğu korumak sağlık ve esenlik için çok önemlidir. Ayrıca, tip 2 diyabet, kanser ve kalp-damar hastalıkları vb. gibi çeşitli hastalıkların risklerini azaltarak yaşamımızda önemli bir role sahiptir. Fiziksel uygunluk, kalbin, akciğerlerin ve vücut kaslarının performansını içerir (Bailey and Martin, 1994). Bedenlerimiz ile gerçekleştirdiğimiz aktiviteler, zihnimizle yapabileceklerimizi de etkilediğinden, zindelik, zihinsel uyanıklık ve duygusal istikrar gibi nitelikleri etkiler. Bu nedenle, fitness sadece fiziksel olarak uygun olmayı değil, aynı zamanda bir kişinin zihinsel durumunu da ifade eder.

Son zamanlarda, çocuklar için fiziksel uygunluğun değerlendirilmesi (9 ila 14 yaş arası) popüler bir araştırma alani haline gelmiştir, bu nedenle bu çalışmamız aynı zamanda uygunluk seviyelerini değerlendirmek ve izlemek için bu yaş aralığındaki çocuklara odaklanmıştır. Fitness eğitimi ve öğrenci fitness değerlendirmeleri, öğrencilere fitness seviyelerini değerlendirme, izleme ve geliştirme fırsatı sunar.

Fiziksel testler, sağlık ve fiziksel aktivite arasındaki bağlantıyı gösterir. Fiziksel aktivitenin ölçülmesi, fiziksel uygunluk oranının tespiti için gereklidir, ancak doğrudan uygulanan ölçümlerin kalabalık test grupları ile yürütülmesinin zorluğu ve eş zamanlı sadece bir ölçüme imkan vermesi gibi olumsuzlukları nedeniyle uygulamada güçlüklere neden olmaktadır. Bununla birlikte doğrudan ölçüm teknikleri yardımıyla doğru sonuçlar elde edilebilmesi için profesyonel olarak eğitilmiş personeller ve uzun zaman periyotları gerekmektedir. Söz konusu faktörler değerlendirildiğinde, araştırmacılar için doğrudan ölçüm yöntemlerinin yerine fiziksel uygunluğun tahmini için makine öğrenmesi metodlarının kullanılması önerilebilir. Bu alanda yapılan çalışmaların makine öğrenmesi metodu sayısı ile ilgili kısıtlamalar içermesi ve bu kısıtlamaların tahmin modellerinde kullanılan tahmin değişkenlerine ait performansın ayrıntılı bir şekilde kavranmasını engellemesi nedeniyle detaylı çalışma yapma ihtiyacı gözlemlenmiştir. Bu sebepten dolayı fiziksel uygunluk ölçümünde kullanılan tahmin değişkenlerinin sonuca etkilerinin kapsamlı bir şekilde kavranabilmesi için daha yeni ve fazla sayıda tahmin modeli üreten çalışmaların yürütülmesine ihtiyaç duyulmuştur. Literatürde istatistiksel ve makine öğrenme yöntemlerini kullanan çeşitli fiziksel uygunluk tahmin modelleri önerilmiştir. Bu çalışmaları genişletmek ve daha doğru sonuçlar geliştirmek için, çeşitli makine öğrenme yöntemleri ve birçok fiziksel uygunluk değerlendirme protokolünü kullanarak, Türk ortaokul öğrencisi için geniş kapsamlı yeni tahmin modelleri geliştirmeye odaklandık.

Söz konusu çalışmanın amacı, Destek Vektör Makineleri (Support Vector Machines, SVM), Radyal Temel Fonksiyon Yapay Sinir Ağları (Radial Basis Function Neural Network, RBFNN), Genel Regresyon Sinir Ağı (General Regression Neural Network, GRNN) ve Tekli Karar Ağaçları (Single Decision Tree, SDT) gibi makine öğrenme yöntemlerini kullanarak Türk ortaokul öğrencileri için yeni tahmin modelleri geliştirmektir. Bu çalışmanın temel amacı, kapsamlı sonuçlar elde etmek için model doğrulama özelliklerine sahip makine öğrenme yöntemlerini kullanarak Türk ortaöğretim öğrencileri için yeni tahmin modelleri geliştirmektir.

SVM yöntemi, yüksek doğruluğu nedeniyle birçok uygulama alanında yaygın olarak kullanılan denetimli makine öğrenme algoritması ve son teknoloji regresyon yöntemi olarak sınıflandırılır.

GRNN yöntemi, geri yayılma ağları olarak yinelemeli bir eğitim prosedürü gerektirmeyen radyal temelli sinir ağlarının bir çeşididir. GRNN'nin temel avantajları hızlı bir öğrenme yöntemi olmasi ve büyük eğitim setlerinde uygulanabilir olmasıdır.

SDT tabanlı makine öğrenmesi mekanizması, kesin koşullara göre olası bir karar vermek için çözümlerin grafiksel bir yaklaşımı olarak gösterilebilir. Karar ağacı öğrenmesi, bir öğeye ilişkin gözlemlerden (dallarla temsil edilir), öğenin hedefe ulaşma karar aşamaları olarak bilinen sonuçlara varmak için bir karar ağacı (öngörücü bir model olarak) kullanır. SDT tabanlı modellerin yeterince büyük örneklerle bile iyi performans gösteriyor olması ve ayrıca karar ağacı mekanizmasının daha az karmaşık olması nedeniyle sonuçları yüksek hassasiyetle ortaya çıkarmaktadır.

RBFNN özel bir yapay sinir ağı türüdür ve popüler bir sinir ağı modeli haline gelmektedir. RBFNN ileri besleme yapısına sahiptir ve birkaç katman içerir bunlar sırasıyla, birinci katman girdi katmanı olarak adlandırılır, ikincisi doğrusal olmayan bir RBFNN aktivasyon işlevi olan gizli (çekirdek) katmanıdır ve son katmana doğrusal çıktı katmanı denir. Her katmandaki düğümler önceki katmana tamamen bağlıdır ve bu tür modeller genel tahmin özelliğine sahiptir.

Bu tez çalışmasında, fiziksel uygunluk tahmini için çeşitli kapsamlı modeller önerilmiş ve fitness değerlendirme kavramlarında geniş kapsamı olan dokuz farklı egzersiz tipi için makine öğrenme yöntemlerine dayalı tahmin modelleri geliştirilmiştir. Veri seti, 30 metre (m) sürat, 20m mekik koşusu, durarak uzun atlama, otur-eriş testi, el kavrama kuvveti sağ/sol, barfix kol kuvveti, denge ve çeviklik testi gibi hedef değişkenlere göre çeşitli testlerin verilerini içermektedir. Fiziksel egzersiz testlerinden önce, bu çalışmaya katılan tüm öğrencilere bir onay formu doldurulmuştur. Öğrencilerin verileri Gazi Üniversitesi Beden Eğitimi ve Spor Yüksekokulu tarafından toplanmıştır. Fiziksel zindeliği kazanmak için öğrencilere fiziksel egzersiz testleri uygulanmıştır. Toplamda, veri seti, hedef değişkenlere göre 300-350 sağlıklı lise öğrencisinin verilerinden oluşmaktadır. Sekiz farklı fiziksel uygunluk tahmin modeli geliştirmek için cinsiyet, yaş, boy, kilo, vücut yağ, kıvrılma sayısı ve 30 saniyedeki şınav sayısı değişkenleri kullanılmıştır.

Bu çalışmanın genel amacı, fiziksel uygunluk tahmini konusundaki önceki çalışmaları genişletmek ve kapsamlı sonuçlar veren farklı makine öğrenme yöntemlerini kullanmaktır. Bildiğimiz kadarıyla, istatistiksel ve makine öğrenmesi tahmin yöntemlerinin yardımıyla fiziksel uygunluğu öngören birçok çalışmanın varlığına rağmen, farklı makine öğrenme yöntemlerinin ve çeşitli tahmin modellerinin performansını karşılaştıran kapsamlı bir çalışma yoktur.

Bu çalışmada değişkenlerin ve etkinlik türlerinin özelliklerinin daha iyi anlaşılmasını sağlayan fiziksel uygunluk değerlendirme testleri veri doğrulama teknikleriyle analiz edilmiştir. Fiziksel uygunluk tahmin modellerini gerçekleştirmek için V-katlama çapraz doğrulama adı verilen model doğrulama özellikleri kullanılmıştır. Tahmin modellerinin performansı, tahmin modellerinin performansını değerlendirmek için Kök Ortalama Kare Hatası (RMSE) kullanılarak değerlendirilmiştir.

Bu çalışmayı yürütmek için dört iyi bilinen makine öğrenme yöntemi, sırasıyla SVM, GRNN, RBFNN ve SDT kullanılmış ve okul çağındaki çocuklar için etkili fiziksel tahmin modelleri geliştirmeyi amaçlamıştır. Makine öğrenme yöntemlerinin performansları bakımından en düşük RMSE'den en yüksek RMSE'e göre SVM, GRNN, RBFNN ve SDT olarak listelenebilir. Tahmin modellerinin gösterilmiş sonuçlarına göre, çoğu durumda, doğruluk oranları açısından diğerleri arasında SVM ve RBFNN, en iyi ve en kötü modeller olarak seçilmiştir.

Tahmin modelleri, tahmin değişkenlerine göre çeşitli kategorilere ayrılmıştır. Cinsiyet, yaş, kilo, boy, vücut yağı, kıvrılma sayısı ve örneklerin şınav sayısı bu çalışmada kullanılan tahmin değişkenleri olarak kabul edilmiştir. Makine öğrenmesi yöntemleri kullanılarak sekiz farklı fiziksel uygunluk tahmin modeli geliştirilmiştir. Sonuçlar söz konusu olduğunda, Curl-up, Push-up ve Vücut Yağı değişkenleri etkin katkı sağlayanlar olarak görülebilir. Ayrıca, tüm fiziksel testler ile model grupları için yaptığımız analizlere dayanarak, tüm tahmin değişkenleri içeren tahmin modeli (Model 8), tüm gruplar içinde tahmin doğruluğu açısından en yüksek performansı göstermektedir.

Fiziksel uygunluk test tipleri ile ilgili olarak, 30m hız testi makine öğrenme yöntemi veya tahmin modelinden bağımsız olarak en düşük hata (RMSE) oranını vermiştir (ortalama 0.49). Öte yandan, durarak uzun atlama testi'nin RMSE'leri en yüksek hata oranlarıyla en kötü sonuçları vermektedir (ortalama 20,60). Geliştirilen modellerden elde ettiğimiz bulgulara göre, tahmin değişkenleri (kıvrılma, şınav ve vücut yağı) içeren modeller en düşük RMSE'lerle en iyi sonucu verdiği gözlemlenmiştir. Elde edilen sonuçlar değerlendirildiğinde, başta SVM olmak üzere makine öğrenme metodlarının kabul edilebilir ve doğru sonuçlar vermesi nedeniyle fiziksel uygunluk tahmininde kullanılabileceği kanıtlanmıştır.

Fiziksel uygunluk tahmini için makine öğrenme yaklaşımı nispeten yeni bir alan ve eğitim alanında elde edilen sonuçlar neticesinde umut verici bir şekilde ilerlemektedir. Bu tez, belirtilen öğrenci grupları için fiziksel uygunluk tahmininin yüksek oranda elde edilen doğruluk oranlarını göstermek için makine öğrenme yöntemlerinin kullanımının avantajlarını göstermiştir. Öte yandan ilerdeki çalışmalar için, gözlemlerimize dayanarak, farklı makine öğrenme yöntemleri ve çeşitli tahmin değişkenleri içeren ek veri setleri, fiziksel uygunluk tahmini için sonuçların doğruluğunu artırabilir.

**ACKNOWLEDGMENTS**

Foremost, I would like to express my sincere gratitude to my advisor Assoc. Prof. Dr. M. Fatih AKAY, for his supervision guidance, motivation and his valuable time for this MSc thesis.

I would like to thank members of the MSc thesis jury, Asst. Prof. Dr. Onur ÜLGEN and Asst. Prof. Dr. Buse Melis ÖZYILDIRIM, for their suggestions and corrections.

I would also like to thank Cukurova University Scientific Research Projects Center for supporting this work (Project no: FYL-2018-10430).

Last but not the least, I would like to thank my family for their endless support and encouragements for my life and career.

**CONTENTS** **PAGE**

**LIST OF TABLES** **PAGE**

# XIX

**LIST OF FIGURES** **PAGE**

## LIST OF ABBREVIATIONS

| | | |
|---|---|---|
| SVM | : | Support Vector Machines |
| RBF | : | Radial Basis Function |
| RBFNN | | Radial Basis Function Neural Network |
| GRNN | : | General Regression Neural Network |
| SDT | : | Single Decision Tree |
| TB | : | Tree Boost |
| ADA | : | Ada Boost |
| KNN | : | K Neighbors Classifier |
| SGD | : | Stochastic Gradient Descent |
| SVC | : | Support Vector Classification |
| DT | : | Decision Tree |
| RF | : | Random Forest |
| LR | : | Logistic Regression |
| NB | : | Naive Bayesian |
| NN | : | Neural Network |
| MLR | : | Multiple Linear Regression |
| RMSE | : | Root Mean Square Errors |
| SEE | : | Standart Error of Estimate |
| R | : | Multiple Correlation Coefficient |
| CBR | : | Case-Based Reasoning |
| MET | : | Metabolic Equivalent |
| BMI | : | Body Mass Index |

XXIII

# l. INTRODUCTION

## 1.1. Overview of Physical Fitness

Physical fitness can be defined as the quality of life and state of being physically fit. The modern definition of fitness describes a person can perform any specific function. This has led to interrelation of human adaptability to cope with various physical situations. In theory, physical fitness involves the performance of the heart, lungs, and the muscles of the body. Physical fitness can be defined by examining four basic parts: Cardiorespiratory endurance, muscular strength, muscular endurance and flexibility. Cardiorespiratory endurance is the ability to deliver oxygen and nutrients to tissues and remove wastes over sustained periods of time (Hoffman, 2006). Long runs and swimming are among the methods employed in measuring this component. Muscular strength is the ability of a muscle to exert force for a short period of time. Upper body strength, for example, can be measured by various weight-lifting exercises. Muscular endurance is the ability of a muscle, or group of muscles, to sustain repeated contractions or continue applying force against a fixed object (Chin et al, 2002). Push-ups are often used to test endurance of arm and shoulder muscles. Flexibility is the ability to move joints and use muscles through their full range of motion. The sit and reach is a good measure of flexibility of the lower back and the back of the upper legs.

Regarding specific functions, fitness is attributed to persons who possess significant aerobic or anaerobic ability, i.e. strength or endurance. Physical fitness has a significant effect on our daily activities and maintaining healthy lifestyles. In order to have a sufficient health level evaluation, the measurement of physical fitness has become a fundamental component for health monitoring as part of public health measures (Witten et al, 2016). A well-rounded fitness consideration might improve a person in all aspects of well-being. Since what we do with our bodies also affects what we can do with our minds, fitness influences mental alertness and emotional stability. Physical fitness also cites mental, social and emotional health as an important part of overall fitness (Pink, 2015). This approach is often represented in three points as physical, emotional, and mental fitness. Physical fitness can also prevent or treat many chronic health conditions brought on by unhealthy lifestyle or aging (Glassman, 2002). Moreover, according to some researches, workout lifestyle can also help people to have better sleep periods and possibly alleviate some mood disorders in certain individuals (Hu et al, 2010). Developed researches were demonstrated benefits of fitness through the significant role of being physically fit.

Physical fitness assessments give an opportunity to observe individuals' fitness level and measurement of physical activity is essential to fitness performance rate. However, the evaluation requires improved professional equipment, well-educated staff and long assessment time. Due to these disadvantages of direct measurements of physical fitness, researchers focused on machine-learning methods in order to compute physical fitness rates. In literature section, details of the case studies that physical fitness prediction models using machine-learning methods have been reviewed.

### 1.2. Motivation and Purpose of this Thesis

As we mentioned above physical fitness is important for the survival of human-beings and a necessary component for daily activities. Fitness does not only refer to being physically fit, but also refers to a person's mental state as well. Moreover, fitness education and fitness assessments offer students an opportunity to assess, track and improve their fitness level. Physical tests show the link between fitness, health and physical activity. Professional equipment can produce highly accurate physical evaluations. However, as mentioned before, their utilization is associated with several difficulties and limitations such as following points:

- High cost and low availability of some of the necessary measurement equipment. Reducing the utilization of such equipment within research projects at educational institutions or in certain sports research centers.

- The impracticality and lack of mobility of these equipment limit their usage in specific places and purposes only.
- Individual-only usability.
- Additionally, the evaluation requires a professional test zone, qualified staff and long time in order to be interpreted.

In the light of the drawbacks mentioned above, rather than directly measuring the physical fitness, researchers have recently focused on machine-learning approaches to determine physical fitness predictions. On that perspective, various studies have been examined and observed. Depends on the observations that we made, these literature studies have important limitations in scope of the prediction models and the techniques utilized. This can be considered as an obstacle since it prevents having detailed comprehension of the performance for each predictor variables used to have more precise results in order to conduct prediction mechanism.

The fundamental aim of this thesis is to develop new prediction models for specified group of subjects by using machine-learning methods including Support Vector Machines (SVM), Radial Basis Function Neural Network (RBFNN), General Regression Neural Network (GRNN), and Single Decision Tree (SDT). It is also aimed to extend the previous studies on the physical fitness prediction topic and utilize different machine-learning methods and more comprehensive prediction models by using data validation methods such as 10-fold cross-validation. Root mean square errors (*RMSEs*) were used to have comprehensive results and compute the prediction methods accuracy.

## 1.3. Review of the Literature Studies

As a result of literature reviews, some related studies that utilized machine-learning methods to predict physical fitness rate can be encounter as case studies. In (Fergus et al, 2015), a supervised machine-learning approach was implemented for physical activity measurements. The main concept of this study shows physical activity measures are vital to maintaining healthy lifestyles and making suggestions about physical activity levels. The study provides appropriate activities to measure physical activity by using a multi-layered neural network to classify physical activities according to the activity type. Results show the possibility of obtained results for an overall accuracy when three and four feature combinations were used. In (Dijkhuis et al, 2018), an activity tracker to record participants' daily step count was used as an input for a coaching session to train eight different machine-learning algorithms. Hourly estimations were handled by daily steps threshold for the number of 48 subjects. According to findings, the Random Forest (RF) algorithm showed the best performance in the process of automated personalized coaching. In (Mobyen et al, 2013), a Case-Based Reasoning (CBR) an approach to identify the physical activity of a group of elderly based on pulse rate was proposed. The CBR approach was compared with the two popular classification techniques including SVM and Neural Network (NN) on 24 subjects. In (Barnekow et al, 1998), the associations of self-reported measures of physical activity from a mail survey with an objective measure of physical fitness were investigated. From this group, males who also had a clinical examination within 60 days of the return of their questionnaire served as the subjects. The subjects of the study completed a maximal physical fitness assessment. Reported exercise participation values were converted to estimates of energy expenditure and combined into overall indices of the physical activity participation. Multiple regression analyses were used to determine the individual contributions of the physical activity indices in predicting maximal treadmill performance (physical fitness). Significant predictors of the physical fitness were age, an index of running, walking and jogging participations, and the response to a question on frequency of sweating. These results indicate that the exercise behaviour can be accurately estimated in large populations by using simple questions in a mail survey. In (Jaakkola et al, 2016), fundamental movement skills and physical fitness scores were assessed in early adolescence to predict self-reported physical activity. The effects of previous physical activity, sex, and BMI were investigated as main concept. Adolescents' fundamental movement skills, physical fitness, self-report physical activity and BMI were collected at baseline, and their self-report energy expenditure and intensity of physical activity were collected by using the International Physical Activity Questionnaire for an identified period of time. Results show that sex and BMI have an impact on the fundamental movement skills, while physical fitness affects energy expenditure. In (Reichherzer et al, 2017), the data analysis methods were used to train a classifier for records with the individuals. Four different machine-learning algorithms including Decision Tree (DT), Random Forest (RF), SVM, Naive Bayesian (NB) were used to make predictions for 29 participants by the evaluation of physical activities. Results of the study show that the methods allow individuals and their activities to be tracked.

Summaries of the some recent studies can be found in Table 1.1.

Table 1.1. Summary of studies

| Study | Methods | Writers |
|---|---|---|
| 'Development of Physical Fitness Prediction Models for Turkish Secondary School Students Using Machine Learning Methods' | SVM, RBFNN, TB | Akay et al, 2018 |
| 'Multiple Linear Regression Based Physical Fitness Prediction Models for Turkish Secondary School Students' | MLR | Akay et al, 2018 |
| 'Predicting Physical Fitness Using Support Vector Machines and Blood Test Results' | SVM | Akay et al, 2018 |
| 'A Machine Learning Approach to Measure and Monitor Physical Activity in Children to Help Fight Overweight and Obesity' | MLP | Fergus et al, 2015 |
| 'Personalized Physical Activity Coaching : A Machine Learning Approach' | RF, ADA, DT, KNN, LR, NN, SGD, RF, SVC | Dijkhuis et al, 2018 |
| 'Physical Activity Identification using Supervised Machine Learning and based on Pulse Rate' | CBR, SVM, NN | Ahmed M.U. and Loutfi A., 2013 |
| 'Using Machine Learning Techniques to Track Individuals and Their Fitness Activities' | DT, RF, SVM, NB | Reichherzer et al, 2017 |

According to the findings and reviews from the literature, following case studies can be examined as a part of the thesis by using the material included:

- The first related case study shown in (Akay et al, 2018). The SVM, RBFNN and Tree Boost (TB) have been applied to generate new prediction models for predicting the physical fitness of Turkish secondary school students by using machine-learning methods. The dataset comprises of the data for various numbers of subjects according to the target variables such as the test scores of the 30m speed sprint, 20m stage run, balance and agility. The predictor variables (gender, age, height, weight, body fat, number of curl-ups and push-ups in 30 seconds) have been used to develop the prediction models. To do that, eight physical fitness prediction models have been created with the variables listed above. The authors claimed that the SVM model yields lower RMSEs than the ones obtained from the RBFNN and the TB prediction models. As conclusion, the results indicate that the predictor variables body fat, push-up and curl-up play a significant role when used all together for physical fitness prediction.

- The second case study shown in (Akay et al, 2018) by using the MLR method. The authors aimed to develop prediction models for specified group of subjects and the datasets comprised the target variables including the test scores of the 30m speed, the 20m stage run, the balance and the hand-grip (right/left). The predictor variables (gender, age, body mass index (BMI), body fat, number of curl-up and push-ups in 30 seconds) have been used to develop the prediction models. So, eight physical fitness prediction models for each target have been created with the predictor variables listed above and the performance of the prediction models has been calculated by using standard error of estimate (*SEE*). At the end, their study shows that the MLR-based prediction models are suitable for physical fitness prediction.
- The third study has been carried out in (Akay et al, 2018) named `Predicting physical fitness using support vector machines and blood test results`. In this study SVM and blood test results have been

applied with seven prediction models by utilizing the combinations of the blood test results. The seven prediction models have been divided into a couple of categories according to the number of predictor variables employed in the prediction model. To compare performance of the SVM-based models, prediction models based on the RBFNN and TB have also been developed. The performance of all prediction models has been calculated by *SEE*. The results indicated that the SVM-based prediction models outperform the RBFNN-based and the TB-based prediction models regardless of the independent category for the applied prediction models. Furthermore, the authors claimed that among the seven prediction models developed, the prediction models that include the predictor variables cholesterol, hemoglobin and triglyceride play a significant role in physical fitness prediction based on blood tests. The category of models with three predictor variables on average yields the best results while the one with only one predictor variable shows the worst.

### 1.4. Contributions of the Thesis

As a conclusion or literature reviews, some of the studies concentrated on classifying physical activity levels rather than predicting the actual test results. On the other hand, some of the developed models are not suitable for determining the physical activity level of the Turkish students due to the genetic factors depending on the location. Additionally, the developed models require a subject to complete several exercises in order to have his/her physical activity level classified. Differences between the research proposal and the literature case studies have been summarized in this section as following.

The most distinct contributions of this thesis compared to the previous study can be listed as follows:

1. In literature, to the best of our knowledge, despite of the existence of several studies which predict physical fitness with the help of statistical as well as machine-learning prediction methods, there is no comprehensive study that compares the performance of different machine-learning methods and various prediction models by using Turkish subjects.
2. This study has a significant role in encouraging the use of machine-learning methods for physical fitness prediction.
3. In this thesis, various comprehensive models for physical fitness prediction were proposed and prediction models based on machine-learning methods have been developed for nine different exercise types which have wide coverage in concepts of fitness assessments.
4. In this thesis, various predictor variables including (gender, age, weight, height, body fat, number of curl-ups and number of push-ups) were utilized. Moreover, using nine different exercise types with different combinations of predictor variables made possible to create more detailed prediction models. More specifically, around 160 different prediction models have been developed.
5. Around a number of 300-350 healthy secondary school students participated in the development of the precise dataset used in this thesis.

In the main concept, the purpose of this study is to extend the previous works on the physical fitness prediction topic and utilize different machine-learning methods with comprehensive results.

### 1.5. An Overview for the Dataset

The precise dataset comprises of the data of 300-350 healthy secondary school student subjects according to the target variables. The predictor variables which are (gender, age, height, weight, body fat, number of curl-ups, and number of push-ups in 30 seconds) were used to develop eight different physical fitness prediction models. The participants (specified subjects) were assigned to perform the following core stabilization assessments 30 meter (m) Speed, 20m Shuttle Run, Standing Long Jump, Sit & Reach, Hand-Grip (right/left), Flexed Arm Hang, Balance and Agility tests to predict their physical fitness. The students who participated in the experiments were assessed with the usage of fitness equipment are given in Figure 1.1 through Figure 1.8.

In 30m speed test, length of non-slip running area is 30 meter with indicated start and end lines. The subject requires run from start line and back to the same line to finish the running area as soon as possible.

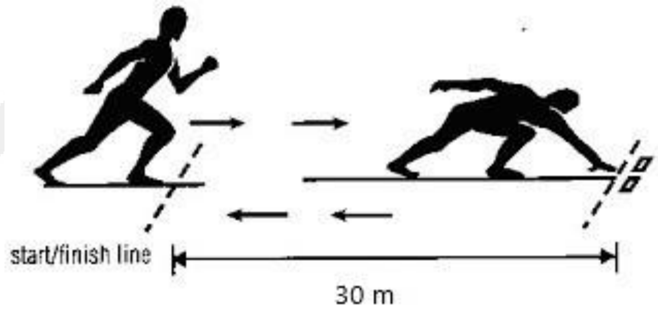


Figure 1.1. 30m Speed Test

The purpose of the 20m Shuttle test is to estimate the participant's maximum oxygen consumption capacity. The subject requires run 20 meter round trip between two signals period. The subject needs to start with first signal and back again to the starting line then continued with the other signals while timer is on.

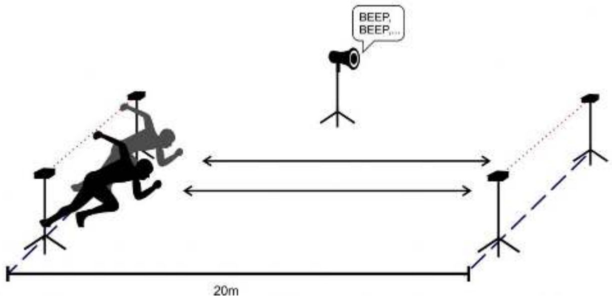


Figure 1.2. 20 m Shuttle Run Test

In standing long jump test, the participant is measured in cm between the line at the end of the long jump at which the two legs are connected to each other and the last distance left by the athlete without taking any speed.

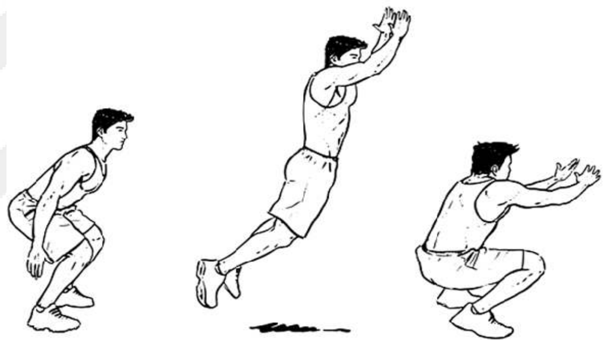

Figure 1.3. Standing Long Jump Test

Sit and reach testing is one of the way to measure tightness in lower back and hamstring muscles. The test is performed as it is shown in the figure 1.4. Tightness is measured by the number of inches from hands to reaching point in forward.

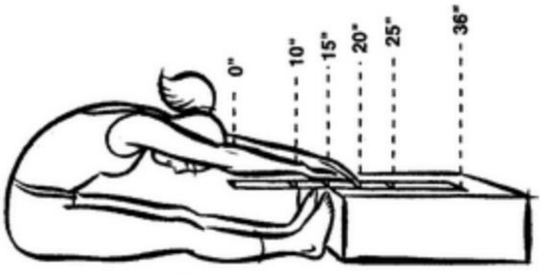


Figure 1.4. Sit & Reach Test

In hand-grip test, the dynamometer is set according to the hands of the participant to measure strength.


Figure 1.5. Hand-Grip Test (Right/Left)

In flexed arm hang test, the round horizontal bar is set in a way with the shoulders wide, while the thumb at the bottom and the other fingers at the bottom above the bar. The hold time is recording with a timer for each subject.

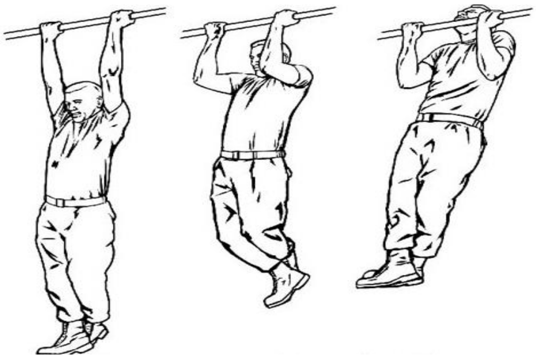

Figure 1.6. Flexed Arm Hang Test

The balance test is useful way to determine the balance of subjects while the subjects are balanced with one foot, the time starts counting.

Figure 1.7. Balance Test

In agility test, participant starts running straight from the ‘A’ funnel to the ‘B’ funnel and touched the funnel with right hand. After that, the participant keeps running towards the ‘C’ hunter and touches the funnel with left hand. Then the participant should run to the right side and touches ‘D’ hunter with right hand. From here the subject needs to come back to the ‘B’ hunter with a side run, left hand touching and back to ‘A’ hunter while time is recorded.

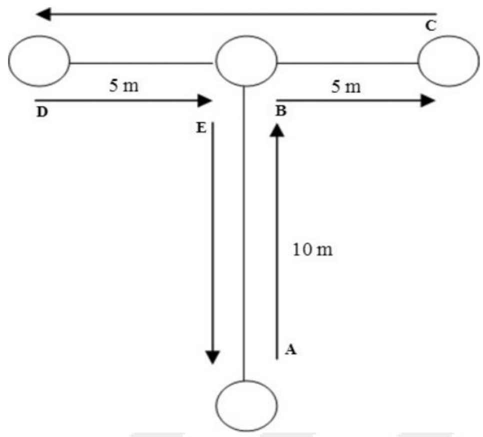


Figure 1.8. Agility Test

Statistical information about the datasets with the maximum and minimum rates, average points and standard deviation rates for each predictor and/or target variables are given in Table 1.2 through Table 1.10 regarding to datasets.

Table 1.2. Statistical Information of the 30M Speed Test

| **Variable** | **Min. Value** | **Max. Value** | **Mean Value** | **Standard Deviation Value** |
|---|---|---|---|---|
| Gender | 0 | 1 | 0.49 | 0.5 |
| Age (Year) | 11 | 16 | 12.75 | 0.9 |
| Height (m) | 130 | 182 | 152.99 | 8.45 |
| Weight (kg) | 21 | 107.6 | 46.509 | 12.41 |
| BMI (kg/m2) | 13.5 | 43.7 | 19.82 | 3.97 |
| Curl-up | 0 | 26 | 11.88 | 5.07 |
| Push-up | 0 | 30 | 8.86 | 7.04 |
| Body Fat | 9.6 | 41.7 | 23.03 | 6.54 |
| 30m Speed (s) | 4.4 | 8.8 | 5.98 | 0.6 |

Table 1. 3. Statistical Information of the 20M Shuttle Run Test

| Variable | Min. Value | Max. Value | Mean Value | Standard Deviation Value |
|---|---|---|---|---|
| Gender | 0 | 1 | 0.49 | 0.5 |
| Age (Year) | 11 | 16 | 12.81 | 0.91 |
| Height (m) | 130 | 182 | 153.11 | 12.09 |
| Weight (kg) | 21 | 107.6 | 46.87 | 12.47 |
| BMI (kg/m2) | 14 | 32.5 | 19.79 | 3.72 |
| Curl-up | 0 | 26 | 11.89 | 5.12 |
| Push-up | 0 | 30 | 8.99 | 7.03 |
| Body Fat | 9.6 | 41.7 | 23.02 | 6.55 |
| 20m shuttle Run(s) | 3.2 | 7.1 | 6.05 | 0.8 |

Table 1.4. Statistical Information of the Standing Long Jump Test

| Variable | Min. Value | Max. Value | Mean Value | Standard Deviation Value |
|---|---|---|---|---|
| Gender | 0 | 1 | 0.49 | 0.5 |
| Age (Year) | 11 | 16 | 12.70 | 0.92 |
| Height (m) | 130 | 182 | 152.89 | 8.39 |
| Weight (kg) | 21 | 107.6 | 46.49 | 12.43 |
| BMI (kg/m2) | 13.5 | 43.7 | 19.81 | 4.08 |
| Curl-up | 0 | 26 | 11.67 | 5.001 |
| Push-up | 0 | 30 | 8.84 | 7.09 |
| Body Fat | 9.6 | 41.7 | 23.05 | 6.61 |
| Standing Long Jump(m) | 13.5 | 198.5 | 121.46 | 21.50 |

Table 1.5. Statistical Information of the Sit & Reach Test

| Variable | Min. Value | Max. Value | Mean Value | Standard Deviation |
|---|---|---|---|---|
| Gender | 0 | 1 | 0.49 | 0.5 |
| Age (Year) | 11 | 16 | 12.74 | 0.901 |
| Height (m) | 130 | 182 | 152.96 | 8.43 |
| Weight (kg) | 21 | 107.6 | 46.58 | 12.23 |
| BMI (kg/m2) | 13.5 | 43.7 | 19.80 | 3.97 |
| Curl-up | 0 | 26 | 11.87 | 5.07 |
| Push-up | 0 | 30 | 8.87 | 7.07 |
| Body Fat | 9.6 | 41.7 | 23.01 | 6.52 |
| Sit & Reach | 0 | 265 | 17.53 | 14.74 |

Table 1.6. Statistical Information of the Hand-Grip (Right) Test

| Variable | Min. Value | Max. Value | Mean Value | Standard Deviation |
|---|---|---|---|---|
| Gender | 0 | 1 | 0.49 | 0.5 |
| Age (Year) | 11 | 16 | 12.75 | 0.90 |
| Height (m) | 130 | 182 | 152.99 | 8.44 |
| Weight (kg) | 21 | 107.6 | 46.63 | 12.23 |
| BMI (kg/m2) | 13.5 | 43.7 | 19.81 | 3.97 |
| Curl-up | 0 | 26 | 11.86 | 5.07 |
| Push-up | 0 | 30 | 8.83 | 7.05 |
| Body Fat | 9.6 | 41.7 | 23.02 | 6.53 |
| Hand-Grip(cm) | 10.1 | 43.1 | 22.54 | 5.43 |

Table 1.7. Statistical Information of the Hand-Grip (Left) Test

| Variable | Min. Value | Max. Value | Mean Value | Standard Deviation |
|---|---|---|---|---|
| Gender | 0 | 1 | 0.49 | 0.5 |
| Age (Year) | 11 | 16 | 12.75 | 0.90 |
| Height (m) | 130 | 182 | 152.99 | 8.44 |
| Weight (kg) | 21 | 107.6 | 46.73 | 12.70 |
| BMI (kg/m2) | 13.5 | 43.7 | 19.81 | 3.97 |
| Curl-up | 0 | 26 | 11.86 | 5.07 |
| Push-up | 0 | 30 | 8.83 | 7.05 |
| Body Fat | 6.53 | 41.7 | 23.03 | 6.65 |
| Hand-Grip (cm) | 10.2 | 40.6 | 21.94 | 5.32 |

Table 1.8. Statistical Information of the Flexed Arm Hang Test

| Variable | Min. Value | Max. Value | Mean Value | Standard Deviation |
|---|---|---|---|---|
| Gender | 0 | 1 | 0.49 | 0.5 |
| Age (Year) | 11 | 16 | 12.76 | 0.908 |
| Height (m) | 130 | 182 | 153.09 | 8.48 |
| Weight (kg) | 5.9 | 107.6 | 46.45 | 12.40 |
| BMI (kg/m2) | 14 | 43.7 | 19.78 | 3.96 |
| Curl-up | 0 | 26 | 11.88 | 4.99 |
| Push-up | 0 | 30 | 8.98 | 6.98 |
| Body Fat | 9.6 | 41.7 | 22.95 | 6.51 |
| Flexed Arm | 0 | 71.15 | 10.89 | 12.91 |

Table 1.9. Statistical Information of the Balance Test

| Variable | Min. Value | Max. Value | Mean Value | Standard Deviation |
|---|---|---|---|---|
| Gender | 0 | 1 | 0.49 | 0.5 |
| Age (Year) | 11 | 16 | 12.75 | 0.9 |
| Height (m) | 130 | 182 | 153.1 | 8.56 |
| Weight (kg) | 5.9 | 107.6 | 46.44 | 12.48 |
| BMI (kg/m2) | 14 | 43.7 | 19.77 | 3.96 |
| Curl-up | 0 | 26 | 11.82 | 5.08 |
| Push-up | 0 | 30 | 9 | 7 |
| Body Fat | 9.6 | 41.7 | 22.94 | 6.53 |
| Balance(s) | 0 | 73.07 | 6.33 | 7.02 |

Table 1.10. Statistical Information of the Agility Test

| Variable | Min. Value | Max. Value | Mean Value | Standard Deviation |
|---|---|---|---|---|
| Gender | 0 | 1 | 0.49 | 0.5 |
| Age (Year) | 11 | 16 | 12.75 | 0.9 |
| Height (m) | 130 | 182 | 153.02 | 8.45 |
| Weight (kg) | 21 | 107.6 | 46.67 | 12.25 |
| BMI (kg/m2) | 13.5 | 43.7 | 19.82 | 3.98 |
| Curl-up | 0 | 26 | 11.91 | 5.04 |
| Push-up | 0 | 30 | 8.88 | 7.04 |
| Body Fat | 9.6 | 41.7 | 23.04 | 6.54 |
| Agility | 6.64 | 22.77 | 15.40 | 1.95 |

## 2. OVERVIEW OF METHODS

According to the points mentioned above, in this thesis, SVM, GRNN, RBFNN and SDT have been used to predict the physical fitness rates.

### 2.1. Support Vector Machines

SVM method is classified as a supervised machine-learning algorithm and invented by Vladimir N. Vapnik and Alexey Chervonenkis in 1963 (Vapnik, 2000). SVM is a state-of-the-art regression method, which is widely utilized in many application areas due to its high accuracy (Kohl et al., 1988). In order to perform a regression analysis, SVM constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space. The values of cost (C), epsilon (ε) and gamma (γ) determine the performance of SVM based models (Wang, 2005). SVM has the most usage compared to the other machine-learning methods in different fields. Generally, the SVM-based models are declared implicitly to be superior to other machine-learning methods especially in the sports area (Akay et al (2016), Abut et al (2015)). Because of these pros, the SVM-based model has been chosen in this study as the main regression method for the assessment of results based on prediction of physical fitness.

#### 2.1.1. SVM-based Model (Linear Approach)

A training set of N considered as data points, $S = \{x_k, y_k\}$ that $x_k \in R^n$ is defined as an input vector and $y_k \in R$ is an output vector. The SVM-based model problems are related with hyperplanes which separate the data and SVM-based model outputs an optimal hyperplane which can be categorized as a new example. The equation of the hyperplanes are given using a vector $w$ and a bias $b$ define the equations of hyperplanes and $w^t - x + b = 0$ illustrates the decision function. An optimal hyperplane is created to maximize the margin of separation $\rho$ . The SVM-based model technique aims demonstrates a classifier in the equation below.

$$f(x) = sign(w^t.x + b) \tag{2.1.}$$

The *w* and *b are limited parameters with* given equation in Equation 2.2:

$$\min_i |w.x_i + b| \geq 1. \tag{2.2.}$$

After the division of vectors, in case the distance between the nearest vector and the hyperplane is maximum, then it needs to be divided by a hyperplane. Therefore, in order to be in standard form, a division hyperplane with a concept form must satisfy the constraints given below in Equation 2.3.

$$y_i\left(w.x_i + b\right) \geq 1, \; i = 1, 2, ..., n. \tag{2.3.}$$

As an input point $x_i$ having the distance *d* from the hyperplane $(\mathbf{w}, b)$ is,

$$d((w,b), xi) = \frac{y_i(x_i.w + b)}{\|w\|} \geq \frac{1}{\|w\|} \tag{2.4.}$$

Form of the $\rho$ can be shown like below,

$$\rho = \frac{2}{\|w\|} \tag{2.5.}$$

Therefore, SVM-based model searches for a separating hyperplane by minimizing

$$\Phi(w) = \frac{1}{2}(w.w) \tag{2.6.}$$

As it is illustrated above, $\Phi(w)$ can be minimized through the structural risk minimization by performing principle,

$$\|w\|^2 \leq c. \tag{2.7.}$$

The standard hyperplanes can be defined as *h*, in space that have n-dimension by limitation in the equation 2.8.

$$h \leq \min\left[\left(R^2 c\right), d\right] + 1 \tag{2.8.}$$

in which R represents a radius of a hypersphere surrounded by training vectors. As a result of this, minimizing equation principle in Equation 2.6 shows the balance for minimization of the upper bound.

The limitations that demonstrated in Equation 2.3 can be reduced by illustration of slack variables

$\xi_i \geq 0,\ i = 1, 2, ..., n$; hence (2.3) can be rewritten as

$$y_i.(w.x_i + b) \geq 1 - \xi_i,\ i = 1, 2, ..., n. \tag{2.9.}$$

In concepts of this situation, the optimization problem occurs as defined below Equation 2.10.

$$\Phi(w,\xi)=\frac{1}{2}(w.w)+C\sum_{i=1}^{n}\xi_i \, . \tag{2.10.}$$

A specified fixed constant which called C as user positive constant, shown in Equation 2.10 and the point of Lagrangian function is utilized in the solution of the problem indicated.

$$L(w,b,\alpha,\xi,\gamma)=\frac{1}{2}(w.w)+C\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\alpha_i\left|y_i(w.x_i+b)-1+\xi_i\right|-\sum_{i=1}^{n}\gamma_i\xi_i \tag{2.11.}$$

Lagrange multipliers are $\alpha_i \geq 0,\ \xi_i \geq 0, i=1,2,...,n$ and shown in Equation 2.11 and must be solved in terms of $\mathbf{w},\ b,$ and $\xi_i$. Classical Lagrangian duality empowers the first issue that turning into its dual problem (Equation 2.11), which renders the solution. In order to solve the dual problem, required formulation shown in Equation 2.12.

$$\max_{\alpha}\left[\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j\left(x_i.x_j\right)\right] \tag{2.12.}$$

Illustrated constraints,

$$\sum_{i=1}^{n}\alpha_i y_i=0,\ 0\leq\alpha_i\leq C,\ i=1,2,...,n. \tag{2.13.}$$

There is a unique solution for this classic quadratic optimization problem as demonstrated below with respect of the Kuhn-Tucker theorem in Equation 2.16.

$$\alpha_i \left[ y_i \left( w.x_i + b \right) - 1 \right] = 0, \; i = 1, 2, ..., n. \quad (2.14.)$$

If $\mathbf{x}_i$ satisfies (Equation 2.15), then (Equation 2.14) will have non-zero Lagrange multipliers

$$y_i \left( \mathbf{w}.\mathbf{x}_i + b \right) = 1. \quad (2.15.)$$

The samples in Equation 2.15 are named support vectors (SVs). The SVs determine the hyperplane as a small subset of the training dataset. Because of this, if the optimal solution $\alpha_i^*$ does not take the value of zero, the classifier function can be formulated as Equation 2.16.

$$f\left(x\right) = \text{sgn}\left\{ \sum_{i=1}^{n} \alpha_i^* y_i \left( x_i.x \right) + b^* \right\} \quad (2.16.)$$

In Equation 2.16 $b^*$ is the solution of (14) for any non-zero $\alpha_i^*$.

### 2.1.2. The SVM-based Model (Non-linear)

As a common word, division of the datasets accurately is not possible with a linear separating hyperplane. However, the datasets can mapped into a higher dimensional field with a nonlinear dimension, so they can be divided accurately. Therefore, to converts the input vector x with the dimension *d*, the following equation $z = \phi(x)$ defined. With the dimension *d* into a vector $z$ is defined and $\phi()$ is selected parameter is given divisible with a hyperplane that defined $\{\phi(x_i, y_i)\}$ as new training data.

The following function which defines the mapped data points from the input space into a higher dimension is shown as $\varphi(.) : R^n \rightarrow R^{nh}$.

By using the same constraints, the optimal function transforms the equation below,

$$\max_{\alpha}\left[\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K\left(\mathbf{x}_i,\mathbf{x}_j\right)\right] \tag{2.17.}$$

The kernel function is demonstrated in below,

$$K\left(\mathbf{x}_i,\mathbf{x}_j\right) = \left\{\varphi\left(\mathbf{x}_i\right).\varphi\left(\mathbf{x}_j\right)\right\} \tag{2.18.}$$

And the kernel function is shown by equation below,

$$K\left(\mathbf{x}_i,\mathbf{x}_j\right) = \exp\left\{-\frac{\left|\mathbf{x}_i - \mathbf{x}_j\right|^2}{2\sigma^2}\right\} \tag{2.19.}$$

The polynomial kernel function is shown by equation below,

$$K\left(\mathbf{x}_i,\mathbf{x}_j\right) = \left(\mathbf{x}_i.\mathbf{x}_j + 1\right)^q,\ q = 1,2,... \tag{2.20.}$$

the $\sigma$ and q parameters that given above have to be predetermined. And after that, the classifier function can be defined as following,

$$f\left(\mathbf{x}\right)=\operatorname{sgn}\left\{\sum_{i=1}^{n}\alpha_i^* y_i K\left(\mathbf{x}_i,\mathbf{x}\right)+b^*\right\} \tag{2.21.}$$

As shown in the equation above, the $b^*$ parameter is the solution of the classifier function for non-zero $\alpha_i^*$ situation.

## 2.2. Generalized Neural Network

The GRNN prediction model is developed by Nadara (Nadaraya, 1964) and Watson (Watson, 1964) and also known as the Nadara-Watson Estimator or the Kernel Regression Estimator. GRNN is a variation of the radial basis neural networks that does not require an iterative training procedure as back propagation networks. The principal advantages of the GRNN are being a fast learning method and available for application on large training sets. Moreover, the GRNN is particularly advantageous with sparse data in a real-time environment (Hannan et al, 2010). A GRNN has four layers which are input layer, pattern layer, summation layer and output layer. The number of input units in the input layer depends on the total number of the observation parameters. The first layer is connected to the pattern layer and in this layer each neuron presents a training pattern and its output. The pattern layer is connected to the summation layer and the summation layer has two different types of summation, which are the single division unit and summation units. The summation and output layer together perform a normalization of output. In training of network linear activation functions are used in hidden and output layers. As per the basic principle of neural network it needs a training data to train itself. Training data should contain input-output mapping and the output is estimated using weighted average of the outputs of the training dataset, where the weight is calculated using the euclidean distance between the training data and the test data. Minimum and maximum values for sigma, search step and the type of kernel function influences the performance of GRNN based models (Specht, 1991).

An GRNN-based model is illustrated as following (Figure 2.1).

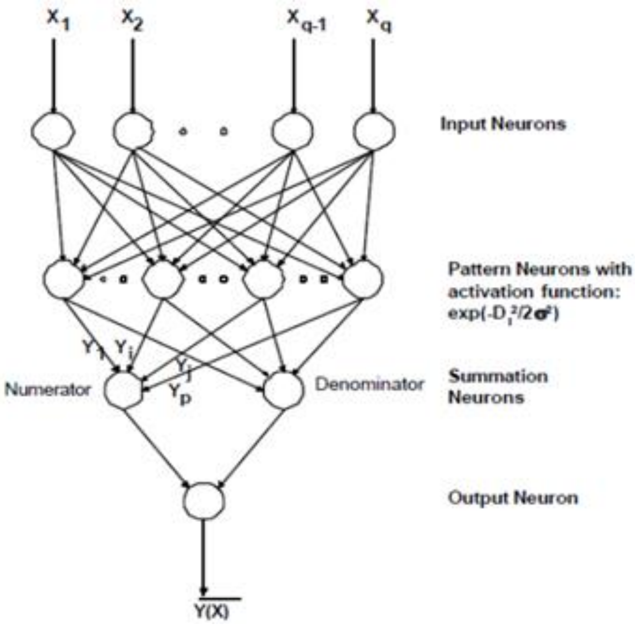


Figure 2.1. Basic Structure of General Regression Neural Network

In the illustrated graph above, the Regression network has an input neuron which feeds the input to the next layer, the pattern layer that calculates the Euclidean distance and activation function. The summation layer that feeds both the numerator & denumerator to the next output layer and the output layer contains one neuron which calculates the output by dividing the numerator part of the Summation layer by the denominator part.

In main principle, GRNN stands on the below equation:

$$Y(x) = \frac{\sum Yi^{e^{-(\frac{di^2}{2\sigma^2})}}}{\sum e^{-(\frac{di^2}{2\sigma^2})}}$$

Where, $di^2 = (x - x_i)^T \, (x - x_i)$

According to the performed integrations, *x* value is used as an input sample and *Xi* illustrates a training sample. The estimate *Y(x)* can be visualized as an average of the observed values, *Yi* represents each observed value according to its Euclidean distance from the *x*. Basically, this activation function theoretically is the weight for that input. Here we have only one unknown parameter which is spread constant $\sigma$. On the other hand, a value of $\sigma$ allows the estimated density for non-Gaussian form so that can be tuned by training process to reach an optimum value where the error will be very small. Training procedure is to find out the optimum value of $\sigma$ as it becomes very large, *Y(x)* assumes the value of the sample mean of the trained data *Yi* associated with the closest to *x* are given heavier weight. Therefore it is necessary to find the $\sigma$ density estimate. When the distribution is not known, the best practice is to find the position where the *MSE* (Mean Squared Error) is minimum. To do that, the training needs to be divided into two parts as a training sample and a test sample. Then, GRNN needs to be applied on the test data based on the training data to find out the *MSE* for various $\sigma$, and detect the minimum *MSE* and corresponding value of $\sigma$.

### 2.3. Radial Basis Function Neural Network

The RBFNN is a special type of artificial neural network and becoming a popular neural network model with diverse patterns. The RBFNN has the feed-forward structure and includes several layers. The first layer is called the input layer, the second is the hidden (kernel) layer with a non-linear RBFNN activation function, and the last layer is called a linear output layer. The nodes within each layer are fully connected to the previous layer and these kinds of models have the general estimation feature (Park and Sandberg, 1991). The input variables are each assigned to the nodes in the input layer and pass directly to the hidden layer without weights. Also, these units are completely connected with the output units as RBF has the feed-forward structure. The number of neurons in the hidden layer, radius of the RBFNN and the regularization parameter lambda (λ) affect the performance of RBFNN-based models (Ghosh-dastidar et al, 2008). In the RBFNN, several hidden nodes with RBFNN activation functions are connected in a feed-forward parallel architecture. The parameters associated with the RBFNNs are optimized during the network training (Hannan et al, 2010). While the data processing period is ongoing, nonlinear transformation represents by the hidden layer for the feature extraction while a linear combinations of output node represents by the output layer. The RBFNN method has fast learning algorithm advantage (Korürek & Doğan, 2010). As stated before, RBFNN method is defined as a structure named feed-forward with a layer which consists of hidden neurons. And these neurons are regulated with the output nodes. Following equation (Equation 2.31) indicates the output units ($\psi_j$) from a linear combination of the main functions.

$$\psi_j(x) = K\left(\frac{\left\| x - c_j \right\|}{\sigma_j^2}\right) \tag{2.22.}$$

A function that demonstrated as $f : R^n \rightarrow R^1$ is estimated by using an RBFNN based method. And an input vector is demonstrated as $x \in R^n$, $\psi(x, c_j, \sigma_j)$ is the jth function with centre $c_j \in R^n$, width $\sigma_j$, and the linear output vector is $w = (w_1, w_2, ..., w_m) \in R^M$. M represents the main number function. The M centers $c_j \in R^n$ are connected to obtain $c = (c_1, c_2, ..., c_m) \in R^{nM}$. Lastly, the widths are connected to obtain $\sigma = (\sigma_1, \sigma_2, ..., \sigma_M) \in R^M$. The output of the network for $x \in R^n$ and $\sigma \in R^M$ is shown below in equation 2.23.

$$F(x, c, \sigma, \omega) = \sum_{j=1}^{M} (\omega_j . \psi(x, c_j, \sigma_j)) \qquad (2.23.)$$

It is assumed that $y = (y_1, y_2, ..., y_n)$ is the weighted output vector and $(x_i, y_i) : i = 1, 2, ..., N$ is a training series. For each parameters $c \in R^{nM}$, $\omega \in R^M, \sigma \in R^M$ and for random weights $\lambda_i, i = 1, 2, ..., N$, that taken as positive numbers to point out importance of definite domains of the input space, set (Equation 2.23) (Wright et al, 2013).

$$E(c, \sigma, \omega) = \frac{1}{2} \sum_{i=1}^{N} \left[ \lambda_i (y_i - F(x_i, c, \sigma, \omega)) \right]^2 \qquad (2.24.)$$

**2.4. Single Decision Tree Method**

The single decision tree based mechanism can be illustrated as a graphical approach of the solutions to make a possible decision according to precise conditions. Decision tree learning uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value, known as decision leaves. In general, the instance space is divided into a single variable and the value of the predictor variable accordingly.

The range of the proper target variable stated by linked leaf from one class to another or the numeric predictor variables holds the linked to each leaf by showing a probability of vectors to have an exact value. Observations with SDT in various domains show that it performs well even with large enough samples (Lopez-Vallverdu et al, 2012). Also, due to decision tree mechanism has less complexity, it's easier to comprehend results with high accuracy. Because of these advantages, decision trees are preferred in most of the cases. The complexity is generally determined by certain parameters including number of samples, the total number of nodes and leaves. Therefore, the main goal of this method is to create a model that predicts the value of a target variable based on several input variables with high accuracy and less complexity.

The decision tree is a flowchart -like tree structure where each internal node represents a test on an attribute and each branch represents- an outcome of the model while leaf node denotes class distributions. As mentioned above, the top node in the tree is named the root node. Leaf nodes give a classification, that applies to all instances to reach the leaf or a probability distribution over all possible classification. To classify an uncertain instance, it is routed down the tree according to the values of the attributes in successive nodes, and when a leaf is reached the instance is classified according to the class assigned to the leaf (Hans and Kamber, 2006). It can be concluded that, in the decision tree construction process, the important thing to consider is the selection of the attribute used for splitting the example set into different classes.

Figure 2.2 is an illustration of the decision tree mechanism, applied in real life. There is a chart that illustrates options for weekend plans, but there are a few unknown factors that will determine that plan.

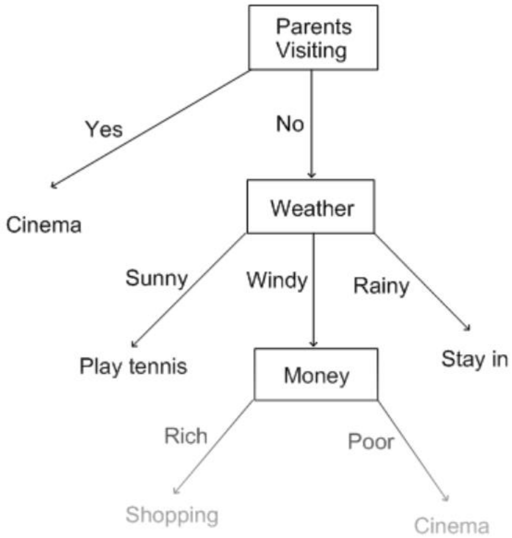

Figure 2. 2. Decision tree example

Decision trees can also be used to solve multi-class classification problems. So, there could be infinite boundaries for a continuous variable and continuous features can be turned to the categorical variables (i.e. lesser than or greater than a certain value) before a split at the root node the choice is made depending on which boundary will result in the most information gain (Chandra, 2011).

# 3. DEVELOPMENT OF PREDICTION MODELS

In this thesis, the physical prediction rates were calculated by using SVM, GRNN, RBFNN and SDT machine-learning methods. Then, based on these prediction rates, several models were developed according to prediction variables by using RMSEs results. The models were evaluated by using one of the validation features named 10-fold cross validation.

## 3.1. Support Vector Machine Prediction Models

The accuracy of SVM prediction models depend on the value of C, the type of kernel function and the parameters of this function. In terms of accuracy, to have an effective SVM-based model, the cost parameter C, the ε-insensitive loss function and gamma parameters require to have their values in the optimal range (Chuang et al, 2011). The kernel parameters are important due to their role as a bridge between the input space and the high-dimensional feature space (Wang et al, 2007). The C parameter acts as the tradeoff for the ratio of non-separable samples. On the other hand, the C parameter examines the tradeoff between the margin of the maximization and SVM errors of training data. While a large value of C exhibits similar hard-margin SVM, a small value of C makes the decision surface regular by increasing the quantity of errors. However, misclassification cost is demonstrated when the C value is large. To measure estimation of quality, the Vapnik function (Vapnik, 2000) is the general operated function. To be able to control the width range of the Vapnik function, also named as ε-insensitive zone, it is required to fit the training data and exploit the parameter of ε. A large value of ε implies a small number of dedicated support vectors due to the number of selected support vectors are biased by the ε.

The aim is to achieve optimal values of the parameters (C and gamma) so that the training data can be estimated with a minimum error by the prediction model. There are many methods that can be used for deciding the values of the optimal parameters in order to minimize the error. Several techniques like grid search techniques and swarm optimization methods are used to evaluate these values in the optimal range. But, rather than swarm optimization technique, the grid search method is the most effective one for achieving the optimum values for C and gamma parameters for medium-size dataset. In this method, the specific parameters are differentiated by constant value through an interval of values, and the success of each value is evaluated. Also, the parameter values in grid search are altered within the tree parameters and a specific interval indicating that the minimum error rates are detained. The cross validation option within the grid search can be utilized in order to develop the generalization capability of the SVM model. In the process of n-fold cross validation, the cross validation is recurred n times with each of the n subsets utilized precisely as the validation data. The n results from the folds ought to be associated to generate a single prediction. To build the prediction models in this study, grid search employing 10-fold cross validation was utilized to indicate the optimal values.

The demonstrated flow-chart of the SVM-based models are given in Figure 3.1. In the beginning, the standardized dataset is developed to have zero mean and unity variance by using the ranker search method and score of each attribute was sorted in a descending order. The dataset was reduced by one feature once at a time as the attribute with the lowest score was removed before being passing on to the SVM. Throughout these processes creating new train and train subsets can be deployed. As mentioned, through the validation of models and increasing of the reliability of results, a 10-fold cross validation has been operated on the dataset, which in turn yielded the computation of *RMSEs* for each prediction.

Table 3.1. List of the utilized parameters for the SVM model

| **Utilized Parameter** | **Value of the Prameter** |
|---|---|
| Cost parameter for the SVM model (*C*) | [$2^{-6}$ - $2^{16}$] |
| Epsilon parameter for the SVM model(**ε**) | [0.01-150] |
| Gamma parameter for the SVM model (*γ*) | [$2^{-10}$- $2^{8}$] |
| Kernel Function of the model | RBFN |

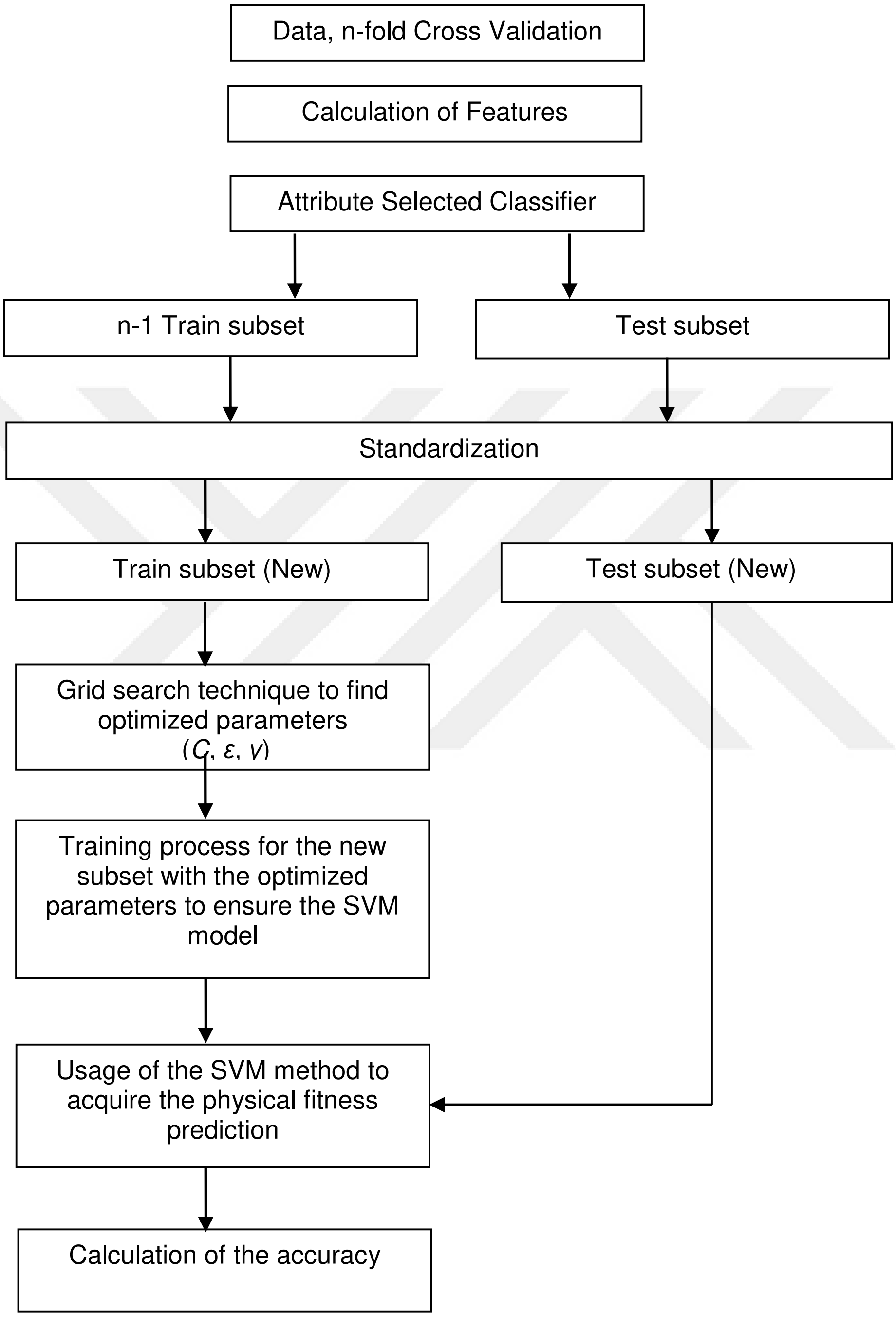


Figure 3.1. Flow-chart illustration for SVM-based model

### 3.2. GRNN-based Prediction Models

The main concept of this model depends on the nonparametric model where the prediction for an input data named x is given by the average of the target outputs of the training data (Hardle, 1990). The average can be generalized by weighting the nodes in order to calculate their distance from input data called x by using the same kernel function.

In general, the GRNN model is applicable for prediction and regression analysis and it may be considered as a useful tool to discover the relationship between dependent and independent variables for unknown and complex structures. Furthermore, GRNN models support linear and non-linear relationships. GRNNs need supervised training and the selection methodologies to be able to find the optimal value for the σ parameter to obtain reliable results (Specht, 1991).

The structure of the model is identical for one variant to the others and has an advantage for the fast learning in order to obtain optimal regression for the samples that have very large data. Particularly, the GRNN is advantageous with sparse data in real-time environments as the regression surface is instantly defined in whole structure on all samples. On the other hand, the disadvantage of the GRNN model is required a large amount of computation in order to estimate a new output vector. To overcome this disadvantage, the clustering method can be considered (Cherkassky, 1990).

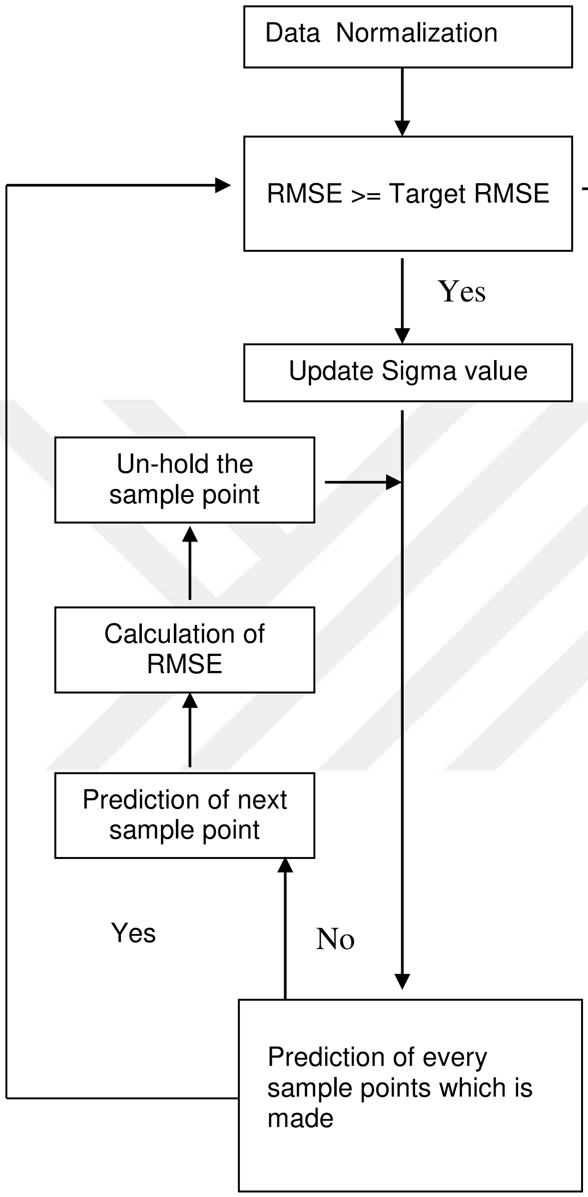


Figure 3.2. Flow chart of GRNN based prediction models

### 3.3. RBFN Networks Prediction Models

Performance of RBFN networks depend on numerous factors, but their success especially relies on the choice of the network parameters. An RBFN comprises of three layers concept as an input layer, a hidden layer part and an output layer part. Neuron number of the hidden layer must be specified before the parameter selection phase. The Gaussian transfer function that has outputs of the hidden layer is inversely correlated to the distance from the center of the neuron to employ.

In order to develop a powerful RBFNN model, various points need to be considered. The first step is selecting the RBFN centers randomly from the training data or using uncontrolled techniques for selecting the RBFN centers. The following step is reading the content of the dataset, and standardizing the variables in order to add a new neuron and adjust the output layer weight in the network. Calculation of the network's output error needs to be implemented at the end and if the error is too large, the calculation of the error needs to be reproduced. In case an acceptable error occurs, the performance of the RBFN networks for the training and test data is determined. However, this method mostly leads to over-fitting with a negative effect on the achievement of trained neural networks. The model parameters include the population size, radius of the model, the neuron's maximum number and the regularization parameter (λ) connected to the performance of an RBFNN. Important parameters of the RBFNN prediction models and their values are given in Table 3.2.

Table 3.2. List of the utilized parameter rates for RBFNN-based models

| **Parameters** | **Rates** |
|---|---|
| Regularization parameter of RBFNN (λ) | [0.001 - 25] |
| Model's Population size | [100 - 400] |
| Model's Radius value | [0.001- 450] |
| Number of neurons (Max) | [95-100] |

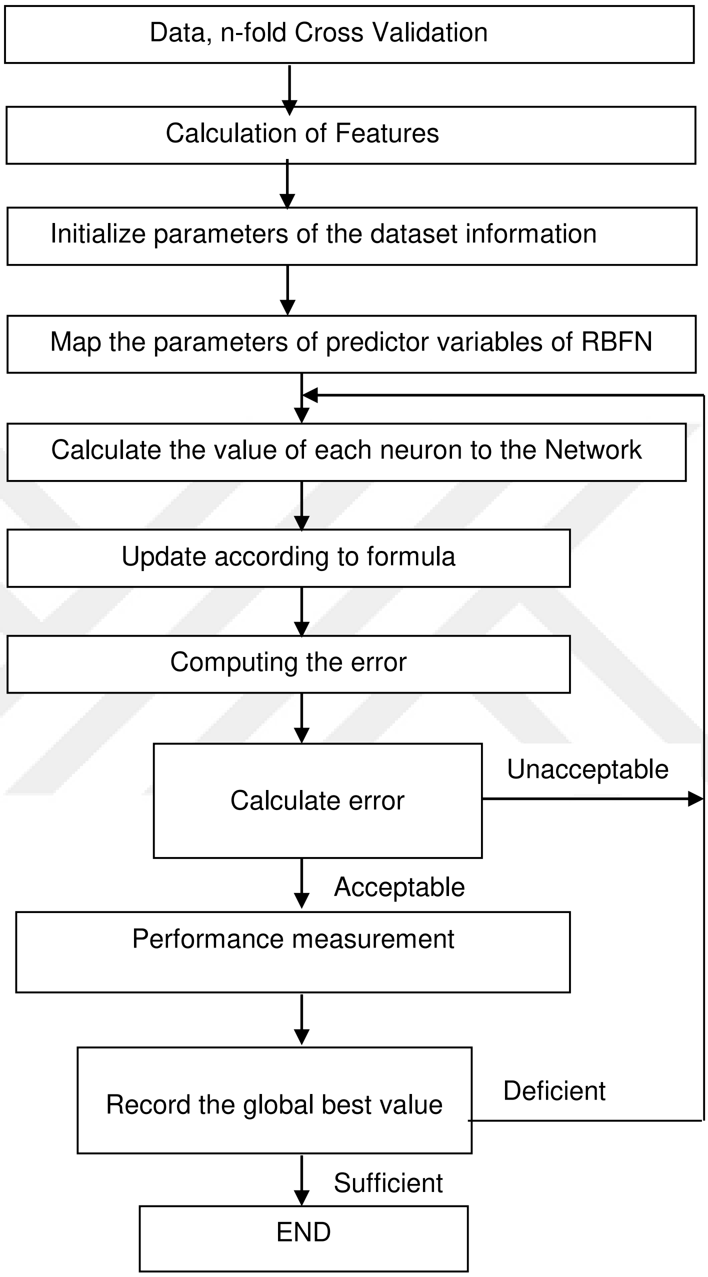


Figure 3.3. Flow-chart for RBFNN models

### 3.4. Single Decision Tree Model

The least rows in a node, least split node size and the highest three level parameters are important factors for SDT models. Because of this effect, the performance of the models is relative to these parameters. The least rows in a node specify the least rows number to fall in a node after splitting. The minimum split node size defines a threshold value for a node to be split. The highest tree levels refer to the maximum possible number of developed levels when a tree mechanism is built. To develop efficient SDT models, the mentioned structural parameters must have their optimal values.

The list of the utilized parameter rates to construct efficient model mechanism is given in Table 3.3.

Table 3.3. Utilized parameters rates for SDT model

| **Parameters** | **Rates of the Parameter** |
|---|---|
| Minimum split node size for the SDT Model | [2 - 15] |
| Minimum rows in a node for the SDT Model | [2 - 25] |
| Maximum tree levels for the SDT model | [5- 10] |

# 4. RESULTS AND DISCUSSIONS

## 4.1. Study Approach

The combinations of the predictor variables were developed in eight categories and the first developed category includes gender, age, height and weight. The second, third and fourth categories include the predictor variables (curl-up, push-up and body fat) were used with the first category. For the rest of the categories, the predictor variables were used in a specific order. To appreciate the performance efficiency for different models, the *RMSE* and *R* were applied through the validation method called 10-fold cross validation. The equation of *RMSE* is given in equation 4.1.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y - Y')^2} \tag{4.1}$$

In (4.1), Y is the value that has been measured, Y' is the value that has been predicted and *n* is the number of samples where in a test subset.

Table 4.1. Combination of the predictor variables

| Model No | Predictor Variables |
|---|---|
| 1 | Gender, Age, Weight, Height |
| 2 | Gender, Age, Weight, Height, Curl-up |
| 3 | Gender, Age, Weight, Height, Push-up |
| 4 | Gender, Age, Weight, Height, Body Fat |
| 5 | Gender, Age, Weight, Height, Curl-up, Push-up |
| 6 | Gender, Age, Weight, Height, Curl-up, Body Fat |
| 7 | Gender, Age, Weight, Height, Push-up, Body Fat |
| 8 | Gender, Age, Weight, Height, Curl-up, Push-up, Body Fat |

## 4.2. Results of the Study

Demonstrations of the test results are given with tables in following.

Table 4.2. RMSEs for the 20m Shuttle Run Test with prediction models

| Variables | SVM | GRNN | RBF | SDT |
|---|---|---|---|---|
| Gender, Age, Weight, Height | 2.119 | 2.124 | 2.177 | 2.323 |
| Gender, Age, Weight, Height, Curl-up | 2.067 | 2.088 | 2.136 | 2.325 |
| Gender, Age, Weight, Height, Push-up | 2.071 | 2.091 | 2.174 | 2.164 |
| Gender, Age, Weight, Height, Body Fat | 2.066 | 2.094 | 2.222 | 2.098 |
| Gender, Age, Weight, Height, Curl-up, Push-up | 2.045 | 2.048 | 2.145 | 2.193 |
| Gender, Age, Weight, Height, Curl-up, Body Fat | 2.041 | 2.053 | 2.141 | 2.098 |
| Gender, Age, Weight, Height, Push-up, Body Fat | 2.054 | 2.067 | 2.188 | 2.098 |
| Gender, Age, Weight, Height, Curl-up, Push-up, Body Fat | 2.036 | 2.044 | 2.221 | 2.098 |

Table 4.3. RMSEs for the 30m Speed Test with prediction models

| Variables | SVM | GRNN | RBF | SDT |
|---|---|---|---|---|
| Gender, Age, Weight, Height | 0.516 | 0.531 | 0.558 | 0.554 |
| Gender, Age, Weight, Height, Curl-up | 0.470 | 0.485 | 0.532 | 0.496 |
| Gender, Age, Weight, Height, Push-up | 0.493 | 0.494 | 0.502 | 0.511 |
| Gender, Age, Weight, Height, Body Fat | 0.486 | 0.490 | 0.517 | 0.516 |
| Gender, Age, Weight, Height, Curl-up, Push-up | 0.449 | 0.459 | 0.478 | 0.488 |
| Gender, Age, Weight, Height, Curl-up, Body Fat | 0.449 | 0.454 | 0.487 | 0.516 |
| Gender, Age, Weight, Height, Push-up, Body Fat | 0.465 | 0.472 | 0.580 | 0.503 |
| Gender, Age, Weight, Height, Curl-up, Push-up, Body Fat | 0.436 | 0.441 | 0.497 | 0.476 |

Table 4.4. RMSEs for the Standing Long Jump Test with prediction models

| Variables | SVM | GRNN | RBF | SDT |
|---|---|---|---|---|
| Gender, Age, Weight, Height | 19.939 | 20.873 | 23.858 | 20.741 |
| Gender, Age, Weight, Height, Curl-up | 18.740 | 19.878 | 21.167 | 19.565 |
| Gender, Age, Weight, Height, Push-up | 19.289 | 20.144 | 36.790 | 20.594 |
| Gender, Age, Weight, Height, Body Fat | 19.481 | 20.331 | 20.216 | 20.386 |
| Gender, Age, Weight, Height, Curl-up, Push-up | 18.746 | 18.747 | 20.753 | 20.275 |
| Gender, Age, Weight, Height, Curl-up, Body Fat | 18.524 | 18.705 | 19.705 | 20.633 |
| Gender, Age, Weight, Height, Push-up, Body Fat | 18.741 | 19.750 | 22.200 | 20.171 |
| Gender, Age, Weight, Height, Curl-up, Push-up, Body Fat | 18.409 | 18.484 | 23.222 | 20.150 |

Table 4.5. RMSEs for the Sit & Reach Test with prediction model

| Variables | SVM | GRNN | RBF | SDT |
|---|---|---|---|---|
| Gender, Age, Weight, Height | 14.689 | 14.684 | 14.892 | 15.034 |
| Gender, Age, Weight, Height, Curl-up | 14.683 | 14.684 | 14.978 | 16.150 |
| Gender, Age, Weight, Height, Push-up | 14.542 | 14.301 | 15.102 | 16.009 |
| Gender, Age, Weight, Height, Body Fat | 14.735 | 14.686 | 15.024 | 15.377 |
| Gender, Age, Weight, Height, Curl-up, Push-up | 14.588 | 14.395 | 15.158 | 15.962 |
| Gender, Age, Weight, Height, Curl-up, Body Fat | 14.690 | 14.686 | 15.266 | 15.763 |
| Gender, Age, Weight, Height, Push-up, Body Fat | 14.598 | 14.289 | 15.053 | 16.734 |
| Gender, Age, Weight, Height, Curl-up, Push-up, Body Fat | 14.595 | 14.286 | 15.071 | 16.759 |

Table 4.6. RMSEs for the Hand-Grip (Left) Test with prediction models

| Variables | SVM | GRNN | RBF | SDT |
|---|---|---|---|---|
| Gender, Age, Weight, Height | 3.723 | 3.860 | 3.800 | 4.026 |
| Gender, Age, Weight, Height, Curl-up | 3.638 | 3.719 | 3.919 | 4.074 |
| Gender, Age, Weight, Height, Push-up | 3.511 | 3.642 | 3.780 | 3.987 |
| Gender, Age, Weight, Height, Body Fat | 3.648 | 3.691 | 3.762 | 4.039 |
| Gender, Age, Weight, Height, Curl-up, Push-up | 3.478 | 3.570 | 3.627 | 4.115 |
| Gender, Age, Weight, Height, Curl-up, Body Fat | 3.564 | 3.548 | 3.778 | 4.129 |
| Gender, Age, Weight, Height, Push-up, Body Fat | 3.471 | 3.606 | 3.776 | 3.947 |
| Gender, Age, Weight, Height, Curl-up, Push-up, Body Fat | 3.443 | 3.515 | 3.991 | 4.091 |

Table 4.7. RMSEs for the Hand-Grip (Right) Test with prediction models

| Variables | SVM | GRNN | RBF | SDT |
|---|---|---|---|---|
| Gender, Age, Weight, Height | 4.199 | 4.291 | 4.242 | 4.567 |
| Gender, Age, Weight, Height, Curl-up | 4.132 | 4.198 | 4.433 | 4.558 |
| Gender, Age, Weight, Height, Push-up | 4.076 | 4.119 | 4.316 | 4.448 |
| Gender, Age, Weight, Height, Body Fat | 4.186 | 4.259 | 4.308 | 4.528 |
| Gender, Age, Weight, Height, Curl-up, Push-up | 4.045 | 4.076 | 4.225 | 4.409 |
| Gender, Age, Weight, Height, Curl-up, Body Fat | 4.112 | 4.144 | 4.188 | 4.470 |
| Gender, Age, Weight, Height, Push-up, Body Fat | 4.057 | 4.116 | 4.116 | 4.392 |
| Gender, Age, Weight, Height, Curl-up, Push-up, Body Fat | 4.007 | 4.066 | 4.375 | 4.436 |

Table 4.8. RMSEs for the Flexed Arm-Hang Test with prediction models

| Variables | SVM | GRNN | RBF | SDT |
|---|---|---|---|---|
| Gender, Age, Weight, Height | 11.103 | 10.892 | 12.187 | 11.755 |
| Gender, Age, Weight, Height, Curl-up | 10.541 | 10.417 | 11.991 | 11.576 |
| Gender, Age, Weight, Height, Push-up | 10.456 | 10.268 | 15.246 | 11.681 |
| Gender, Age, Weight, Height, Body Fat | 10.018 | 9.794 | 11.552 | 10.035 |
| Gender, Age, Weight, Height, Curl-up, Push-up | 9.859 | 9.885 | 12.731 | 11.301 |
| Gender, Age, Weight, Height, Curl-up, Body Fat | 9.730 | 9.580 | 12.108 | 10.035 |
| Gender, Age, Weight, Height, Push-up, Body Fat | 9.692 | 9.314 | 10.671 | 10.163 |
| Gender, Age, Weight, Height, Curl-up, Push-up, Body Fat | 9.444 | 9.078 | 10.516 | 10.213 |

Table 4.9. RMSEs for the Balance Test with prediction models

| Variables | SVM | GRNN | RBF | SDT |
|---|---|---|---|---|
| Gender, Age, Weight, Height | 6.653 | 6.632 | 7.410 | 7.198 |
| Gender, Age, Weight, Height, Curl-up | 6.607 | 6.601 | 6.998 | 7.198 |
| Gender, Age, Weight, Height, Push-up | 6.620 | 6.575 | 7.100 | 6.970 |
| Gender, Age, Weight, Height, Body Fat | 6.704 | 6.620 | 6.948 | 7.356 |
| Gender, Age, Weight, Height, Curl-up, Push-up | 6.621 | 6.560 | 7.651 | 6.721 |
| Gender, Age, Weight, Height, Curl-up, Body Fat | 6.677 | 6.604 | 7.439 | 7.362 |
| Gender, Age, Weight, Height, Push-up, Body Fat | 6.620 | 6.539 | 7.444 | 6.792 |
| Gender, Age, Weight, Height, Curl-up, Push-up, Body Fat | 6.630 | 6.532 | 9.696 | 6.744 |

Table 4.10. RMSEs for the Agility Test with prediction models

| Variables | SVM | GRNN | RBF | SDT |
|---|---|---|---|---|
| Gender, Age, Weight, Height | 1.760 | 1.781 | 1.831 | 1.822 |
| Gender, Age, Weight, Height, Curl-up | 1.634 | 1.669 | 2.486 | 1.768 |
| Gender, Age, Weight, Height, Push-up | 1.684 | 1.693 | 1.871 | 1.793 |
| Gender, Age, Weight, Height, Body Fat | 1.740 | 1.742 | 1.873 | 1.789 |
| Gender, Age, Weight, Height, Curl-up, Push-up | 1.588 | 1.628 | 1.690 | 1.784 |
| Gender, Age, Weight, Height, Curl-up, Body Fat | 1.619 | 1.657 | 1.769 | 1.768 |
| Gender, Age, Weight, Height, Push-up, Body Fat | 1.683 | 1.684 | 1.843 | 1.851 |
| Gender, Age, Weight, Height, Curl-up, Push-up, Body Fat | 1.590 | 1.626 | 1.903 | 1.748 |

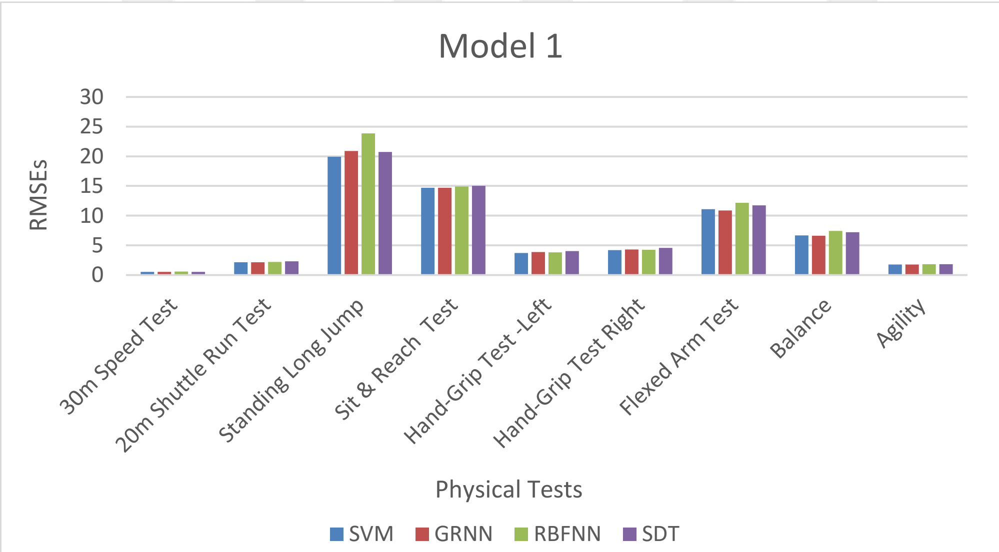


Figure 4.1. RMSEs for physical fitness prediction models without additional predictor variables (Model 1)

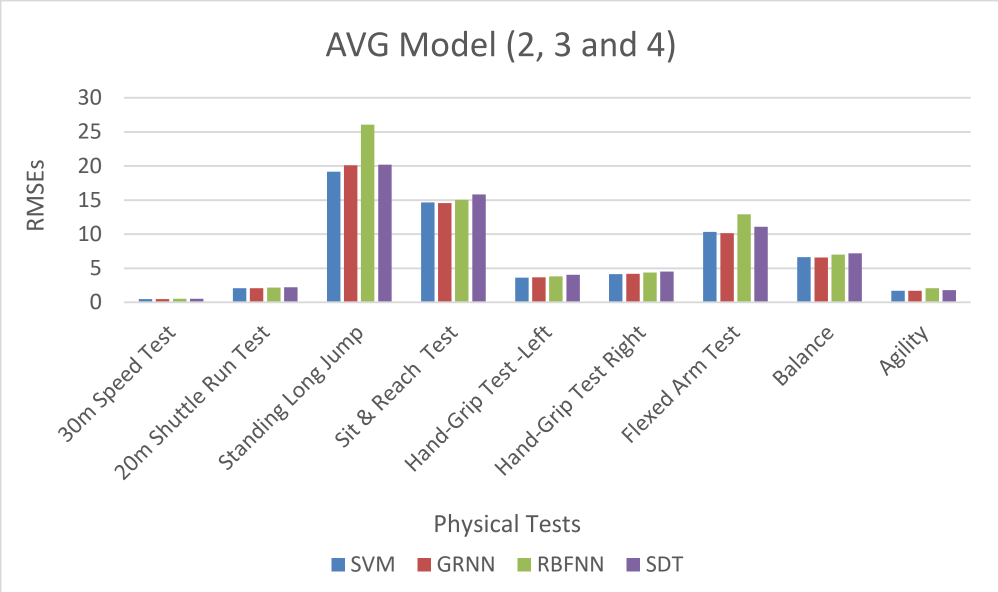


Figure 4.2. *RMSE*s in average for physical fitness prediction models with one additional predictor variables (Average of Model 2, 3 and 4)

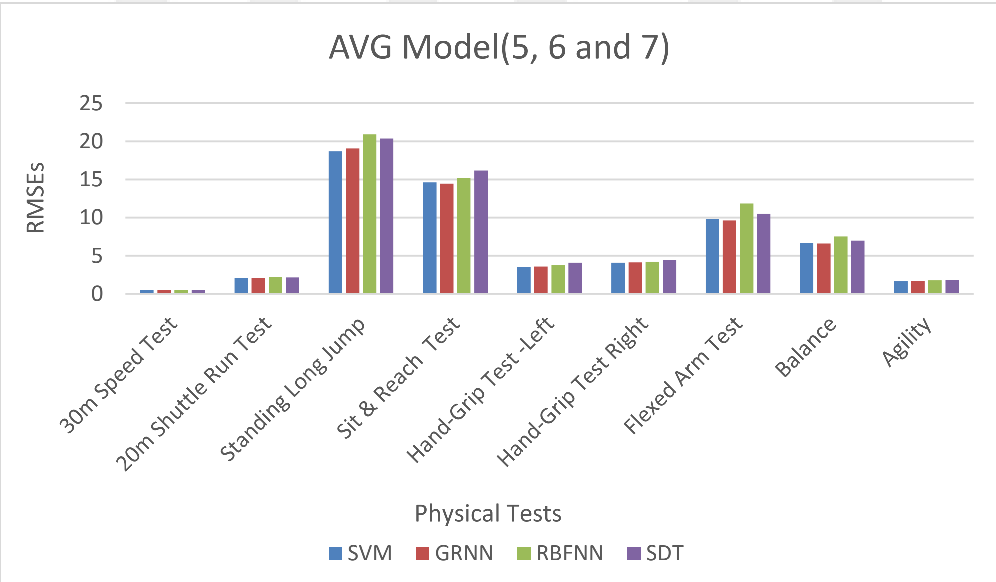


Figure 4.3. *RMSE*s in average for physical fitness prediction models with two additional predictor variables (Average of Model 5, 6 and 7)

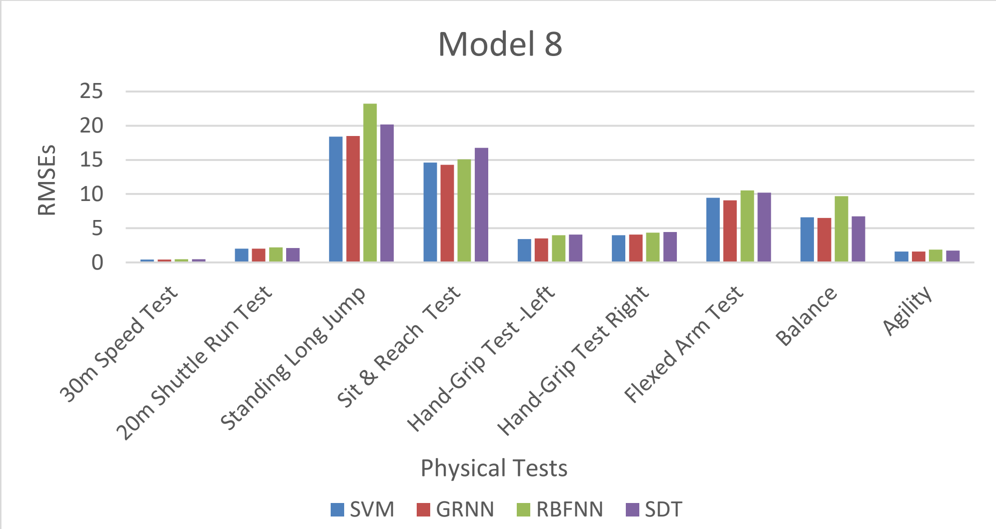


Figure 4.4. RMSEs for physical fitness prediction models with three additional predictor variables (Model 8)

Table 4.11. Physical Assessment results for Physical fitness prediction methods demonstrated below

| Physical Tests | SVM | GRNN | RBFNN | SDT |
|---|---|---|---|---|
| **30m Speed Test** | 0.471 | 0.478 | 0.519 | 0.508 |
| **20m Shuttle Run Test** | 2.063 | 2.076 | 2.176 | 2.175 |
| **Standing Long Jump** | 18.984 | 19.614 | 23.489 | 20.314 |
| **Sit & Reach Test** | 14.640 | 14.501 | 15.068 | 15.974 |
| **Hand-Grip Test - Left** | 3.559 | 3.644 | 3.804 | 4.051 |
| **Hand-Grip Test-Right** | 4.102 | 4.159 | 4.275 | 4.476 |
| **Flexed Arm Test** | 10.105 | 9.904 | 12.125 | 10.845 |
| **Balance** | 6.642 | 6.583 | 7.586 | 7.043 |
| **Agility** | 1.662 | 1.685 | 1.908 | 1.790 |

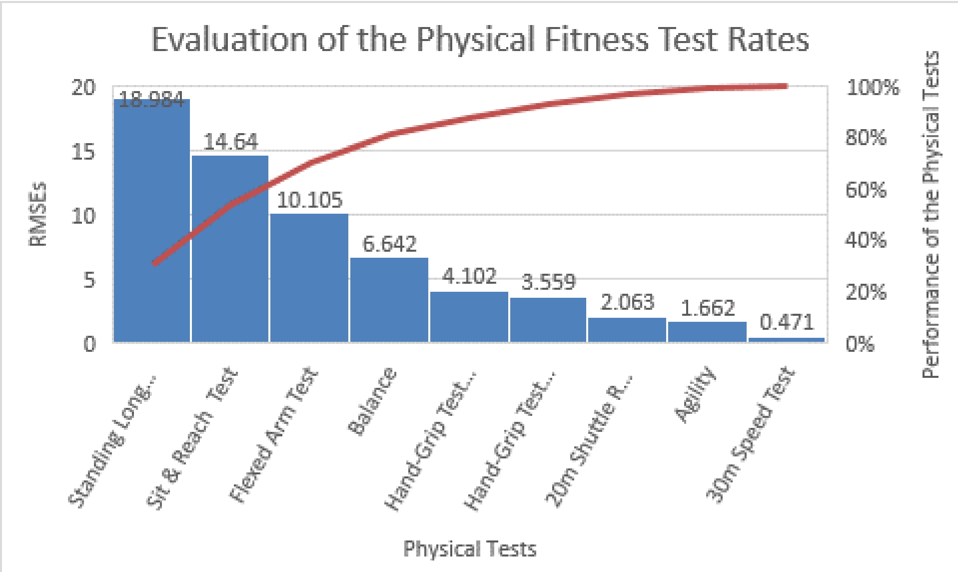


Figure 4.5. Evaluation of the physical fitness assessment rates in general

### 4.3 Discussions on the Study

#### 4.3.1 Overall Discussions via Test Results

The obtained results display that SVM-based and GRNN-based prediction models outperform RBFNN-based and SDT-based models. Moreover, SVM and GRNN models have high prediction rates in terms of accuracy regardless of the utilized prediction category or the physical assessment test applied to the participants. In more details:

- In terms of the performances of machine-learning methods from the lowest *RMSE* to the highest ones can be listed as SVM, GRNN, RBFNN and SDT.
- The SVM-based prediction model without any additional predictor variables including gender, height, weight and age for the physical fitness prediction with all physical assessment test results yields an *RMSE* value of 7.18 which is the lowest *RMSE* among all the

prediction models. On the other hand, the RBFNN-based prediction model for the same situation yields *RMSE* value of 7.88 which is the highest *RMSE* among all prediction models. Therefore, these two prediction models are respectively the best and worst for the prediction model without any additional predictor variables in terms of prediction accuracy.

- Among the categories with one additional variable which are Model 2, 3 and 4 yield average *RMSE* value of 6.97 which is the lowest *RMSE* for the SVM-based methods. Moreover, other prediction method's performances for the same group are the GRNN-based methods yield average *RMSE* value of 7.06, the SDT-based methods yield average *RMSE* value of 7.48 and the RBFNN-based methods yield average *RMSE* value of 8.22 which is the highest *RMSE* among the others.
- Among the categories with two additional variables which are Model 5, 6 and 7 yield average *RMSE* value of 6.82 which is the lowest *RMSE* for the SVM-based methods. Moreover, other prediction method's performances for the same group are the GRNN-based methods yield average *RMSE* value of 6.83, the SDT-based methods yield average *RMSE* value of 7.43 and the RBFNN-based methods yield average *RMSE* value of 7.52 which is the worst performance among the others.
- Among the category with includes all variables (Model 8) yields average *RMSE* value of 6.73 which is the lowest RMSE for the SVM-based methods in all physical test assessments. Additionally, other prediction method's performances for the same group are the GRNN-based methods yield average *RMSE* value of 6.67, the SDT-based methods yield average *RMSE* value of 7.41 and the RBFNN-based

methods yield average *RMSE* value of 7.94 which is the worst performance among the others.

- Based on analysis that we made for the model groups with all physical tests, prediction model with all predictor variables (Model 8) shows the highest performance in terms of prediction accuracy for all samples.
- The comparable *RMSE*s obtained by all prediction models, and according to that analysis, the SVM-based prediction models are considered as the best model by the other methods (GRNN, RBFNN, SDT models on the physical fitness prediction) in terms of the average of *RMSE*.
- The prediction models with variables curl-up, push-up and body fat yield lower *RMSE*s than the ones that do not. Whereas, prediction model with default predictors (Model 1) yields the worst result in all circumstances in terms of the average of *RMSE*.
- Regarding physical fitness test types, the *RMSE*s of the 30m speed test yields the best assessment results due to its lowest error rates (0.49 in average), and its independency on the machine-learning method or the prediction model was utilized.
- On the other hand, the *RMSEs* of the Standing Long Jump test yields the worst results with its highest error rates (20.60 in average).
- The rest of the physical assessment tests results shown in following as descending order in terms of the performances. Respectively; the Agility test gives 1.76 value of *RMSE* in average, the 20m Shuttle Run test gives 2.12 value of *RMSE* in average, the Hand-grip Left test gives 3.76 value of *RMSE* in average, the Hand-grip Right test gives 4.25 value of *RMSE* in average, the Balance test gives 6.96 value of *RMSE* in average, the Flexed Arm test gives 10.95 value of *RMSE* in

average and the Sit & Reach test gives 15.04 value of *RMSE* in average.

### 4.3.2 Discussions for the Different Categories of the Prediction Models

To sum up, throughout the developed models, the ones that include the predictor variables (curl-up, push-up and body fat) lead to the lowest *RMSE*s. On the other hand, the prediction models which include the variables (gender, age, weight and height) lead to the highest *RMSE*s. Also, the category (Model 8) which includes all predictor variables indicates the best performance in general. Whereas, the category (Model 1) of prediction models without all these variables yields the worst performance in terms of *RMSEs*. Furthermore, the obtained results reveal the following:

- The SVM-based model with an average *RMSE* value of physical assessments yields the lowest *RMSE* with 6.91. On the other hand, the RBFNN-based prediction model with in average *RMSE* value of physical assessments yields the highest *RMSE* 7.88. According to the demonstrated results of the prediction models in most of the cases, SVM and RBFNN were selected the best and the worst models respectively, among the others in terms of accuracy rates.
- Both prediction models include curl-ups, push-ups and body fat, and the model without these predictor variables occupy the first and last places respectively in terms of model performances regarding the *RMSEs* in average.
- Among all prediction model categories, the ones which includes all predictor variables outperform all other prediction models. Whereas the categories with the gender, age, height and weight variables yield the worst results in terms of the average of *RMSE*.

- The physical assessment tests used for physical prediction with four machine-learning methods including SVM, GRNN, RBFNN and SDT-based can be represented in various results. The 30m Speed test yields the lowest *RMSE* with 0.49 in average. However, Standing Long Jump test yields the highest *RMSE* with 20.60. The performance of the other physical assessments are respectively, Agility that gave the second lowest *RMSE* with 1.76, the 20m Shuttle Run test with 2.12, the Hand-grip (right/left) test in average with 4.00, the Balance test with 6.96, the Flexed-arm test with 10.95 and Sit & Reach test with 15.04. Therefore, 30m Speed test can be represented as the best assessment test for the physical fitness prediction concept while the Standing Long Jump test can be the worst assessment test with the worst performance.

# 5. CONCLUSION

The machine-learning approach for the physical fitness prediction is relatively new and mainly focused on the educational area as we aimed in this study as well to obtained results are clearly promising. In order to carry out this study, four well-known machine-learning methods, respectively SVM, GRNN, RBFNN and SDT, have been used and aimed to develop effective physical prediction models for school-aged children. These prediction models have been split into various categories based on predictor variables quantity. Gender, age, weight, height, body fat, number of curl-ups and number of push-ups of the samples are considered as the predictor variables used in this study and the dataset taken from Gazi University, department of Physical Education and Sport. Traditional training methods have been used for the physical fitness assessment and validation method named 10-fold cross-validation has been used for the figuring of the errors. Also, the *RMSE* approach has been used to compute the errors.

Based on the results obtained, some provisions can be observed on this study. Firstly, the results disclose that machine-learning methods -especially SVM and GRNN- are efficient to use to estimate the physical fitness prediction with acceptable accuracy. The SVM-based and the GRNN-based prediction models outperform the other regression methods utilized in this study, regardless of the prediction models category used. On the other hand, RBFNN and SDT models exhibit the insufficient performance for the physical prediction with high *RMSE* rates. Moreover, the physical assessments used for this thesis made a significant effect on the prediction of physical fitness. The 30m Speed test yields lower *RMSEs* than the other assessments, regardless of the prediction model category or the machine-learning method used. On the contrary, the Standing Long Jump test yields the worst performance according to *RMSEs* average. The prediction model that includes all variables yields the lowest *RMSEs*. In contrary, the ones that do not include any additional variables (Model 1) exhibit the highest *RMSEs*.

This thesis has showed the efficiency of machine-learning methods to indicate high accuracy of the physical fitness prediction, still there are many opportunities pending improvement and extended results for related studies. Different machine-learning methods and additional datasets that include various predictor variables can improve the accuracy of future results for physical fitness prediction. The rest of the supervised or unsupervised machine-learning methods can be used for detailed evaluation of the physical fitness.

**REFERENCES**

Ahmed M.U., Loutfi A., 2013. Pyhsical Activity Identification using Supervised Machine Learning and based on Pulse Rate. International Journal of Computer Science and Applications, 4(7).

Akay, M. F., Abut, F., Çetin, E., Yarım, İ., & Sow, B., 2015. Data-Driven Modeling of Quadriceps and Hamstring Muscle Strength Using Support Vector Machines. Third International Symposium on Engineering, Artificial Intelligence & Applications, North Cyprus, Nov 4-6, 2-4.

Akay, M. F., Abut, F., Çetin, E., Yarım, İ., & Sow, B., 2016. Support vector machines for predicting the hamstring and quadriceps muscle strength of college-aged athletes. Turkish Journal of Electrical Engineering and Computer Science, DOI:10.3906/elk-1603-304.

Akay, M. F., Bozkurt, Ö., Çetin, E., Yarım, İ., 2018. Multiple Linear Regression Based Physical Fitness Prediction Models for Turkish Secondary School Students, CYICER'08 Proceedings of the 7th Cyprus International Conference on Educational Research ISSN 2421-8030.

Akay, M. F., Çetin, E., Yarım, İ., Bozkurt, Ö., Erdem, S., 2018. Development of Physical Fitness Prediction Models for Turkish Secondary School Students Using Machine Learning Methods, Gümüşhane University Journal of Science and Technology Institute, Nov 7-10. http://dx.doi.org/10.17714/gumusfenbil.435897.

Akay, M. F., Çetin, E., Yarım, İ., Bozkurt, Ö., Erdem, S., 2018. Predicting Physical Fitness Using Support Vector Machines and Blood Test Results

Bailey DA., Martin AD., 1994. Physical activity and skeletal health in adolescents. Pediatric Exercise Science 6:330-47.

Brian Pink, 2015. "Participation in Sport and Physical Recreation, Australia,". Australian Bureau of Statistics. 18 February.

Chandra, B. 2011. Heterogeneous Node Split Measure for Decision Tree Construction, 872–877.

Chin M., K, Girandola R., N., Yang J., Cruz A., Liu Y., K. 2002. The body mass and body composition of Hong Kong school chidren. The 44thIchper-SD World Congress, s. 17, Taipei, Taiwan.

Chuang, C.C., Lee, Z.C., 2011. Hybrid robust Support Vector Machines for regression without liers, Applied Soft Computing.

Dogan, U., Glasmachers, T., & Igel, C., 2016. A unified view on multi-class support vector classification. Journal of Machine Learning Research, 17:1-32.

Ghosh-dastidar, S., Adelı, H., Dadhmer, N., 2008. Principal Component Analysis Enhanced Cosine Radial Basis Function Neural Network for Robust Epilepsy and Seizure Detection. Biomedical Engineering, 55(2):512 – 518.

Greg Glassman, U.S. Department of Health & Human Services ,2002. "Physical Activity Fundamental To Preventing Disease", 20 June.

Han, J., & Kamber, M., 2006. Data Mining: Concepts and Techniques (2nd ed.). Morgan Kaufmann.

Hannan, S.A., Manza, R.R., Ramteke, R.J., 2010. Generalized Regression Neural Network and Radial Basis Function for Heart Disease Diagnosis, International Journal of Computer Applications, Volume 13, Issue 2, ISBN: 10.5120/1325-1799.

Hardle, W., 1990. Applied Nonparametric Regression, Econometric Society Monographs, 19, Cambridge University Press.

Harold W. Kohl, Steven N.Blair, Ralph S. Paffenbarger, Caroline A. Macera, 1988. Jennie J.Kronenfeld, "A mail survey of physical activity habits as related to measured physical fitness ", American Journal of Epidemiology, Volume 127, Issue 6, 1 June 1988, Pages 1228–1239.

Ian H. Witten, Eibe Frank, Mark A.Hall and Christopher J.Pal, 2016. Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann series in data management systems, 4th Edition: ISBN-13: 978-0128042915.

Ishwaran, H., & Rao, J. S., 2009. Decision Tree: Introduction. In M. Kattan (Ed.). Encyclopedia of medical decision making, 323-328, California: Sage Inc.

Jaakkola T., Yli-Piipari S., Huotari P., Watt A., Liukkonen J. (2016). Fundamental movement skills and physical fitness as predictors of physical activity, A 6-year follow-up study. Scandinavian Journal of Medicine & Science in sports, Volume 26, issue 1, 74-81.

Jay Hoffman, 2006. "Norms for Fitness, Performance, and Health ", Human Kinetics Publishers, Vol: 2, 97–104, 1st Edition: ISBN-13: 9780736054836.

Jie Hu, Debra C. Wallace, Anita S. Tesh, 2010. “Physical Activity, Obesity, Nutritional Health and Quality of Life in Low-Income Hispanic Adults with Diabetes”, J Community Health Nurse, 2010 April, 27(2):70-83.

Korürek, M., & Doğan, B., 2010. ECG beat classification using particle swarm optimization and radial basis function neural network. Expert systems with Applications, 37(12):7563-7569.

Lipo Wang, 2005. "Support Vector Machines: Theory and Applications", Springer Science & Business Media, Volume 177, 1st Edition: ISBN: 978-3-540-32384-6.

López-Vallverdú, J. A., Riaño, D., & Bohada, J. A. (2012). Improving medical decision trees by combining relevant health-care criteria. Expert Systems with Applications, 39(14), 11782–11791.doi:10.1016/j.eswa.2012.04.073.

M.Barnekow, G.Hedberg, U.Janlert, E.Jansson, "Prediction of physical fitness and physical activity level in adulthood by physical performance and physical activity in adolescence - An 18-year follow-up study", Scandinavian Journal of Medicine & Science in sports, Volume 8, issue 5, October 1998, pages 299-308.

Nadaraya, E., A., 1964. On Estimating regression, Theory of Probability and its Applications 10, 186-168.

P.Fergus, A.Hussain, J.Hearty, S.Fairclough, L.Boddy, K.A. Machintoch, G.Stratton, N.D. Ridgers and Naeem Radi, 2015. “A Machine Learning Approach to Measure and Monitor Physical Activity in Children to Help Fight Overweight and Obesity”, Intelligent Computing Theories and Methodologies: 11th International Conference, ICIC 2015, Fuzhou, China, August 20-23, Proceedings, Part II, pages 676-688.

Park, J., Sandberg, I.W., 1991. Universal approximation using Radial-basis Function Networks. Neural Computation, 3, 246-257.

Specht, D.F., 1991. A General Regression Neural Network, IEEE transactions on neural networks, vol. 2. No. 6. page 568-576

Talko Bernhard Dijkhuis, Frank Blaauw, Miriam W Van Ittersum, Marco Aiello, 2018. "Personalized Physical Activity Coaching : A Machine Learning Approach", in Sensors, 18(2) :623, February.

Thomas Reichherzer, Mikayla Timm, Nathan Earley, Nathaniel Reyes, 2017. "Using Machine Learning Techniques to Track Individuals & Their Fitness Activities", In A. Bossard, G. Lee, & L. Miller (Eds.), Proceedings of 32nd International Conference on Computers and Their Applications, March, 20-22, Honolulu, Hawaii, USA (pp. 119–124). Winona, MN, USA: ISCA.

Vapnik V., 2000. The Nature of Statistical Learning Theory. 2nd ed. New York: Springer.

Wang, G., Yeugn, D., Lochovsky, F.H., 2007. A Kernel Path Algorithm for Support Vector Machines, ICML’07 Proceedings of the 24th International Conference on Machine Learning, pages 951-958.

Watson, G., S., 1964. Smooth Regression Analysis, Sankhy, Series A 26, 359-372.

Wright, G. B., Flyer, N., and Fornberg, B., 2013. Radial Basis Functiongenerated Finite Differences: A Mesh-free Method for Computational Geosciences, Handbook of Geomathematics. Springer, Berlin, 1-30.

## CURRICULUM VITAE

Özge BOZKURT was born in Turkey, in 1992. She completed the elementary education at Şanlıurfa, Turkey. She went to the Ç.E.A.Ş Anatolian High School. In 2015, she graduated from the Department of Computer Engineering, Harran University, Şanlıurfa, Turkey. She worked as an Information Technology and Information Security Officer from 2016 to 2018 at Creative Associates International. She started MSc program at the department of Computer Engineering, Cukurova University, Adana, in 2016 under the supervision of Assoc. Prof. Dr. Fatih Akay. She completed MSc program at the Department of Computer Engineering in 2019.