İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

HIZLI MODA SEKTÖRÜNDE MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE SATIŞ MİKTARLARININ TAHMİN EDİLMESİ

YÜKSEK LİSANS TEZİ

Sinem ÖZTÜRK

Endüstri Mühendisliği Anabilim Dalı

Endüstri Mühendisliği Programı

TEMMUZ 2020

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

HIZLI MODA SEKTÖRÜNDE MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE SATIŞ MİKTARLARININ TAHMİN EDİLMESİ

YÜKSEK LİSANS TEZİ

Sinem ÖZTÜRK
(507171126)

Endüstri Mühendisliği Anabilim Dalı

Endüstri Mühendisliği Programı

Tez Danışmanı: Doç. Dr. Seda Yanık ÖZBAY

TEMMUZ 2020

İTÜ, Fen Bilimleri Enstitüsü'nün 507171126 numaralı Yüksek Lisans Öğrencisi Sinem ÖZTÜRK, ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı "HIZLI MODA SEKTÖRÜNDE MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE SATIŞ MİKTARLARININ TAHMİN EDİLMESİ" başlıklı tezini aşağıda imzaları olan jüri önünde başarı ile sunmuştur.

Tez Danışmanı : **Doç. Dr. Seda Yanık ÖZBAY** .....
İstanbul Teknik Üniversitesi

Jüri Üyeleri : **Dr. Öğr. Üyesi Erkan IŞIKLI** .....
İstanbul Teknik Üniversitesi

Doç. Dr. Emre ALPTEKİN .....
Galatasaray Üniversitesi

Teslim Tarihi : 12 Haziran 2020
Savunma Tarihi : 13 Temmuz 2020

Anne ve Babama,

ÖNSÖZ

Aldığım her kararda arkamda duran, eğitim hayatım boyunca desteklerini hiçbir zaman esirgemeyen, her zaman bir üst noktayı bana hedef olarak gösteren biricik annem Seniha Öztürk'e, kıymetli babam Özkan Öztürk'e göstermiş oldukları destekler ve emekler için teşekkür eder, minnet ve şükranlarımı sunarım.

Bu tez çalışmasının araştırma aşamasından nihayete ermesine kadar olan süreçte yardımları ve destekleri için saygıdeğer danışmanım Doç. Dr. Seda Yanık Özbay'a teşekkür ederim.

Haziran 2020

Sinem ÖZTÜRK
Endüstri Mühendisi

İÇİNDEKİLER

Sayfa

KISALTMALAR

YSA : Yapay sinir ağları
SKU : Stok tutma birimi
ÇKA : Çok katmanlı algılayıcı
MAE : Ortalama mutlak hata
RMSE : Hata ortalama karelerin karekökü
RI : Rand endeksi

SEMBOLLER

$\mathbf{j}$	: Sinir hücresi (girdi indisi)
$\mathbf{i}$	: Çok katmanlı algılayıcı katman indisi
$\mathbf{w_{i,j}(t)}$	: Girdi değişkenlerinin ağırlıkları
$\mathbf{\Delta w_{i,j}(t)}$	: Girdi ağırlıkları güncellendikten sonra değerler arasındaki fark
$\mathbf{P_i}$	: Uzman bilgisi
$\mathbf{d(t)}$	: Çıktı değeri
$\mathbf{e(t)}$	: Hata sinyali
$\mathbf{v_j\,(n)}$	: Toplama Fonksiyonu
$\boldsymbol{\varphi}(.)$	: Aktivasyon Fonksiyonu
$\mathbf{B_j}$	: *j*. nörona uygulanan sapma değeri
$\mathbf{M}$	: Nöron *j*'ye gelen toplam girdi (sapma hariç) sayısıdır
$\boldsymbol{\eta}$	: Öğrenme parametresi
$\boldsymbol{\delta_j}\,(\mathbf{n})$	: Lokal değişim ölçüsü

ÇİZELGE LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

HIZLI MODA SEKTÖRÜNDE MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE SATIŞ MİKTARLARININ TAHMİN EDİLMESİ

ÖZET

Talep tahmini, gelecek periyottaki müşteri ihtiyaçlarını belirlemek için geçmiş verilerin kullanıldığı bir süreçtir. Bu tahmin geleceğe ilişkin olduğundan büyük oranda belirsizlik içermektedir ve bu sebeple gelecek periyot talebini kesin olarak kestirmek çoğu zaman mümkün değildir. Ancak küçük bir hata payı ile elde edilen tahminler karar vermede önemli rol oynamaktadırlar.

Bu araştırma kapsamında makine öğrenmesi yöntemleri moda sektöründe faaliyet gösteren bir firma ürünlerinin talep tahmini için kullanılmıştır. Moda sektöründe üretilen ürünlerin kullanım ömrüne göre teslim süresinin uzun olması, her bir sezon için talep edilen ürünlerin çok hızlı değişikliğe uğraması, müşteri beklentilerin en hızlı ve en iyi şekilde karşılanma ihtiyacı, pazardaki rekabet gücü ve daha birçok etken özellikle bu sektördeki talep tahmininin doğru yapılmasına olan gereksinimi oldukça önemli kılmaktadır.

Bununla birlikte hava şartlarına karşı çok hassas olan mevsimsel satışlar, üretilen ürün çeşitliliğinin oldukça fazla olması, ürün tasarımlarının her daim güncel tutulması ve çoğu ürünün bir sonraki koleksiyonda yer almaması talep tahminini oldukça karmaşık bir hale getirmektedir. Tüm bu kısıtlamalar, hazır giyim şirketleri için satış tahmin sistemlerini çok özel ve karmaşık hale getirmektedir.

Yapılan literatür araştırması sonucunda son yıllarda perakende sektöründeki talep tahmini problemi çözümü için makine öğrenme yöntemlerine sıkça başvurulduğu görülmüştür. Bu çalışmalardan hareketle uygulama kapsamında makine öğrenmesi yöntemleri karşılaştırmalı olarak değerlendirilmiş ve veri setine en iyi yanıt veren, veri setini en iyi öğrenebilen algoritma seçimi üzerinde durulmuştur.

Çalışmada yapay sinir ağlarından çok katmanlı algılayıcı, bir karar ağacı alt yapısı ile çalışan rassal orman algoritması ve kümeleme yöntemlerinden *K*-ortalamalar ve *X*-ortalamalar algoritmaları kullanılmıştır. *K*-ortalamalar algoritması kendi başına küme sayısını belirleyemez ve bu da bir dezavantaj oluşturur. *X*-ortalamalar algoritması aslında *K*-ortalamalar algoritma mantığı ile çalışan ve optimum *k* değeri belirleme işlemini kullanıcıya değil de kendi başına halledebilen bir algoritma olduğu için optimum küme sayısı *X*-ortalamalar algoritması ile belirlenmiş olmaktadır.

Çalışma kapsamında kullanılan veri seti çoğunlukla nominal değerlerden oluştuğu için bu değerlerin regresyon uygulamasına yanıt verebilmesi için ikilik tabandaki sayılara çevrilmişlerdir. Nümerik değişkenlerin de birbirleri cinsinden ifade edilebilmesini sağlama amacı ile normalleştirme yaklaşımı uygulanmıştır. Uygulama kapsamında kullanılan normalleştirme yöntemi Min-Maks yöntemidir.

Uygulamanın ilk aşaması kısa dönemlik gerçekleşen ürün sipariş verilerini kullanarak bir sonraki sipariş edilebilecek ürünü ve ürün miktarını tahmin etmektir. Bunu gerçekleştirmek için eldeki veri rassal bir şekilde eğitim ve test veri setlerine ayrılmış ve eğitim veri seti çok katmanlı algılayıcı ve rassal orman algoritması ile çalıştırılmıştır.

Yapılan ilk denemenin sonucunda çok katmanlı algılayıcının rassal orman algoritmasına göre performans kriterleri gözetilerek daha iyi sonuç verdiği gözlemlenmiştir. Bu üstünlük halinin rastlantısal bir durum olup olmadığını anlamak için de birbirlerinden farklı 10 adet eğitim ve test setleri oluşturulup her iki algoritma da bulunan veri setleri ile çalıştırılmıştır. Bunun sonucunda da bulunan performans kriterlerine *t*-testi uygulanarak bulunan sonucun tesadüfi olmadığı, 10 denemenin her birisinde çok katmanlı algılayıcı rassal orman algoritmasına göre daha iyi sonuç verdiği belirlenmiştir.

Yapılan uygulamanın ikinci aşamasında, ürün özelliklerine bakılmaksızın tüm ürünleri aynı veri seti içerisinde bulunduran ilk aşamadaki süreçten ziyade birbirlerine benzer özellik gösteren ürünleri kümelemenin ve bu işlem sonunda her bir küme için ayrı ayrı tahmin çalışması yapmanın tahmin performansını iyileştirip iyileştirmediği hesaplanmıştır. Kümeleme yaklaşımında da iki yöntem (*K*-ortalamalar, *X*-ortalamalar algoritması) ile çalışılmış ve her iki yöntemden elde edilen sonuçlar rand endeks değeri hesaplanarak birbiri ile karşılaştırılmış ve daha iyi bir endeks değerine sahip olduğu için *K*-ortalamalar algoritması ile belirlenen kümelere tahmin çalışması uygulanmıştır. Çalışma sonunda elde edilen performans değerleri birinci aşamada bulunan performans değerleri ile karşılaştırılmış ve kümelemenin tahmin performansını iyileştirmediği gözlemlenmiştir.

Uygulamanın üçüncü aşamasında ise amaç, seçilen bir ürüne ait geçmiş yılların sipariş verileri kullanılarak gelecek periyottaki sipariş miktarını tahmin etmektir. Çalışmanın bu kısmında mevcut olan tüm sezonlardaki ortak bulunan bağımsız değişken değerleri belirlenmiş ve bunun sonucunda veri setinde belirli sadeleştirmeler yapılmıştır. Veri seti yine rassal olacak şekilde eğitim ve test veri seti olarak iki kısımda incelenmiştir. Eğitim veri seti ile yapılan regresyon çalışmasının sonucunda rassal orman algoritması performans değerlendirme kriterleri açısından daha tahmin edici bir sonuç vermiştir.

Bu üstünlük halinin rastlantısal bir durum olup olmadığını anlamak için de birinci aşamada yapıldığı gibi birbirlerinden farklı 10 adet eğitim ve test setleri oluşturulup her iki algoritma da bulunan veri setleri ile çalıştırılmıştır. Bunun sonucunda da bulunan performans kriterlerine *t*-testi uygulanarak bulunan sonucun tesadüfi olmadığı, 10 denemenin her birisinde rassal orman algoritmasının çok katmanlı algılayıcı algoritmasına göre daha iyi sonuç verdiği belirlenmiştir.

Sonrasında 2004 ve 2005 yıllarındaki dört sezon ele alınarak 2006 yılındaki ilkbahar yaz sezonundaki ceket miktarı tahmin edilmeye çalışılmıştır. Bulunan tahmin sonucu 2006 yılı ilkbahar yaz sezonunda gerçekleştirilen ceket üretim miktarları ile karşılaştırılmış ve sonucun tutarlı olduğu gözlemlendiği için 2007 sonbahar kış sezonundaki ceket talep miktarı tahmin edilmeye amacıyla sırasıyla 2005 ve 2006 yıllarındaki sezon verileri kullanılmıştır.

Uygulama kapsamında kullanılan çok katmanlı algılayıcı ve rassal orman algoritması parametreleri seçimi optimal bir seçim yöntemi bulunmadığı için birçok defa yapılan deneme sonuçları karşılaştırılıp en iyi değeri veren parametre optimal parametre değeri olarak belirlenmiştir. Regresyon çalışmaları için belirlenen performans kriterleri;

korelasyon katsayısı, ortalama mutlak hata (MAE), hata ortalamalarının karekökü (RMSE) olmaktadır.

Değişkenliğin getirdiği riskin çok yüksek olduğu perakende sektöründeki en önemli problemlerden biri olan talep tahmini konusunda makine öğrenme yöntemlerinin ne kadar etkin bir biçimde sonuç verdiği bu çalışma kapsamında gözlemlenmiştir. Ancak algoritmalar aynı veri seti üzerinde benzer etkiyi yaratmamaktadır ve bundan dolayı da veri seti için algoritma seçimi büyük önem taşımaktadır.

Önerilen metodoloji, firma için müşteri taleplerinin tahmin edilmesinde başarılı bir karar destek sistemi olarak düşünülmektedir. Gelecekteki çalışmalarda topluluk öğrenmesi (ensemble learning) gibi farklı makine öğrenmesi teknikleri ile uygulama yapılabilir.

FORECASTING OF SALES QUANTITIES BY MACHINE LEARNING METHODS IN FAST FASHION SECTOR

SUMMARY

Demand forecasting is a process where historical data is used to determine customer needs in the next period. Since this forecasting is related to the future, it contains a great deal of uncertainty and therefore it is not possible to predict the future period demand precisely. However, predictions obtained within a small margin of error play an important role in decision making.

Within the scope of this research, machine learning methods were used to predict demand for products of a company operating in the fashion industry.

Since delivery time which is based on the lifetime of products produced in the fashion industry is long, products demanded for each season are changed very quickly, customer needs have to be met in the fastest and best way, to maintain competitive power in the market, it is critical to predict demand accurately in this industry.

However, seasonal sales, which are very sensitive to weather conditions, high degree of variety of the products produced, and the fact that product designs are always up to date and that most products are not included in the next collection make demand forecasting a very complicated task. All these restrictions make sales forecast systems very special and complex for garment companies.

A thorough review of the related literature revealed that machine learning methods are frequently used in order to solve the demand forecasting problem in the retail sector. Based on these studies, some machine learning methods were evaluated comparatively within the scope of this study, and the selection of the algorithm that fit the data set best and the one that was able to learn the data set best was emphasized.

In this study, multi-layer perceptron algorithm from artificial neural networks, random forest algorithm working with a decision tree infrastructure and *K*-means and *X*-means algorithms from clustering methods were employed. The *K*-means algorithm cannot determine the number of clusters on its own, which creates a disadvantage. The logic of the *X*-means algorithm is the same as that of the *K*-means algorithm; however, unlike the *K*-means algorithm finding the optimum number of clusters is not left to the user, and the *X*-means algorithm determines it on its own.

Since the data set used in this study mostly consists of nominal inputs, they were converted to dummy variables so that they can respond to the regression application well. A normalization procedure, which is called the Min-Max method, was applied to ensure that numerical variables are also be expressed on the same scale.

The first stage of the application utilized short-term product order data to predict which product and how much of it would be ordered in the next period. To accomplish this, the available data was randomly divided into training and test datasets and the training datasets were run with a multi-layer perceptron and random forest algorithm. As a result of the first trial, it was observed that the multi-layer sensor gave better results

by considering the performance criteria according to the random forest algorithm. In order to understand whether this superiority is a random situation, 10 different training and test sets were created and run with datasets with both algorithms. As a result, it was determined that the result found by applying t-test to the performance criteria was not obtained by chance. It was determined that multilayer perceptron gave better results in each of 10 trials rather than random forest algorithm.

In the second stage of this study, products that show similar characteristics in the data set were grouped in the same clusters. At the end of the clustering process, the effect of performing prediction study for each cluster separately was calculated. In clustering approach, two methods (*K*-means, *X*-means algorithm) were studied and the results obtained from both methods were calculated by comparing the rand index value. Since rand index value was greater than X-means algorithm, *K*-means algorithm was applied to the data set. The performance values obtained at the end of the study were compared in terms of performance values found in the first stage and it was observed that clustering could not improve the predictive performance.

In the third stage of this study, the aim was to estimate the order amount of the next period by using the historical data of a predetermined product. In this part of study, common variable values found in all seasons were determined and certain simplifications were made in the data set. The data set was examined randomly in two parts as a training data and test data. As a result of the regression study conducted with the training data, the random forest algorithm gave a more predictive result in terms of performance evaluation criteria. As what was done in the first stage, in order to understand whether this superiority was a coincidence or not, 10 different training and test sets were created and run with datasets with both multi-layer perceptron and random forest algorithms. In conclusion, it was determined that the result found by applying *t*-test to the correlation coefficient was not obtained by chance, and the random forest algorithm gave better results compared to the multi-layer perceptron algorithm in each of the 10 trials.

Afterwards, an estimation study was conducted by using the sales data of spring-summer and fall-winter seasons in 2004 and 2005 as training data set and the sales data of spring-summer season in 2006 as test data set. The actual sales observed in the spring and summer season of 2006 were compared with the predicted values found in the same time period and it was seen that the comparison result was consistent. The sales forecasts of all seasons were calculated by using the data of the previous two years. The estimated values and the actual values were compared, and the difference between these two values was controlled.

Since there is no optimal selection method for multi-layer sensor and random forest algorithm parameters, the trial and error results were compared many times and the parameter giving the best value was determined as the optimal parameter value. Performance criteria determined for regression studies; correlation coefficient, mean absolute error (MAE), root mean square (RMSE).

Algorithms also differ according to the variable types included in datasets. Some methods can handle both categorical variables and numeric variables, while others can only process categorical or only numeric values. For this reason, one of the biggest points to be considered is to determine the algorithm well according to the data set.

In this study, how effectively the machine learning methods have yielded results in demand forecasting, which is one of the most important problems in the retail sector, where the risk brought by variability is very high. However, the algorithms do not have

a similar effect on the same data set, and therefore the selection of the algorithm is of great importance for the data set.

The proposed methodology is considered as an effective decision support system for forecasting of sales quantities for the company. In future studies, different machine learning techniques can be applied.

1. GİRİŞ VE PROBLEM TANIMI

Tedarik zinciri terimi, bir ürünün üretiminde karşılaşılan ilk süreçten nihai tüketiciye nihai satışa kadar mal akışını tanımlamak için kullanılır (Aksoy ve diğ, 2012). Bir tedarik zincirinin tedarik, üretim ve dağıtım olmak üzere üç temel süreci vardır. Bu süreçler sadece üreticileri ve tedarikçileri değil aynı zamanda taşıyıcılar, depolar, perakendeciler ve müşterileri de kapsamaktadır. Süreçler arasında bilgi, ürün ve kaynak ileri ve geri yönlü akmaktadır. Tedarik zincirinin performansını etkileyen birçok faktör bulunmaktadır. En önemli faktör ise talep tahmin doğruluğudur. Günümüzde perakende sektöründe faaliyet gösteren firmaların yüzleştikleri en büyük problemlerden biri de talep kestirimini iyi yapamamaları sonucunda meydana gelen stoksuz kalma ya da aşırı stok tutma gibi kar marjını önemli ölçüde etkileyen sonuçlardır.

Talep tahmini, gelecek periyottaki müşteri ihtiyaçlarını belirlemek için geçmiş verilerin kullanıldığı bir süreçtir. Bu tahmin geleceğe ilişkin olduğundan büyük oranda belirsizlik içermektedir ve bu sebeple gelecek periyot talebini kesin olarak kestirmek çoğu zaman mümkün değildir. Ancak küçük bir hata payı ile elde edilen tahminler karar vermede önemli rol oynamaktadırlar.

Tekstil / giyim sektöründe üretilen ürünlerin kullanım ömrüne göre teslim süresinin uzun olması, her bir sezon için talep edilen ürünlerin çok hızlı değişikliğe uğraması, müşteri beklentilerin en hızlı ve en iyi şekilde karşılanma ihtiyacı, pazardaki rekabet gücü ve daha birçok etken özellikle bu sektördeki talep tahmin doğruluğu gereksinimini oldukça önemli kılmaktadır. Bununla birlikte hava şartlarına karşı çok hassas olan mevsimsel satışlar, üretilen ürün çeşitliliğinin oldukça fazla olması, ürün tasarımlarının her daim güncel tutulması ve çoğu ürünün bir sonraki koleksiyonda yer almaması talep tahminini oldukça karmaşık bir hale getirmektedir. Tüm bu kısıtlamalar, hazır giyim şirketleri için satış tahmin sistemlerini çok özel ve karmaşık hale getirmektedir.

Geçmişte tekstil sektöründe talep tahmini yapılırken geleneksel olarak bir yılda üretilecek ürünler ilkbahar-yaz ve sonbahar-kış sezonu olarak ayırılırdı ve her sezon başı bütün ürünler satışa sunulurdu.

Her sezonun sonunda ise satılmamış ürünler indirime girerek elde kalan ürünler temizleniyordu. Ancak günümüzde hızlı moda terimi tekstil sektörüne hâkim olduğundan bu yana ürün sirkülasyonları çok hızlı olmaya başladı. Bununla birlikte talep tahmini de tüm sezonu kapsayacak şekilde değil belirli periyotlarda yapılmaya başlandı. Sezonda başta satılan ürünler kullanılarak tek sezonun satışlarını ve koleksiyonun kendisini inceleyip o koleksiyona ek gelecek ürünün ne kadar satılacağı daha çok önem arz eder oldu.

Satış tahmin yöntemleri alanında istatistiksel yöntemler yakın geçmişte en çok kullanılan yöntemler arasında yer almaktaydı. Bu istatistiksel teknikler, üstel yumuşatma, Holt-Winters modeli, Box & Jenkins modeli, regresyon modelleri veya otoregresif entegre hareketli ortalama (ARIMA) modellerine sahip iyi bilinen çeşitli modelleri içerir. Bazı alanlarda tatmin edici sonuçlar vermesine ve basitliği ve hızlı cevap verebilmesi adına popüler olarak kullanılmasına rağmen, istatistiksel yöntemler moda sektörü için uygulandığında birkaç sorun ortaya çıkmaktadır.

Çoğu zaman serisi yöntemi büyük tarihsel veri setleri, parametrelerinin karmaşık bir optimizasyonu ve kullanıcının belirli bir deneyimini gerektirdiğinden ve genellikle doğrusal bir yapı ile sınırlı olduğundan bu yöntemler tekstil / konfeksiyon ortamı için ve daha genel olarak herhangi bir moda sektöründe kullanım açısından uygun değildir, bu nedenle yapay zeka (AI) gibi daha sofistike yöntemlerle karşılaştırıldığında performansları genellikle daha kötüdür (Brahmadeep ve Thomassey, 2016).

Bu çalışmada Türkiye'de tekstil sektöründe geniş bir üretim ve satış hacmine sahip olan spor ürünleri satan uluslararası en büyük firmalardan birine ait toplam dört yıldan oluşan sipariş verileri göz önünde bulundurularak yapılan ilk uygulamada sezon içi değişen talep tahminini ele alan ve ikinci olarak ise belirlenen bir ürün için geçmiş sezon satış verileri kullanılarak gelecek sezondaki satış miktarı tahminini ele alan bir sistem geliştirilmesi amaçlanmıştır.

Çalışmanın ikinci bölümünde makine öğrenme yöntemleri ile talep tahminine ilişkin literatürdeki çalışmalar incelenmiştir ve çalışma kapsamında kullanılacak olan yapay sinir ağları, karar ağacı algoritmaları ve kümeleme algoritmalarına ait çalışmalar yer almaktadır.

Çalışmanın üçüncü bölümünde inceleme kapsamında karşılaştırılacak olan çok katmanlı algılayıcı ve karar ağacı algoritması, rassal orman algoritması, *K*-ortalamalar ve *X*-ortalamalar kümeleme algoritmaları ve teorik alt yapıları ayrıntılı şekilde açıklanmıştır. Bunların devamında da tahmin ve kümeleme performansının ölçümü için gerekli kriterler belirlenmiş ve formülleri açıklanmıştır. Bu bölümde makine öğrenme yöntemlerinde sıklıkla kullanılan normalleştirme yöntemlerinden de bahsedilmiştir.

Tahmin performansının ölçümü için belirlenen kriterler korelasyon katsayısı, mutlak ortalama hata (MAE) ve hata ortalama kareler toplamının karekökü olan RMSE değerleri dikkate alınmıştır. Bunların yanı sıra kümeleme performans analizi için kullanılan performans kriteri rand endeks değeri olarak belirlenmiş ve bu bölümde açıklamaları yapılmıştır.

Çalışmanın dördüncü bölümünde ise uygulama kapsamında yapılan çalışmalar ayrıntıları ile anlatılmıştır. Bu bölümde başlangıç olarak problem tanıtımı, metodolojiden ve uygulamanın yapılacağı programdan bahsedilmiştir. Daha sonra veri setleri tanıtılmış ve veri dönüştürme işlemlerinde yapılan çalışmalardan bahsedilmiştir. Uygulama üç aşamadan oluşmaktadır.

Birinci aşamada dönüştürülen sezonluk veri seti %90 eğitim %10 test veri seti olarak rassal bir şekilde program tarafından ikiye ayrılmıştır. Devamında elde edilen eğitim veri seti regresyon çalışması ile hem çok katmanlı algılayıcı hem de rassal orman yöntemi ile hata ölçümü yapılarak en düşük hatayı veren yöntem seçildikten sonra, tahmin çalışması yapılarak koleksiyona eklenebilecek bir sonraki ürünün nitelikleri ve talep miktarı belirlenmiştir. İkinci aşamaya gelindiğinde, incelenen mevcut sezonda üretilen bütün ürünleri kapsamakta olduğundan benzer özellik gösteren ürünlere kümeleme çalışması uygulanması tahmin performansını iyileştirip iyileştiremeyeceği üzerine çalışılmıştır.

Kümeleme çalışması için belirlenen *K*-ortalamalar ve *X*-ortalamalar kümeleme yöntemlerinin uygulanma sonuçları birbirleri ile karşılaştırılmıştır. Bu karşılaştırma

rand endeks değeri hesaplanarak elde edilmiştir. Devamında, endeks değerine göre daha iyi olan kümeleme algoritmasında elde edilen her kümeye bir önceki aşamada sonucu daha iyi verdiği için çok katmanlı algılayıcı ile regresyon çalışması uygulanmış ve kümeleme yapılmadığı haldeki performans kriterleri kümelemenin verdiği performans kriterleri ile kıyaslanmıştır.

Üçüncü aşamaya gelindiğinde ise, belirlenen bir ürün için geçmiş sezonlardaki sipariş verileri göz önünde tutularak ürüne ait gelecek sezonda gerçekleşebilecek tüm sipariş miktarı tahmin edilmeye çalışılmıştır. Bunu gerçekleştirmek için iki algoritma ile değerlendirilmeye tabi tutulmuş ve bu veri seti için en iyi sonucu veren yöntem ile tahmin çalışması yapılarak gelecek sezondaki belirlenen ürüne ait toplam satış miktarı tahmin edilmeye çalışılmıştır. Sonrasında 2004 ve 2005 yıllarındaki dört sezon ele alınarak 2006 yılındaki ilkbahar yaz sezonundaki ceket miktarı tahmin edilmeye çalışılmıştır. Bulunan tahmin sonucu 2006 yılı ilkbahar yaz sezonunda gerçekleştirilen ceket üretim miktarları ile karşılaştırılmış ve sonucun tutarlı olduğu gözlemlendiği için 2007 sonbahar kış sezonundaki ceket talep miktarı tahmin edilmeye amacıyla sırasıyla 2005 ve 2006 yıllarındaki sezon verileri kullanılmıştır.

Çalışmanın son bölümünde ise uygulamadan alınan sonuçlar değerlendirilmiş ve kullanılan yöntemlerin problem seçimindeki etkileri doğrultusunda gelecek çalışmalara yön verebilme adına geliştirme önerileri verilmiştir.

2. LİTERATÜR ARAŞTIRMASI

2.1 Talep Tahmininde Yapay Sinir Ağları

Sinir ağı teknolojisi, insan beyninin bilgi ve organizasyon becerilerinin kazanılmasını taklit etmek amacıyla geliştirilmiştir. Verilerin düzenlenmesi, sınıflandırılması ve özetlenmesi konusunda önemli destek sunmaktadır. Ayrıca, giriş verileri arasında kalıpları ayırt etmeye yardımcı olur, az sayıda varsayım gerektirir ve yüksek derecede tahmin doğruluğu sağlar. Bu özellikler sinir ağı teknolojisini finanstan tıbba kadar çeşitli alanlarda tanınma, sınıflandırma, tahmin ve optimizasyon için potansiyel olarak umut verici bir araç haline getirmektedir (Kwon, 2011).

Yapay sinir ağları (YSA), insan beynindeki nöronların öğrenme sürecini simüle etmek için tasarlanmış bilgisayar tabanlı sistemlerdir. Yapay sinir ağları, sistemi yöneten fiziksel ve kimyasal yasalar hakkında gerçek bilgiye sahip olmadan bir dizi deneysel veriden öğrenme yeteneğine sahiptir. Bu nedenle, veri sistemlerinin YSA uygulaması özellikle sistemlerin doğrusal olmamaları ve karmaşık davranışlar göstermesi durumunda yüksektir (Kwon, 2011). Yapay sinir ağları son dönemlerde tekstil sektöründe talep tahmini problemlerinde sıklıkla kullanılmakta ve başarılı sonuçlar vermektedir.

Yu ve diğ. (2011) yaptıkları çalışmada, moda sektöründe faaliyet gösteren bir firma için talep tahminini hızlı ve daha doğru yapabilme amacıyla istatistiksel yöntemlerin kombinasyonu ile yapay zekâ uygulamasının sonuçlarını performans ölçütlerince karşılaştırma gibi bir yöntem önermişlerdir. Satış miktarının ürün rengi, bedeni, fiyatı gibi kriterlerle ilişkisini anlamak için 10 ürün üzerinde inceleme yapılmış ve zamanın maliyet üzerindeki etkisi hesaplanmıştır. Sonuç performansı MSE değeriyle ölçülmüştür. Çıkan sonuçlara bakıldığında yapay zekâ uygulaması ile yapılan tahmin hatasının istatistiksel yöntemlerin verdiği hatadan daha düşük olduğunu dolayısıyla tahmin performansı bakımından daha iyi olduğu gözlemlenmiştir.

Huang ve Liu (2017) yaptıkları çalışmada, çalışmada, yeni giyim ürünleri için iki aşamalı akıllı perakende tahmin sistemi tasarlamışlardır. İlk aşamada stok çıkışı dikkate alınarak orijinal satış verileriyle talep tahmin edilmektedir. Uyarlanabilir nöro-

bulanık çıkarım sistemi (ANFIS) talebi tahmin etmek için ikinci aşamaya girmiştir. Bu arada, yeni ürünlerin sınırlı verileri nedeniyle bir veri seçim süreci sunulmaktadır. Ampirik veriler Kanadalı bir hızlı moda şirketindendir. Sonuçlar, talep ve satışlar arasındaki ilişkiyi ortaya koymakta, talep tahmin sürecini bir tahmin sistemine entegre etmenin gerekliliğini ortaya koymakta ve ANFIS tabanlı tahmin sisteminin geleneksel YSA tekniğinden daha iyi performans gösterdiğini göstermektedir. Üzerinde çalışılan firma uzmanlarıyla görüşerek ve önceki literatürü gözden geçirerek talep tahmininin yapılabilmesi için beş faktör belirlenmiştir. Bu değişkenler orijinal fiyat seviyesi, promosyon seviyesi, yükselme seviyesi, boyut tercihi, takvim faktörü gibi değişkenlerdir. Beş değişken belirlendikten sonra veriler eğitim ve test üzere iki bölümde incelenmiştir. Farklı parametreleri ile deneme yapılmış ve en iyi sonucu veren parametre ile de talep tahmini yapılmıştır. Değerlendirme kriteri olarak MSE ve MAPE hata değerleri seçilmiştir. Sonuç olarak da MSE ve MAPE değerlerine bakılarak ANFIS sisteminin belirsizliğin çok yüksek olduğu giyim sektöründe kullanımının uygun olduğu sonucuna karar vermişlerdir.

Brahmadeep ve Thomassey (2016) yaptıkları çalışmada mağazalardaki ürünleri yenileme simülasyonu ile uzun ve kısa vadeli tahminler yapan iki aşamalı bir tahmin sistemi geliştirmişlerdir. Tahmin doğruluğu ve ayrıca kalan envanter ve toplam satış, modellerin kalitesini değerlendirmek için kullanılan üç performans göstergesidir. Önerilen sistemin denenmesi için kullanılan veriler, bir Fransız perakendecinin veri tabanından alınmıştır. Bu veriler satışları ve uzun vadeli tahmin için kullanılan tanımlayıcı kriterleri içerir. 482 adet ürünün kısa dönem satış verilerini kullanarak bu verilerden yapay zeka uygulaması kullanılarak kısa dönemli talep tahmini yapılmaya çalışılmıştır. Uzun vadeli satış tahmin modeli için ana sorun, geçmiş veri olmadan doğru tahminler yapmaktır. Böyle bir durumda, satış ve ilgili ürün tanımlayıcı kriterleri arasında ilişki kurmak için genellikle karar ağaçları veya sinir ağları gibi bir veri madenciliği tekniği kullanılır. Böylece, yeni ürünlerin satış tahminleri tanımlayıcı kriterlerinden hesaplanabilir. Bu çalışmada 142 yeni ürünün uzun vadeli tahminini elde etmek için önerilen sistemi kümeleme ve karar ağaçlarına dayanarak uygulanmıştır. Hızlı moda ürünleri için, kısa vadeli tahminler hem sınırlı bir sürede hem de sınırlı miktarda veri üzerinde gerçekleştirilmelidir. Değerlendirme kriteri olarak hataların karelerinin toplamının karekökü (RMSE) ele alınmıştır. Sonuç olarak önerilen model hızlı giyim ürünlerinin tahmini konusunda başarıyla uygulanmıştır.

Hui ve Choi (2016) yaptıkları incelemelerinde, giyim perakendecilerinin satış tahminleri yapması ve giyim ürünleri için uygun stok seviyesini belirlemek için yapay zekanın ne derece başarılı olabileceğini teorik olarak ele almışlardır. Yazarlara göre tekstil perakendeciliğinde karşılaşılan sorunlar birçok perakendecinin karşılaştığı sorun, belirli müşteriler hakkında çok fazla bilginin eksik olmasıdır ve bu tür veri ve bilgilerin miktarı çok büyüktür. Moda perakendecilerinin, bireysel müşterilerin satın alma davranışlarını ayrıntılı olarak anlamaları ve ekonomi değişimleri ve moda trend değişiklikleri gibi çevresel değişkenlerden etkilenen satın alma değişikliklerini takip etmesi zordur. Bu nedenle, veri madenciliği, potansiyel müşterilerin satın alma tercihini belirlemeye ve moda pazarlama ekibinin daha iyi müşteri hizmeti ve elde tutma için onları hedeflemesine olanak tanıyan yaklaşımlardan biridir. Uzun gelişim süreci ve geniş uygulama yelpazesi ile YSA modeli, özellikle giyim sektöründeki iş süreçlerinde sıkça karşılaşılan, doğrusal olmayan ilişkilere sahip veri setleri için umut verici bir modelleme tekniğini temsil etmekte olduğu yapılan inceleme sonucunda görülmüştür. Model spesifikasyonu açısından, YSA'ların veri kaynağı hakkında çokça geçmiş veri gerektirmediğine, ancak genellikle tahmin edilmesi gereken çok sayıda ağırlık içerdiğinden, büyük bir eğitim veri seti gerektirdiğine değinmişlerdir.

Giri ve diğ. (2019) yaptıkları çalışmada, kadın giyim ürünlerinin geçmiş satış verilerini ve bunlara karşılık gelen görüntüleri kullanarak bir sonraki periyottaki yeni ürün miktarını tahmin etmeyi amaçlamışlardır. Çalışma için 320 farklı ürün resmi dikkate alınmıştır ve her görüntü benzersiz bir öğe tanımlamaktadır. Görüntülerin nitelik resimlerinin çıkarılması için geliştirilen V3 derin sinir ağı kullanılmıştır. Görüntü işleme sürecinden sonra veri seti 2049 adet değişken ile talep miktarını tahmin etmeye çalışmaktadır. Regresyon modeli için birçok uygulamada öğrenme başarısı kanıtlanmış çok katmanlı algılayıcı (MLP) modeli kullanılmaktadır. Değerlendirme kriteri olarak ortalama mutlak hata (MAE), hataların karelerinin ortalaması (MSE), hataların karelerinin ortalamasının karekökü (RMSE) kullanılmıştır. Sonuç olarak belirli bir grup ürün satışının birkaç hafta boyunca oldukça doğru bir şekilde tahmin edildiği sonucunu gözlemlemişlerdir.

Ngai ve diğ. (2014) yaptıkları çalışmada, tekstil ve hazır giyim tedarik zincirlerinde karar desteği ve akıllı sistemlerin uygulanması ile ilgili araştırma makalelerinin kapsamlı bir incelemesini sunmaktadır. Veriler 1994'ten 2009'a kadar 35 dergide yayınlanan 77 makaleden elde edilmiştir. Makaleler, tekstil üretimi, hazır giyim

üretimi ve dağıtım / satış olmak üzere üç temel sektöre uygulanabilirliklerine göre sınıflandırılmıştır. Bu çalışmada tanımlanan kategorilere ayrılmış dergi makalelerinin kapsamlı bir listesi, araştırmacı ve uygulayıcılara, bir tekstil ve hazır giyim tedarik zincirinin çeşitli aşamalarına karar desteği ve akıllı sistemlerin uygulanması konusunda bilgiler ve ilgili referanslar sunmaktadır. Geliştirilen sınıflandırma çerçevesi ışığında, sektörde yapay sinir ağlarının talep tahmini konusunda diğer uygulamalara nazaran daha fazla kullanıldığını gözlemlemişlerdir.

Semenov ve diğ. (2017) yaptıkları çalışmada, otomat makineleri adı verilen otomatik perakende sistemlerinin sorunlarını ele almışlardır. Çalışmanın amacı, yapılması maksimum kar getirecek bir otomat çeşidinin oluşturulmasıdır. Semenov ve diğ. (2017) sinir ağı yoluyla yapay zeka oluşturma yöntemlerinden birine odaklanmışlardır. Perakende satışları tahmin etmek için sinir ağlarını kullanma deneyimi, öngörülen ve gerçek veriler arasındaki benzerliği garanti etmeyi mümkün kılar. Yapay sinir ağı kullanılarak yapılan tahmin çalışması sonucunda tahmin sonuçları ile gerçek değerler arasında %10'luk bir hata payı ile tahmini başarılı şekilde gerçekleştirmişlerdir.

Slimani ve diğ. (2016) yaptıkları çalışmada, Fas'ta bulunan bir süpermarketin geçmiş talep günlük talep tahminleri üretebilmek için yapay sinir ağları kullanılmıştır. Uygulanan yöntem yapay sinir ağları içerisinde oldukça popüler olan çok katmanlı algılayıcı yapısıdır. Sonuçlar, sinir ağına yeni girdiler eklemenin, vaka çalışmasında, kısa vadeli talep tahmininin doğruluğu üzerinde olumlu bir etkiye sahip olduğunu göstermektedir. Kullanılan veriler eğitim ve test verileri olmak üzere ikiye ayrılarak incelenir. Eğitim seti olarak ele alınan periyodun ilk üç ayı eğitim, geriye kalan üç ayı ise test seti olarak kullanılmıştır. Çok katmanlı algılayıcı için kullanılan parametreler birkaç deneme sonrasında belirlenmiştir. Değerlendirme kriteri olarak hataların karelerinin ortalaması (MSE) kullanılmıştır. Sonuç olarak yapay sinir ağı kullanılarak elde edilen tahminin gerçek tahmine yakın olduğunu gözlemlemişlerdir.

Sharma ve diğ. (2019) yaptıkları çalışmada, regresyon analizi, ileri makine öğrenmesi tekniklerinden olan karar ağaçları ve yapay sinir ağları gibi çeşitli modelleme tekniklerinin, e-ticaret platformlarındaki satışları tahmin etmede etkinliğini göstermeyi ve Amazon'da ki kitap satışları için hangi değişkenlerin önemli olduğunu incelemeyi amaçlamaktadır. Uygulama için seçilen kriterler güncel fiyat, indirim tutarı, indirim oranı, inceleme hacmi, sayfa sayısı, olumlu duygu gücü, olumsuz duygu gücü, pozitif duyarlılık gücü, negatif duyarlılık gücü gibi kriterlerdir. Belirlenen

uygulamalar çerçevesinde bu kriterlerin hangisinin kitap satışları üzerinde daha büyük bir etkiye sahip olduğu araştırılmak istenmektedir.

Ying ve Hanbin (2010) yaptıkları çalışmada, bir üretim işletmesinin 12 aylık geçmiş verisini kullanarak gelecek dönem için tahmin çalışması yapmışlardır. Kullanılan yöntem yapay sinir ağlarından olan çok katmanlı algılayıcıdır. Veriler %70 programın eğitilmesi için %30 ise programın testi için kullanılmıştır. Uygulamada kullanılan fiyat, marka bilinirliği, pazarlama ve hizmet seviyesi, bölge ekonomisi ve politikası gibi bağımsız karar değişkenleri talebi tahmin etmek için kullanılmıştır. Yapılan çalışmanın sonucunda ise tahminin küçük bir hata oranı ile yapılması yapay sinir ağlarının talep tahminindeki başarısını ortaya koymaktadır.

2.2 Talep Tahmininde Karar Ağaçları

Loureiro ve diğ. (2018) yaptıkları çalışmada, derin öğrenme yaklaşımı ile elde edilen satış tahminlerini bir dizi teknikle, yani karar ağaçları, rasgele orman, destek vektör regresyonu, yapay sinir ağları ve doğrusal regresyon ile elde edilenlerle karşılaştırmaktadır. Çalışma sonucunda derin öğrenme kullanan model, moda perakende pazarındaki satışları tahmin etmek için iyi bir performansa sahip olduğu gözlemlenmiştir, ancak dikkate alınan değerlendirme metriklerinin bir kısmı için rastgele orman algoritması (random forest) daha iyi sonuçlar göstermiştir.

Huang ve Pan (2013) yaptıkları çalışmada, işletmelerin operasyonel performanslarının sınıflandırılması konusu üzerinde durmuşlardır. Genel olarak, işletme performansı analizi genellikle finansal tahmin veya kredi derecelendirme modelleri tarafından yapılır. Bu çalışmada, Tayvan borsa ve OTC'de örnek veri olarak listelenen geleneksel endüstri kamu kuruluşlarının 287 özel teşebbüsünden alınan verileri için hem veri madenciliğini hem de yapay zekayı birleştiren hibrit bir metodolojinin tek bir modelin benzersiz gücünden faydalanması önerilmektedir. İlk aşamada ağ giriş değişkenlerini seçmek için geleneksel temel bileşenler analizi gibi veri madenciliği tekniğini kullanılmıştır. İkinci aşamada olasılıksal sinir ağı dahil olmak üzere çeşitli farklı modeller de göz önünde bulundurulmuştur. Bu çalışma olasılıksal sinir ağı modelinin, genetik algoritma tarafından ayarlanan parametreden sonra sınıflandırma yeteneğinin, diğer basit yöntemler-geri-yayılma ağları, karar ağaçları ve lojistik regresyon modelinden önemli ölçüde daha iyi performans gösterdiğini göstermektedir. Sonuç olarak, gerçek veri setleriyle yapılan deneysel sonuçlar, birleştirilmiş modelin, tek bir

modelden birinin elde ettiği tahmin sınıflandırma doğruluğunu iyileştirmenin etkili bir yolu olabileceğini göstermektedir.

Nuaimi ve diğ. (2016) yaptıkları çalışmada, Birleşik Arap Emirlikleri'nde yer alan Al Ain şehrinde yaşayan toplumun ihtiyaçları için (yeni okullar, hastaneler, parklar vb.) belediye yönetiminin karar verme sürecini kolaylaştırmak için ihtiyaç değerlendirme sürecinin nasıl iyileştirileceğini araştırmışlardır. 287 kamu kuruluşunun finansal verilerini kullanarak şirketlerin performans analizlerini yapmak için hem veri madenciliğini hem de yapay zekayı birleştiren hibrit bir metodoloji önermeyi amaçlamaktadır. İlk olarak, ağ giriş değişkenlerini seçmek için geleneksel temel bileşenler analizi gibi veri madenciliği tekniğini kullanılmıştır. Devamında, yerine getirilemeyecek taleplerin nedenlerini olasılıksal sinir ağı da dahil olmak üzere çeşitli farklı modeller de dikkate alınarak araştırılmıştır. Model sonuçlarının ölçüm performansı kıyaslanmıştır. Sonuç olarak, gerçek veri setleriyle yapılan deneysel sonuçlar, birleştirilmiş modelin, tek bir modelden birinin elde ettiği tahmin sınıflandırma doğruluğunu iyileştirmenin etkili bir yolu olabileceğini göstermektedir.

Chaudhuri ve diğ. (2017) yaptıkları çalışmada, hisse senedi fiyatlarının oranlarında gelecekteki hareketi tahmin etmek için makine öğrenimi araçlarını kullanmayı amaçlamaktadırlar. Şirket çiftlerinin hisse fiyatı oranının değerinin tahmini modellemesi için üç farklı algoritma, destek vektör makinesi (SVM), rastgele orman (RF) ve uyarlanabilir nöro-bulanık çıkarım sistemi (anfıs) kullanılmıştır. Bu yazıda, 2012-2015 yılları (Haziran'a kadar) Hero Motocorp-Bajaj Auto, TCS-INFOSYS ve PNB (Pencap Ulusal Bankası), BOB (Baroda Bankası) adlı üç çift Hindistan'da faaliyet gösteren şirketin günlük hisse fiyatları dikkate alınmıştır. Eğitim, her yıl için ayrı ayrı yapılmıştır. Her yıl için verilerin %80'i eğitim amaçlı olarak değerlendirilmiştir ve verilerin %20'si test için kullanılmıştır. Buna göre girdi ve çıktı değişkenlerinin değerleri hesaplanmıştır. SVR, RF ve ANFIS'in hisse senedi fiyatları çiftinin değerini tahmin etme kabiliyeti yıl bazında kontrol edilmiştir. Performansı değerlendirmek için MSE ve MAPE benimsenmiştir. Sonuç olarak rassal orman algoritması ile elde edilen hata değeri en küçük çıkmış olmaktadır.

Liao ve Sun (2010) yaptıkları çalışmada, Çin'de yer alan Chao Gölü'nün su kalitesini tahmin etmek ve değerlendirmek için daha iyi ve daha doğru bilgileri dikkate alarak değiştirilmiş bir karar ağacı yöntemi önermektir. Önerilen yapı geliştirilmiş karar ağacıdır ve bu model geleneksel karar ağacı algoritması olan C4.5 ve yapay sinir ağları

algoritmaları ile karşılaştırılmıştır. Bunun sonucunda da geliştirilmiş karar ağacı algoritmasının diğer yöntemlere nazaran su kalitesi tahmininde daha başarılı sonuç verdiği tespit edilmiştir.

Kang ve diğ. (2017) yaptıkları çalışmada, City Bike adlı bir bisiklet firmasının verilerini kullanarak belirli bir süre içindeki bisiklet kiralanma sayısını tahmin etmek için kullanılacak yöntemi seçmeyi amaçlamaktadırlar. Veri seti sürüş mesafesi, sürüş süresi, istasyon enlemi ve boylamı, kullanıcı tipi vb. dahil 15 değişken içermektedir. Tahmin modelini oluşturmak için başlangıç zamanı, kullanıcı tipi, doğum yılı ve cinsiyet değişkenleri kullanılmıştır. Tahmin algoritması için çok katmanlı algılayıcı, karar ağacı ve rassal orman algoritmaları sonuçları karşılaştırılmıştır. Veriler %70 eğitim ve %30 test verisi olarak rassal bir şekilde bölünmüştür. Değerlendirme kriteri olarak hata ortalamalarının karelerinin karekökü (RMSE) kullanılmıştır. 3 algoritma arasından en iyi sonucu veren algoritma rassal orman algoritması olduğu gözlenmiştir.

Bala (2010), Hindistan'ın doğusundaki bir şehir olan Bhubaneswar'da bulunan bir perakende süpermarket mağazasında satılmakta olan 10 ürün için talep tahmini çalışması yürütmüştür. Perakende mağaza, bilgisayarlı envanter takip sistemi ile orta büyüklüktedir ve çoğunlukla bölgedeki küçük bir lüks alana hitap etmektedir. Bu nedenle, müşterilerin detayları ile veri toplamak mümkün olmuştur. Karar ağacı ve ARIMA'ya dayanan önerilen model mevsimsellik dikkate alınarak uygulanmıştır. "ARIMA" ve "sinir ağı ile ARIMA" temelli diğer çeşitli tahmin modelleri de uygulanmıştır. Haftalık tahmin için, kullanılan öngörücüler ödeme günü sayısı, tatil sayısı (hafta sonu dahil), festival / özel gün sayısı, tatil öncesi gün sayısı, festival öncesi / özel gün sayısı, okul tatili sayısı (yaz / kış) gün sayılarıdır. Önerilen model, satışın önemli bir kısmının belirli bir profile veya profillere güçlü bir şekilde atfedildiği ürün kalemleri için uygun olduğu tespit edilmiştir.

2.3 Talep Tahmininde Kümeleme

Genel olarak, kümeleme benzer nesneleri gruplandırmak için denetimsiz tekniklerin kullanılmasıdır. Veri bilimcinin, kümelere uygulanacak etiketleri önceden belirlememesi anlamında, kümeleme teknikleri denetimsizdir. Verilerin yapısı ilgili nesneleri tanımlar ve nesneleri en iyi nasıl gruplandıracağını belirler. Kümeleme, verilerin analizi için sıklıkla kullanılan bir yöntemdir. Kümeleme mantığının temelinde herhangi bir tahmin amacı güdülmez.

Bunun yerine, kümeleme yöntemleri, nesne özniteliklerine göre nesneler arasındaki benzerlikleri bulur ve benzer nesneleri kümeler halinde gruplandırır (Dietrich ve diğ, 2015).

Huber ve diğ. (2017) yaptıkları çalışmada, bir fırın zincirindeki hizmet düzeyini artırmak ve satılmamış malların miktarını sınırlamak amacıyla sipariş sürecini geliştiren bir karar destek sisteminin parçası olarak makale kümeleme ve hiyerarşik öngörmenin uygulanabilirliğini araştırmışlardır. Çalışmanın odak noktası günlük sipariş edilen çabuk bozulabilen tüketici ürünleridir. Fırın zincirinin satış noktası verilerine dayanan bir değerlendirmede kümeleme yaklaşımını günlük talep tahmini çalışması için kullanmışlardır. En son satış noktası verilerine dayanarak farklı organizasyon düzeylerinde hiyerarşik tahminler sağlayarak günlük işlemleri destekleyen bir karar destek sistemi önerilmiştir. Kümeleme analizi, ikame edilebilir kalemlerin benzer gün içi satış modeline sahip olduğunu ve bu da talebin toplu bir seviyede tahmin edilmesini makul kıldığını ortaya koymuştur.

Demiriz (2014) yaptığı çalışmada hem pozitif hem de negatif öğe ilişkilerini içeren kalem grupları içindeki lineer regresyon modelleri aracılığıyla perakende eşyaların haftalık taleplerinin tahmin edilmesi için bir çerçeve önermiştir. Haftalık satış rakamlarına göre benzer olan madde gruplarına çok değişkenli regresyon modelleri uygulanmıştır. Öğeleri gruplandırmak için temel fikir, benzer öğelerin özelliklerinin (ör. fiyat ve envanter seviyeleri) odak öğesi için daha iyi öngörücü özelliklere sahip olabilmesidir. Kümeleme olarak iki farklı yöntem kullanılmıştır. Bunlar *K*-ortalamalar algoritması ve sınırlandırılmış kümeleme algoritmasıdır. Regresyon ile yapılan talep tahmininde her bir nesne için ayrı ayrı, nesnelerin birbirleriyle olan ilişkileri baz alınarak ve iki kümeleme algoritmasının verdiği hata sonuçları karşılaştırılmıştır ve sınırlandırılmış kümeleme yönteminin talep tahminini diğerlerine nazaran daha az hata payıyla karşılaması üzerine kümeleme yaklaşımının tahmin sistemlerinde iyi bir sonuç verdiğini kanıtlamıştır.

Kusrini (2015) yaptığı çalışmada, bir kümeleme algoritması kullanarak ürünleri 'hızlı tüketilen ürünler' ve ‘yavaş tüketilen ürünler ' olmak üzere iki farklı kategoriye ayırabilecek bir model oluşturarak minimum stok ve kar marjını belirleme sürecini desteklemeyi amaçlamaktadır. Bu çalışmada kullanılan yöntem önceden belirlenmiş öğelerin sınıflandırılmasında başarılı bir şekilde sonuç vermesi nedeniyle, kümeleme yöntemi olan, *K*-ortalamalar algoritması kullanılmıştır. Çalışmada kullanılan veriler

Yogyakarta'da faaliyet gösteren bir mini market olan Citramart'tır. Araştırmada kullanılan veriler, 2014 ve 2015 yıllarına ait satış verilerinden alınmıştır. Kümeleme senaryosu zaman (yıl ve ay) ve değişken (kalem sayısı ve işlem değerleri) kombinasyonu için kurulur. Merkeze en yakın olan grup kümesi elemanları hızlı tüketilen ürünler arasında olmakta ve merkeze en uzak olan grup kümesi elemanlarının yavaş tüketilen ürün grubuna ait olduğu gözlemlenmiştir.

Lingxian ve diğ. (2019) yaptıkları çalışmada, öncelikle İngiltere'nin çevrimiçi perakende aktivite verilerini sınıflandırmak için *K*-ortalamalar kümelemesi adı verilen denetimsiz bir öğrenme yöntemi kullanmış ve daha sonra, geçmiş bilgilere dayanarak gelecekteki çevrimiçi perakende dinamiklerini tahmin etmek için Uzun Kısa Süreli Bellek adlı gelişmiş bir yapay zeka zaman dizisi modeli kullanmışlardır. Veriler normalize edildikten sonra %80 eğitim, %20 test verisi olarak kullanılacak şekilde ikiye ayrılmışlardır. Uzun Kısa Süreli Bellek modelinin gelecek periyottaki 20 saatlik verinin fiyat dinamiklerini tahmin etmek için çoğunlukla son 10 saatlik gerçekleşen veri bilgileriyle hesaplanabileceği, bu da perakendecilerin envanter ve servis personeli gibi kaynaklarını yeniden tahsis etmelerini ve satış planlarını zamanında ayarlayabilmelerini sağladığını ortaya koymuşlardır.

Chen ve diğ. (2017) yaptıkları çalışmada, bilgisayar perakende satış tahminleri için kümelenme ve makine öğrenme yöntemlerini birleştirerek kümelenmeye dayalı bir tahmin modeli önermişlerdir. Önerilen yöntem önce eğitim verilerini benzer özellikler gözetilerek bir gruba ayırmak için kümeleme tekniğini kullanmıştır. Daha sonra, her bir grubun tahmin modelini eğitmek için makine öğrenme teknikleri kullanılmıştır. Test verilerine en çok benzeyen veri örüntülerine sahip küme belirlendikten sonra, satış tahmini için kümenin eğitimli öngörme modeli benimsenmiştir. Bu çalışmada, kendi kendini organize eden harita (SOM), büyüyen hiyerarşik kendi kendini organize eden harita (GHSOM) ve *K*-ortalamalar algoritması olmak üzere üç kümeleme tekniği ve destek vektör regresyonu (SVR) ve aşırı öğrenme makinesi (ELM) olmak üzere iki makine öğrenimi tekniği kullanılmıştır. Toplam altı kümelenmeye dayalı tahmin modeli önerilmiştir. Kişisel bilgisayarlar, dizüstü bilgisayarlar ve likit kristal ekranlar için gerçek satış verileri ampirik örnek olarak kullanılmıştır. Deneysel sonuçlar, büyüyen hiyerarşik kendi kendini organize eden harita ile bir aşırı öğrenme mekanizmasını birleştiren modelin, diğer beş tahmin modelinin yanı sıra tek SVR ve

tek ELM modellerine kıyasla her üç ürün için de üstün tahmin performansı sağladığını göstermiştir.

Kumar ve Patel (2010) yaptıkları çalışmada, her bir tahminle ilişkili standart sapma ile bir dizi öğe için satış tahminleri verildiğinde, kümeleme kavramlarını kullanarak tahminleri birleştirmek için yeni bir yöntem önermişlerdir. Öğe kümeleri, satış tahminlerindeki benzerlik temel alınarak tanımlanır ve her bir öğe kümesi için ortak bir tahmin hesaplanır. Ulusal bir perakende zincirinden elde edilen gerçek bir veri kümesinde, tahminleri birleştirme yönteminin, tek tek tahminlerden veya bir üründeki tüm öğeler için tek bir birleşik tahmin kullanma yönteminden çok daha iyi satış tahminleri ürettiğini saptamışlardır.

Thomassey (2010) yaptığı çalışmada, tekstil perakende sektöründeki tahmin problemini ele almıştır. Yazara göre tekstil perakende sektörünün dinamiği diğer sektörlere nazaran daha farklıdır. Bundan dolayı tahmin çalışması yaparken dikkate alınacak hususlar yazara göre; geçmiş veriler arasındaki (satışlar ve ilgili kalemlerin tanımlayıcı kriterleri arasında) ilişkiyi modellemek ve bu ilişkileri, yeni öğelerin açıklayıcı ölçütlerinden gelecekteki satışları tahmin etmek için kullanmaktır. Thomassey (2010) bu nedenle tahmin problemini, bir zaman dizisi tahmin sorunu değil, bir sınıflandırma sorunu olarak ele almıştır. Yazara göre, böyle bir durumda girdiler (yani tanımlayıcı ölçütler) ve çıktı (yani satışlar) arasındaki ilişkiler karmaşık ve / veya doğrusal olmayan olduğunda, makine öğrenme yöntemleri basit ve yorumlanabilir model sınıflandırma modelleri oluşturmak için en uygun yöntemdir. Tahmin doğruluğunun belirli giyim tedarik zinciri üzerindeki etkisini ölçmek için, bir perakendecinin ve bir üreticinin tedarik sürecini 20 ürünün gerçek verileriyle benzetim çalışması yapılmış ve bu örnekte kullanılan tanımlayıcı kriterler fiyat, satışların başlangıç tarihi ve her bir öğenin kullanım ömrüdür. Bu sistemin homojen madde grupları oluşturmak için çok etkili olduğu kanıtlanmıştır.

Çizelge 2.1 : Literatür taraması yapılmış makalelerin özet tablosu.

Makale	Yazarlar	Yıl	Uygulama Alanı	Kullanılan Yöntem	Faktörler	Özet
An intelligent fast sales forecasting model for fashion products	Yong Yu, Tsan-Ming Choi, Chi-Leung Hui	2011	Perakende Sektörü	ELM-FF model, SLFN (single-hidden layer feedforward neural networks), IFSFM Modeli	Satış Miktarı Ürün Rengi Ürün Bedeni Ürün Fiyatı	Çalışmanın amacı moda sektöründe faaliyet gösteren bir firmanın talep tahmini istatistiksel yöntemlerin kombinasyonu ile yapay zekâ uygulamasının sonuçlarını performans ölçütlerince karşılaştırmaktır.
Intelligent Retail Forecasting System for New Clothing Products Considering Stock-out	He Huang, Qiurui Liu	2017	Perakende Sektörü	YSA, ANFIS	Orijinal Fiyat, Promosyon Seviyesi Beden Tercihi Dağıtım Takvim Faktörü	Çalışmada veri kümelerinin satış ve talep modellerini gözlemleyerek, önerilen sistem ile kayıp satışların azaltılması amaçlanmıştır.
Intelligent demand forecasting systems for fast fashion	Brahmadeep S.Thomassey	2016	Perakende Sektörü	Karar Ağacı (C4.5) Algoritması ELN Modeli	Ürün Fiyatı Ürün Satışlarının Başlangıç Zamanı Ürün Yaşam Ömürleri	Çalışmada, kısa dönem ve uzun dönem talebi makine öğrenme yöntemlerinden yapay sinir ağları modeli, kümeleme ve rassal orman algoritmaları kullanılarak tahmin edilmeye çalışılmıştır.

Çizelge 2.1 (devam) : Literatür taraması yapılmış makalelerin özet tablosu.

Makale	Yazarlar	Yıl	Uygulama Alanı	Kullanılan Yöntem	Faktörler	Özet
Using artificial neural networks to improve decision making in apparel supply chain systems	P.C.L. Hui, T.-M. Choi	2016	Perakende Sektörü	Çok Katmanlı Algılayıcı (MLP)	Ürün Yaşam Ömrü Satış Periyodunun Uzunluğu	İncelemede bir firmanın satış tahminle yapması ve ürünler için uygun stok seviyesini belirlemek için yapay zekanın ne derece başarılı olabileceğini teorik olarak ele alınmıştır.
Forecasting New Apparel Sales Using Deep Learning and Nonlinear Neural Network Regression	Chandadevi Giri, Sebastien Thomassey, Jenny Balkow and Xianyi Zeng	2019	Perakende Sektörü	V3 derin sinir ağı Çok katmanlı Algılayıcı	Ürün Rengi Ürün Stili Ürün Tasarımı Ürün Bedeni	Çalışmada, belirli bir ürün grubunun geçmiş satış verilerini ve satılan ürünlerin görsel bilgilerini kullanarak bir sonraki periyot için talep tahmini yapılması amaçlanmıştır.
Research of Artificial Intelligence in the Retail Management Problems	Viktor P. Semenov Vladimir Chernokulsky Natalya V. Razmochaeva	2017	Perakende Sektörü	Çok Katmanlı Algılayıcı (MLP)	Satılan ürün sayısı Satış ve satın alma fiyatları Makinedeki ürün sayısı	Çalışmanın amacı, üretilmesi planlanan ve üretildiği halde maksimum kar getirecek bir otomat makinesinin talep tahmininin yapılmasıdır.

Çizelge 2.1 (devam) : Literatür taraması yapılmış makalelerin özet tablosu.

Makale	Yazarlar	Yıl	Uygulama Alanı	Kullanılan Yöntem	Faktörler	Özet
Configuration of Daily Demand Predicting System Based on Neural Network	Ilham SLIMANI Ilhame EL FARISSI	2016	Perakende Sektörü	Çok Katmanlı Algılayıcı (MLP)	Satılan Ürün Sayısı Ürün Tipi (çikolata, ekmek, peynir vs.) Müşteri Sayısı vb.	İncelemede, Fas'ta bulunan bir süpermarketin günlük talep tahminleri üretebilmek için yapay sinir ağları kullanılmıştır.
Analysis of book sales prediction at Amazon marketplace in India: a machine learning approach	Satyendra Kumar Sharma Swapnajit Chakraborti Tanaya Jha	2019	Perakende Sektörü	Karar Ağacı (C5.0) Algoritması CART ANN	Güncel fiyat İndirim tutarı İndirim oranı Sayfa sayısı Olumlu duygu gücü Pozitif duyarlılık gücü vb.	Çalışmada makine öğrenmesi tekniklerinin e-ticaret platformlarındaki satışları tahmin etmede etkinliğini göstermeyi amaçlamaktadır.
Study on the Model of Demand Forecasting Based on Artificial Neural Network	Zhu Ying Xiao Hanbin	2010	Perakende Sektörü	Çok Katmanlı Algılayıcı (MLP)	Fiyat Marka Bilinirliği Pazarlama ve Hizmet Seviyesi Bölge Ekonomisi ve Politikası	İncelemede bir üretim işletmesinin geçmiş sipariş verisi kullanılarak gelecek dönem için tahmin çalışması çok katmanlı algılayıcı ile yapılmıştır.

Çizelge 2.1 (devam) : Literatür taraması yapılmış makalelerin özet tablosu.

Makale	Yazarlar	Yıl	Uygulama Alanı	Kullanılan Yöntem	Faktörler	Özet
Exploring the use of deep neural networks for sales forecasting in fashion retail	A.L.D. Loureiro V. L. Miguéis Lucas F. M. da Silva	2018	Perakende Sektörü	Karar Ağaçları Rasgele Orman, Destek Vektör Regresyonu, Yapay Sinir Ağları ve Doğrusal Regresyon	Ürün Ailesi Ürün Alt Ailesi Renk Tipi Renk, Moda Mağaza Tipi, Fiyat, Beden, Beklenti Seviyesi	Araştırmada, derin öğrenme yaklaşımı ile elde edilen satış tahminlerini bir dizi makine öğrenmesi tekniği ile karşılaştırarak en iyi sonucu veren yöntemin seçilmesi amaçlanmıştır.
Forecasting classification of operating performance of enterprises by probabilistic neural network	Jui-Ching Huang Wen-Tsao Pan	2013	İşletme Performans Analizi	Temel Bileşen Analizi (PCA), Olasılıksal Sinir Ağı, Karar Ağacı	Alacak Hesabının Ciro Oranı, Envanter Devir Hızı, Varlığın Ciro Oranı, Sabit Kıymet Büyüme Oranı, Toplam Aktif Büyüme Oranı, Öz kaynak Oranı, Sabit Oran, Kazanç Oranı	Yapılan çalışmanın amacı, işletmelerin operasyonel performanslarının sınıflandırılması konusuna hibrit bir yöntem ile yaklaşmaktır.

Çizelge 2.1 (devam) : Literatür taraması yapılmış makalelerin özet tablosu.

Makale	Yazarlar	Yıl	Uygulama Alanı	Kullanılan Yöntem	Faktörler	Özet
Decision Tree Based Demand Forecasts for Improving Inventory Performance	Pradip Kumar Bala	2010	Perakende Sektörü	ARIMA	Tatil/ Festival/ Özel gün gün sayısı (hafta sonu dahil), Tatil öncesi gün sayısı, Okul tatili (yaz / kış) gün sayısı, Mega indirimlerin yapıldığı gün sayısı, Ürünleri alan müşteri tipleri vb.	İncelemede bir süpermarkette günlük tüketilen ürünleri hedef alan ve bu ürünler için kısa dönemli talep tahminini makine öğrenmesi yöntemleri ile tespit edilmeye çalışılmıştır.
Forecasting and Evaluating Water Quality of Chao Lake based on an Improved Decision Tree Method	Hao Liaoa Wen Suna	2010	Çevre ve Teknoloji	Geliştirilmiş Karar Ağacı Modeli ve C4.5	DO, BOD5, PH, sıcaklık, Amonyak-Azot	Çalışmada, Çin'de yer alan bir gölün su kalitesini tahmin etmek ve değerlendirmek için geliştirilmiş bir karar ağacı yöntemi önerilmektir.
Research on the Forecast of Shared Bicycle Rental Demand Based on Spark Machine Learning Framework	Zilu Kang Yuting Zuo Zhibin Huang Feng Zhou Penghui Chen	2017	Perakende Sektörü	Çok Katmanlı Algılayıcı, Karar ağacı Algoritması, Rassal Orman Algoritması	Sürüş Mesafesi, Sürüş Süresi, İstasyon Enlemi ve Boylamı, Kullanıcı tipi, Cinsiyet, Doğum Yılı	Çalışmada bir bisiklet firmasının belirli bir süre içindeki bisiklet kiralanma sayısını tahmin etmek için kullanılacak yöntem seçiminin belirlenmesi amaçlanmıştır

Çizelge 2.1 (devam) : Literatür taraması yapılmış makalelerin özet tablosu.

Makale	Yazarlar	Yıl	Uygulama Alanı	Kullanılan Yöntem	Faktörler	Özet
Predicting the Decision for the Provision of Municipal Services Using Data Mining Approaches	Eiman Al Nuaimi Samira Al Marzooqi Nazar Zaki	2016	Hizmet Sektörü	Karar Ağacı, Destek Vektör Makinesi (SVM), Sinir Ağları (KNN),	Vatandaş İhtiyaçları Emirlik Dışı Toplam Nüfus Emirlik Nüfusu İşin Kapsamı Yapılan İşlemler Yıl	İncelemede bir şehirdeki toplumun ihtiyaçları için ihtiyaç değerlendirme sürecinin nasıl iyileştirileceğini konusuna yapay sinir ağları ile çözüm bulunması amaçlanmıştır.
Application of Machine Learning Tools in Predictive Modeling of Pairs Trade in Indian Stock Market	Tamal Datta Chaudhuri Indranil Ghosh Priyam Singh	2017	Finans	Destek Vektör Makinesi (SVM), Rastgele Orman (RF) Uyarlanabilir Nöro-Bulanık Çıkarım Sistemi	Zamanla Fiyatların Oranı, Fiyatlardaki Hız Oranı, Standart Sapma	Araştırmada bir şirketin günlük hisse fiyatları dikkate alınarak, hisse senedi fiyatlarının oranlarında gelecekteki hareketi tahmin etmek için makine öğrenimi araçlarını kullanmayı amaçlanmıştır.
Cluster-based hierarchical demand forecasting for perishable goods	Jakob Huber, Alexander Gossmann Heiner Stuckenschmidt	2017	Perakende Sektörü	Makale Kümeleme, ARIMA	Bölge Bölge Kategorisi Bölgede Günlük Satılan Ürün Sayısı Mağaza Mağaza Kategorisi Her Mağazada Günlük Satılan Ürün Sayısı	Çalışmada bir fırının sunduğu hizmet düzeyini artırmak ve gün sonunda satılamamış malların miktarını sınırlamak amacıyla makale kümeleme ve hiyerarşik öngörmenin uygulanabilirliğini araştırılmıştır.

Çizelge 2.1 (devam) : Literatür taraması yapılmış makalelerin özet tablosu.

Makale	Yazarlar	Yıl	Uygulama Alanı	Kullanılan Yöntem	Faktörler	Özet
Grouping of Retail Items by Using *K*-Means Clustering	Kusrini Kusrini	2015	Perakende Sektörü	*K*-ortalamalar algoritması	Bir Yıllık Satış Verileri, Bir Aylık Satış Verileri	Çalışmada, bir mini market verileri kullanılarak ürünleri hızlı/yavaş tüketilen ürünler olmak üzere iki farklı kategoriye ayırabilecek bir model oluşturarak minimum stok ve kâr marjını belirleme sürecini desteklemeyi amaçlamaktadır.
Demand Forecasting Based on Pairwise Item Associations	Ayhan Demiriz	2014	Perakende Sektörü	*K*-ortalamalar algoritması, sınırlandırılmış kümeleme algoritması	Fiyat, Envanter Seviyeleri, Haftalık Satış Rakamları	Çalışmada lineer regresyon modelleri aracılığıyla perakende eşyaların haftalık taleplerinin tahmin edilmesi için bir çerçeve önermiştir.
Online Retail Sales Prediction with Integrated Framework of *K*-means and Neural Network	You Lingxian Kou Jiaqing Wang Shihuai	2019	Perakende Sektörü	*K*-ortalamalar algoritması, Uzun Kısa Süreli Bellek Algoritması	Fatura No, Stok kodu, Açıklama, Miktar, Tarih, Birim fiyat, Müşteri Kimliği, Ülke	İncelemede, İngiltere'de gerçekleşen online satış verilerini tahmin etmek için gelişmiş bir yapay zekâ zaman dizisi modeli kullanımı amaçlanmıştır.

Çizelge 2.1 (devam) : Literatür taraması yapılmış makalelerin özet tablosu.

Makale	Yazarlar	Yıl	Uygulama Alanı	Kullanılan Yöntem	Faktörler	Özet
Sales forecasting by combining clustering and machine-learning techniques for computer retailing	I-Fei Chen Chi-Jie Lu	2017	Perakende Sektörü	SOM, GSOM, *K*-ortalamalar algoritması Destek Vektör Reg, ELM	İki Haftalık PC Satış Miktarı, LCD Satış Miktarı NB Satış Miktarı	Araştırmada, bilgisayar perakende satış tahminleri için kümelenme ve makine öğrenme yöntemlerini birleştirerek kümelenmeye dayalı bir tahmin modeli önerilmiştir.
Using clustering to improve sales forecasts in retail merchandising	Mahesh Kumar Nitin R. Patel	2010	Perakende Sektörü	hClust Algoritması, cClust Algoritması, Box-Jenkins Algoritması, Hareketli Ortalama, Üssel Düzgünleştirme	Ürün Numarası (Merchandise ID) Haftalık Satış Miktarı	Yapılan çalışmada, ulusal bir perakende zincirinin satış verileri kullanılarak satış tahminleri ve standart sapmaları benzer olan ürünleri kümelemek gibi yeni bir yöntem önermişlerdir.
Sales forecasts in clothing industry: The key success factor of the supply chain management	Sebastien Thomassey	2010	Perakende Sektörü	Holt winter with seasonality ARMAX Forecast Pro yazılımı FIS temelli uzun vadeli tahmin modeli	Ürün Fiyatları, Tatil Günleri Ürün Periyodu (Halloween, Christmas, Back to school)	Yapılan çalışmada, tekstil perakende sektöründeki tahmin problemi ele alınmıştır. Tahmin problemi, bir zaman dizisi tahmin sorunu değil, bir sınıflandırma sorunu olarak ele alınmıştır.

3. TALEP TAHMİNİ İÇİN KULLANILAN MAKİNE ÖĞRENMESİ YÖNTEMLERİ

3.1 Regresyon

3.1.1 Yapay sinir ağları

Yapay sinir ağları insan beynindeki nöronların öğrenme sürecini simüle etmek için tasarlanmış bilgisayar tabanlı sistemlerdir (Kwon, 2011). YSA'lar son on yılda öngörücü modeller ve örüntü tanıma olarak büyük ilgi görmektedir. Yapay sinir ağları, sistemi yöneten fiziksel ve kimyasal yasalar hakkında gerçek bilgiye sahip olmadan bir dizi deneysel veriden "öğrenme" yeteneğine sahiptir. Bu nedenle, veri sistemlerinin YSA uygulaması özellikle sistemlerin doğrusal olmamaları ve karmaşık davranış gösterdiği yerlerde yüksektir. YSA'ların modelleme ve simülasyondaki erken uygulama örnekleri ilk kez 1943'te basitleştirilmiş nöronların McCulloch ve Pitts tarafından piyasaya sürülmesinden sonra rastlanıldı. Bu nöronlar biyolojik nöronların modelleri ve hesaplama görevlerini yerine getirebilecek devreler için kavramsal bileşenler olarak sunuldu. YSA modelleri, yaşayan canlıların beyninin problemleri çözmede nasıl geliştiğini inceleyen biyolojik bilimlerden ilham almıştır (Kwon, 2011). Son yıllarda sinir ağı teknolojisinde büyük adımlar atıldı. Bu atılım, çok çeşitli bilimsel uygulamalar üzerinde artan araştırmalarla kendini göstermektedir. Yapay sinir ağlarına ilgi, bu konuda çalışan bilim insanlarının sayısına, bu konuya ilişkin çalışmalar için ayrılan finansman miktarlarına, bu konuya ilişkin düzenlenen büyük konferansların sayısına ve sinir ağlarıyla ilişkili makale içeren dergi sayısına yansımaktadır. Yapay sinir ağları, örüntü tanıma, tanımlama, sınıflandırma, konuşma, görme ve kontrol sistemleri gibi çeşitli uygulama alanlarında karmaşık işlevler yerine getirecek şekilde eğitilmiştir.

3.1.1.1 Yapay sinir ağı topolojisi

Güvenilir ve sağlam bir ağ elde etmedeki başarı, ilgili süreç değişkenlerinin seçimine, mevcut veri kümesine ve eğitim amaçlı kullanılan alana bağlıdır. Bu bölümde, transfer fonksiyonları, öğrenme süreçleri ve eğitim algoritmaları dahil yapay sinir ağlarının topolojisini açıklanacaktır. Birkaç çeşit yapay sinir ağı vardır. İki popüler YSA, yaygın olarak kullanılan geri yayılım algoritması tarafından eğitilmiş çok katmanlı ileri beslemeli sinir ağı ve kendi kendini organize eden haritalamadır. Her ağ, katmanlar halinde gruplandırılmış ve paralel bağlantılarla birbirlerine göre yerleştirilmiş yapay nöronlardan oluşur. Bu ara bağlantıların gücü, kendileriyle ilişkili ağırlık ile belirlenir. Her YSA için ilk katman giriş katmanını (bağımsız değişkenler) ve son katman çıkış katmanını (bağımlı değişkenler) oluşturur. Aralarında gizli katmanlar adı verilen bir veya daha fazla nöron katmanı bulunabilir. Yapay bir sinir ağının topolojisi, katmanlarının sayısı, her katmandaki düğüm sayısı ve transfer fonksiyonlarının doğası ile belirlenir. YSA topolojisinin optimizasyonu muhtemelen bir modelin geliştirilmesinde en önemli adımdır.

Bir sinir ağı, genellikle üç katman halinde düzenlenmiş nöronlardan oluşur: Giriş katmanı olarak adlandırılan ilk katman, analize konu kaynaklardan veri alır. İkinci katman, gizli katman ve sinir ağı için içsel bir değerdir ve dış ile doğrudan teması yoktur. Gizli katmandaki nöronların aktivasyon fonksiyonları genellikle doğrusal değildir; ancak, sınırlı bir kural yoktur. Gizli katmanların sayısı ve gizli katmanlardaki nöronların sayısı önceden tanımlanmamıştır ve ayarlanmalıdır. Genel olarak, orta boyutlu gizli katmanlarla başlamak daha iyidir. Üçüncü katman çıktı katmanı olarak adlandırılır.

İlk katmana girilen verilerin derlenmesinden sonra ağ tarafından elde edilen sonuçları verir. Giriş ve çıkış nöronlarının sayısı, tahmini olarak kullanılan değişkenlerin sayısını ve tahmin edilecek değişkenlerin sayısını etkin bir şekilde temsil eder. Gizli katmanlar özellik detektörleri gibi davranır ve teoride birden fazla gizli katman olabilir.

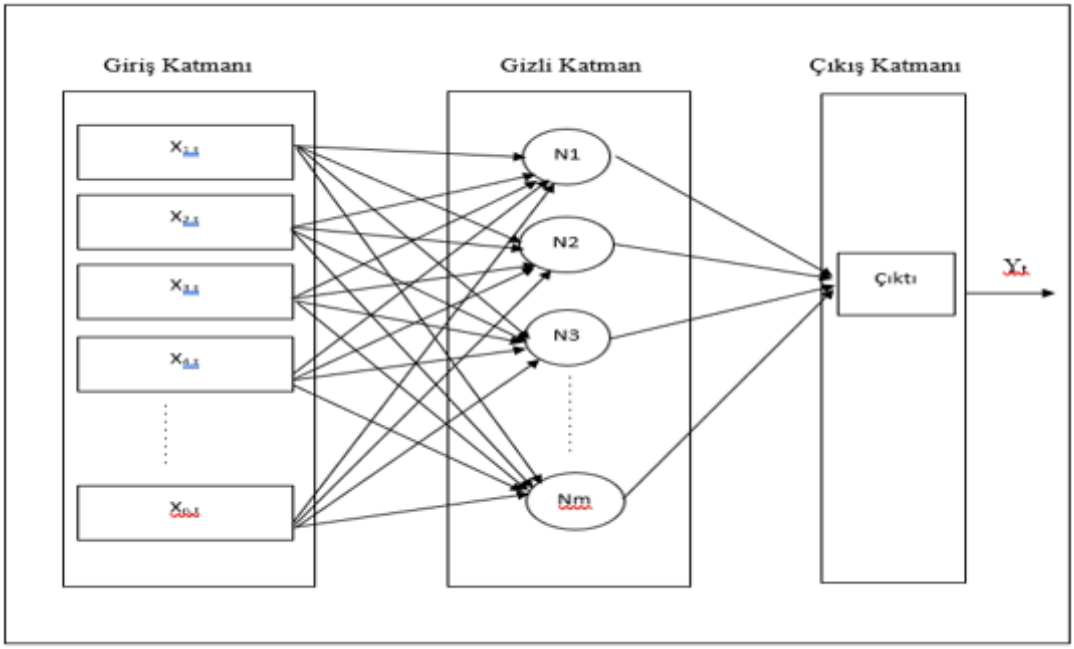
Bununla birlikte, evrensel yaklaşım teorisi, yeterince çok sayıda nöron içeren tek bir gizli katmana sahip bir ağın herhangi bir girdi-çıktı yapısını yorumlayabileceğini düşündürmektedir. Gizli katmandaki nöronların sayısı, sinirsel tahminlerde istenen doğrulukla belirlenir.

Bu nedenle, sinir ağı tasarımı için bir parametre olarak düşünülebilir. İleri beslemeli sinir ağında, belirli bir katmanın tüm nöronları, yanındaki katmanın tüm nöronlarına bağlanır.

Nöronların giriş katmanı bir distribütör gibi davranır ve bu katmana giriş doğrudan gizli katmana iletilir. Gizli ve çıktı katmanlarına girişler, önceki katmandan alınan tüm girdilerin ağırlıklı bir toplamı gerçekleştirilerek hesaplanır. Girdilerin ağırlıklı toplamı, dönüştürüldüğü gizli nöronlara aktarılır (Kwon, 2011). Gizli ve çıktı katmanlarına girişler, önceki katmandan alınan tüm girdilerin ağırlıklı bir toplamı gerçekleştirilerek hesaplanır.

Girişlerin ağırlıklı toplamı gizli nöronlara aktarılır ve burada bir aktivasyon fonksiyonu kullanılarak dönüştürülür. Gizli nöronların çıktısı, sırayla, başka bir dönüşüme uğradığı nöronların çıktısı için girdi görevi görür.

Şekil 3.1 *n* nöron içeren tek bir gizli katmanı olan tipik bir ileri beslemeli sinir ağını göstermektedir. Şekil 3.1'de belirtildiği gibi, her bir yapay nöron, önceki katmanın nöronlarından bir dizi sinyal alan bir temel işlemcidir ve bu girişlerin her birinin, karşılık gelen nöronlar arasındaki bağlantıların gücünü temsil eden ilişkili bir ağırlığa sahiptir.



Şekil 3.1 : Tek gizli katmana sahip tipik bir ileri beslemeli sinir ağı, Kwon (2011).

3.1.1.2 Yapay sinir ağları öğrenme süreci

Bir sinir ağının istenilen özellikleri arasında, ortamından öğrenme yeteneği, ağ performansını artırmak için kesinlikle en önemlisidir.

Öğrenme, bir ağın parametrelerini değiştiren dinamik ve yinelemeli bir süreçtir. Bu, ağın ortamından aldığı sinyallere bir yanıttır.

Çoğu topolojide öğrenme, sinaptik verimlilikte bir değişime veya diğer bir deyişle, bir katmanın nöronlarını bir diğerine bağlayan ağırlıkların miktarında bir değişikliğe yol açar.

$$\Delta w_{i,j}\,(t) = w_{i,j}(t+1) - w_{i,j}(t) \tag{3.1}$$

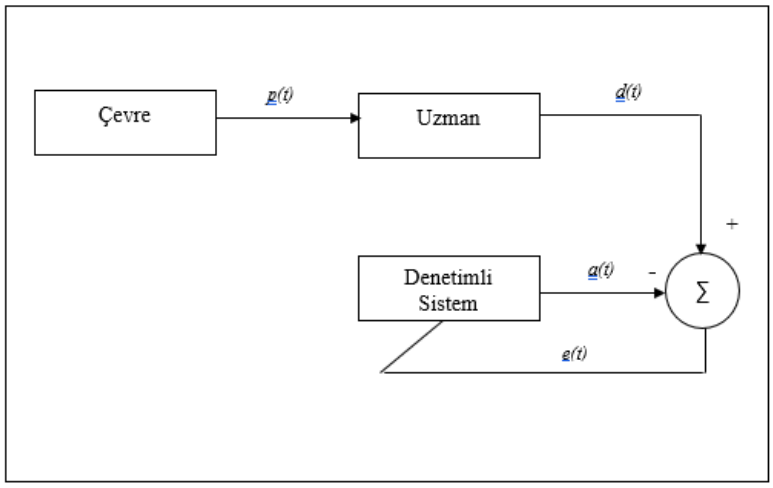
Ve sonra,

$$w_{i,j}(t+1) = w_{i,j}(t) + \Delta w_{i,j}(t) \tag{3.2}$$

Denklem 3.1 ve Denklem 3.2'de $w_{i,j}\ (t+1)$ ve $w_{i,j}\ (t)$ sırasıyla $w_{i,j}$'nin yeni ve eski ağırlık miktarlarını temsil eder. Bu tür bir adaptasyon sürecine ulaşmak için iyi tanımlanmış bir dizi kurala sinir ağının öğrenme algoritması denir (Kwon, 2011).

Denetimli öğrenme

Sinir ağının evrimleştiği ortam hakkında derin bilgiye sahip bir "uzmanın" varlığı ile karakterizedir. Uygulamada, uzman bilgisi, p_i'yi ifade eden ((p_1, d_1), (p_2, d_2)..., (P_Q, d_Q)) belirtilen Q çift giriş / çıkış veri seti biçimindedir. Her çift (p_i, d_i) ağın belirli bir giriş için hedef üretmesi gereken bir duruma karşılık gelir. Bu nedenle, denetimli öğrenme aynı zamanda örneklerle öğrenme olarak tanımlanmıştır. Denetimli öğrenme Şekil 3.2'de şematik olarak gösterilmiştir. Ortam ağ tarafından bilinmemektedir. Bu hem uzmana hem de ağa yönlendirilen bir $p(t)$ girdisi üretir. İçsel bilgisi sayesinde, uzman bu girdi için istenen bir çıktı $d(t)$ üretir. Bu cevabın optimal olduğu varsayılmaktadır. Daha sonra ağın çıkışı ile karşılaştırılır ve tekrar eden bir işlem yoluyla davranışını değiştirmek için ağa yeniden enjekte edilecek olan bir $e(t)$ hata sinyali üretilir. Bu yineleme sonunda ağın uzmanın yanıtını simüle etmesine izin verir. Başka bir deyişle, uzman tarafından çevre bilgisi belli bir durma ölçütüne ulaşıncaya kadar yavaş yavaş ağa aktarılır. Daha sonra uzman ortadan kaldırabilir ve ağın bağımsız çalışmasına izin verilebilir (Kwon, 2011).



Şekil 3.2 : Denetimli öğrenme şeması, Kwon (2011)'den uyarlanmıştır.

Takviyeli öğrenme

Denetimli öğrenmenin bazı sınırlamalarının üstesinden gelir. Bu öğrenme bir tür denetimli öğrenmedir, ancak skaler hata sinyali vektörü yerine bulguları tatmin edici bir ipucu ile vermektedir. Uygulamada, takviyeli öğrenmenin kullanımı karmaşıktır. Ancak, bu tür öğrenme ile denetimli öğrenme arasındaki farkı anlamak önemlidir.

Denetimli öğrenmenin sadece performans indeksini (örneğin ortalama kare hatası) hesaplamakla kalmayıp aynı zamanda sinaptik ağırlıkların adaptasyonu için bir yönü belirten yerel eğimi tahmin eden bir hata sinyali vardır. Bu bilgi tüm farkı yaratan uzman görüşü tarafından sağlanır. Takviye öğrenmede, bir hata sinyalinin olmaması, bu eğimin hesaplanmasını imkânsız hale getirir.

Değişim ölçüsünü tahmin etmek için, ağ eylemleri denemeye ve sonucu gözlemlemeye zorlanır, sonunda sinaptik ağırlıklar için bir değişiklik yönü çıkarır. Bunu yapmak için, memnuniyet endeksi tarafından sunulan ödülü geciktirirken bir deneme / yanılma süreci uygular. Böylece, iki farklı aşama sunulur: ağın rastgele değişim yönleri denediği bir araştırma aşaması ve karar verdiği bir araştırma aşaması. Bu iki aşamalı süreç öğrenmeyi önemli ölçüde yavaşlatabilir. Ayrıca, farklı eylemlerin esası hakkında daha önce öğrenilen bilgileri kullanma arzusu ile bu kararların sonuçları hakkında yeni bilgiler arasında bir ikilem ortaya koymaktadır (Kwon, 2011).

Denetimsiz öğrenme

Denetimsiz veya kendi kendini organize eden olarak adlandırılan bu öğrenme biçimi denetimli durumda olduğu gibi bir sinyal hatası ya da bir memnuniyet endeksi içermemesi ile karakterize edilir. Bu nedenle, girdi sağlayan bir ortam vardır ve bir ağ harici müdahale olmadan öğrenmelidir. Ortamdan gelen girdiyi iç durumunun açıklamasına asimile ederek, bir ağın ana görevi bu durumu olabildiğince iyi modellemektir. Bunu başarmak için, önce bu model için bir kalite ölçüsü tanımlamalı ve daha sonra ağdaki serbest parametreleri, yani sinaptik ağırlıkları optimize etmek için kullanılmalıdır.

Denetimsiz öğrenme tipik olarak nöronların sinaptik ağırlıklarının girdilerin karşılık gelen hedeflerini temsil ettiği bir model oluşturmak için rekabetçi bir sürece dayanır.

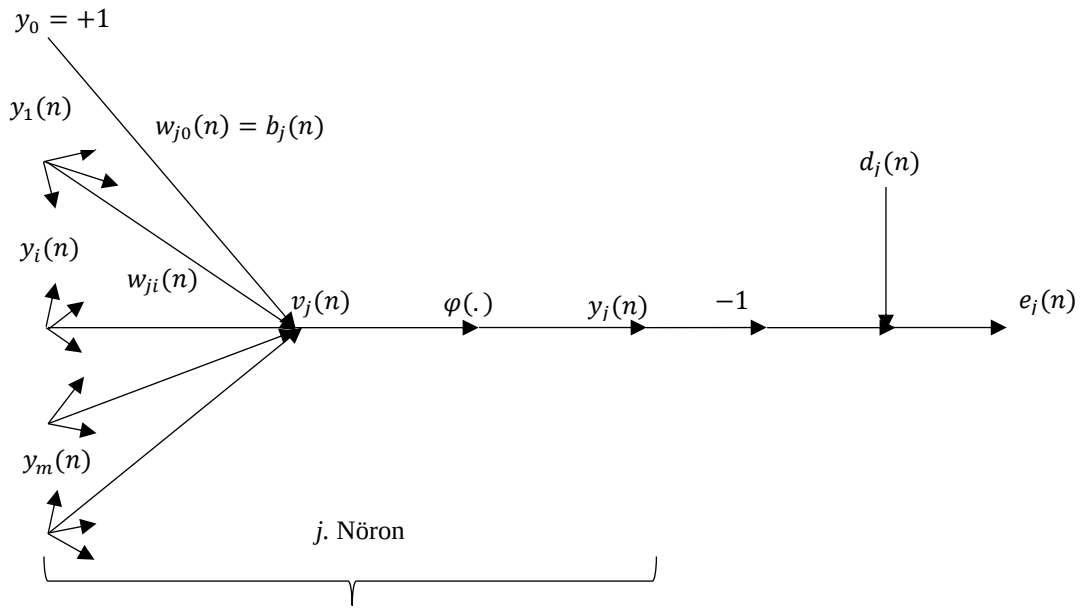
Ortaya çıkan modelin kalitesi, girdiler ve hedefleri arasındaki mesafeyi ölçmek için bir yöntem kullanılarak tahmin edilmelidir. Ortamdan gelen girdiyi iç durumunun açıklamasına uyarlayarak, bir ağın ana görevi bu durumu olabildiğince iyi modellemektir. Bunu başarmak için, önce bu model için bir kalite ölçüsü tanımlamalı ve daha sonra ağdaki serbest parametreleri, yani sinaptik ağırlıkları optimize etmek için kullanılmalıdır.

Öğrenme sürecinin sonunda ağ, özelliklerini kodlamaya izin veren çevresel varyasyonların dahili temsillerini oluşturma ve bu nedenle otomatik olarak benzer cevap sınıfları yaratma yeteneği geliştirmiştir. Denetimsiz öğrenme tipik olarak nöronların sinaptik ağırlıklarının girdilerin karşılık gelen hedeflerini temsil ettiği bir model oluşturmak için rekabetçi bir sürece dayanır. Ortaya çıkan modelin kalitesi, girdiler ve hedefleri arasındaki mesafeyi ölçmek için bir yöntem kullanılarak tahmin edilmelidir.

3.1.1.3 Geri yayılma algoritması çalışma prensibi

Geri yayılım (BP) algoritması 1986 yılında Rumelhart, Hinton ve Williams tarafından ağırlıkların ayarlanması ve dolayısıyla çok katmanlı algılayıcıların eğitimi için önerilmiştir (Graupe, 2013).

Geri yayılma algoritması, Widrow-Hoff öğrenme kuralının çok katmanlı ağlara ve doğrusal olmayan ayırt edilebilir transfer fonksiyonlarına genelleştirilmesiyle oluşturuldu (Kwon, 2011). Giriş vektörleri ve karşılık gelen hedef vektörler, bir işleve yaklaşana kadar, giriş vektörlerini belirli çıkış vektörleriyle ilişkilendirene kadar ağı eğitmek için kullanılır. Standart geri yayılım, ağ ağırlıklarının performans işlevinin eğimi negatifi boyunca hareket ettirildiği bir algoritmadır. Geri yayılım terimi, eğimin doğrusal olmayan çok katmanlı ağlar için hesaplanma biçimini belirtir. Uygun şekilde eğitilmiş geri yayılım ağları, daha önce hiç görmedikleri girdilerle sunulduğunda makul yanıtlar verme eğilimindedir. Tipik olarak, yeni bir giriş, eğitimde kullanılan yeni girişe benzer giriş vektörleri için doğru çıkışa benzer bir çıkışa yol açar. Bu genelleme özelliği, bir ağı temsili bir girdi / hedef çifti kümesi üzerinde eğitmeyi ve ağı olası tüm girdi / çıktı çiftleri üzerinde eğitmeden iyi sonuçlar almayı mümkün kılar. Geri yayılım öğreniminin en basit uygulaması, ağ ağırlıklarını ve sapmalarını, performans işlevinin en hızlı şekilde azaldığı yönde, eğiminin negatifini günceller.



Şekil 3.3 : j. çıkış nöronunun sinyal-akış grafiği, Haykin (2019)'dan uyarlanmıştır.

Bu algoritmayı tanımlamak için, nöron *j*'nin soluna bir nöron tabakası tarafından üretilen bir dizi fonksiyon sinyali tarafından beslendiğini gösteren yukarıdaki Şekil 3.3 kullanılır. Nöron *j* ile ilişkili aktivasyon fonksiyonunun girişinde oluşturulmuş olan lokal alan $v_j(n)$ aşağıdaki Denklem 3.3'te gösterilmiştir.

$$v_j(n) = \sum_{i=0}^{m} w_{ji}(n) y_i(n) \tag{3.3}$$

Burada *m*, nöron *j*'ye gelen toplam girdi (sapma hariç) sayısıdır. Sinaptik ağırlık w_{j0} (sabit giriş $y_0 = +1$'e karşılık gelir), nöron *j*'ye uygulanan sapma b_j'ye eşittir. Bu nedenle, *n*. yinelemede *j* nöronu aktivasyon fonksiyonu çıkışında görünen $y_j(n)$ fonksiyon sinyali Denklem 3.4'te ki gibi olmaktadır.

$$y_i(n) = \varphi_j(v_j(n)) \tag{3.4}$$

Geri yayılım algoritması, sinaptik ağırlıktaki $w_{ji}(n)$ 'ye kısmi türevle orantılı bir düzeltme $\Delta w_{ji}(n)$ uygular. Analizin zincir kuralına göre, bu gradyan Denklem 3.5'te ki şekilde ifade edebilir:

$$\frac{\partial \varepsilon(n)}{\partial w_{ji}(n)} = \frac{\partial \varepsilon(n)}{\partial e_j(n)} \frac{\partial e_j(n)}{\partial y_j(n)} \frac{\partial y_j(n)}{\partial v_j(n)} \frac{\partial v_j(n)}{\partial w_{ji}(n)} \tag{3.5}$$

Kısmi türev $\frac{\partial \varepsilon(n)}{\partial w_{ji}(n)}$, sinaptik ağırlık w_j değeri için ağırlık uzayında arama yönünü belirleyen bir duyarlılık faktörünü temsil eder. Denklemlerin her iki tarafının da diferansiyeli alındığında $e_j(n)$ Denklem 3.6'da ki şekilde elde edilir.

$$\frac{\partial \varepsilon(n)}{\partial e_j(n)} = e_j(n) \tag{3.6}$$

Diferansiyeli alma işlemi bir kez daha uygulandığında $v_j(n)$ Denklem 3.7'de ki şekilde elde edilir.

$$\frac{\partial y_j(n)}{\partial v_j(n)} = \varphi_j'(v_i(n)) \tag{3.7}$$

Son olarak denklemin sol tarafı,

$$\frac{\partial \varepsilon(n)}{\partial w_{ji}(n)} = -e_j(n)\varphi'_j(v_i(n))y_i(n) \tag{3.8}$$

$w_{ji}(n)$ 'ye uygulanan $\Delta w_{ji}(n)$ düzeltmesi delta kuralı tarafından tanımlanır, Denklem 3.8'de kullanılan η geri yayılma algoritmasındaki öğrenme parametresidir.

$$\Delta w_{ji}(n) = -\eta \frac{\partial \varepsilon(n)}{\partial w_{ji}(n)} \tag{3.9}$$

Denklem 3.9'da – işaretinin kullanım sebebi ise ağırlık uzayındaki eğim inişini açıklamaktadır. Denklem 3.8'in Denklem 3.9 içinde kullanılması değişim ölçüsünün $\delta_j(n)$ olarak nitelendirildiği Denklem 3.10'u oluşturur

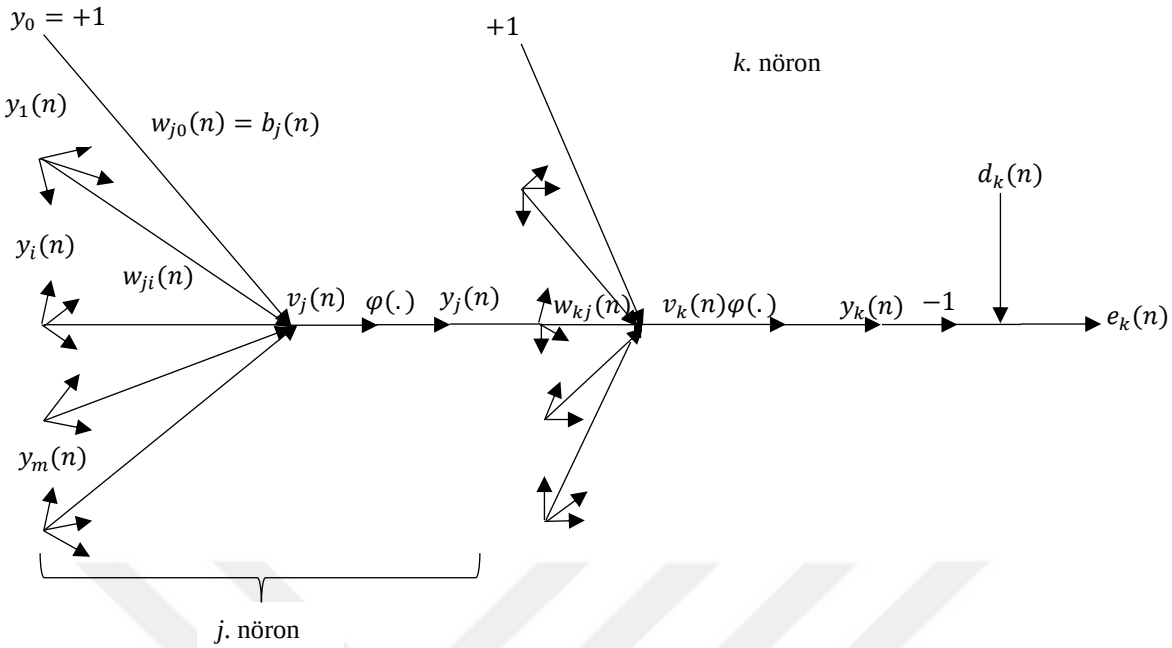
$$\Delta w_{ji}(n) = \eta \delta_j(n) * y_i(n) \tag{3.10}$$

$$\delta_j(n) = \frac{\partial \varepsilon(n)}{\partial v_j(n)} = \frac{\partial \varepsilon(n)}{\partial e_j(n)} \frac{\partial e_j(n)}{\partial y_j(n)} \frac{\partial y_j(n)}{\partial v_j(n)} == e_j(n)\varphi'_j(v_i(n)) \tag{3.11}$$

Lokal değişim ölçüsü sinaptik ağırlıklarda gerekli olan değişiklikleri belirtmektedir. Denklem 3.11'e göre çıkış nöronu j için belirlenen lokal değişim ölçüsü $\delta_j(n)$ bu nöron için karşılık gelen hata sinyali olan $e_j(n)$'nin ve ilişkili aktivasyon fonksiyonun $\varphi'_j(v_i(n))$. türevinin çarpımına eşittir.

Denklem 3.10 ve Denklem 3.11'de $\Delta w_{ji}(n)$ nin ağırlık hesaplamasında j. çıkış nöronundaki $e_j(n)$ hata sinyalinin önemli bir rol oynadığı görülmektedir. Buradan hareketle j. çıkış nöronunun ağdaki bulunduğu yere göre iki farklı senaryo belirlenebilir. Birinci senaryo, j. nöronun çıkış düğümünde olmasıdır, bu durumun işlenmesi kolay olmaktadır çünkü ağın her bir çıkış düğümü, kendisine ait istenen yanıtı sağlayarak ilişkili hata sinyalini hesaplamayı kolaylaştırmaktadır.

Buraya kadar bahsedilen kısım ve denklemler birinci senaryonun oluşması durumunda çalışan algoritmayı anlatmaktadır. Devamında ise j. nöronun gizli nöron olması halindeki algoritma anlatılmıştır.



Şekil 3.4 : *k*. çıkış nöronunun *j*. gizli nörona bağlanmasının sinyal akış grafiği, Haykin (2019).

Nöron *j* ağın gizli bir katmanında bulunduğunda, bu nöron için istenen herhangi bir yanıt yoktur. Buna göre, gizli bir nöron için hata sinyalinin özyinelemeli olarak belirlenmesi ve gizli nöronun doğrudan bağlı olduğu tüm nöronların hata sinyalleri açısından geriye doğru çalışması gerekir; geri yayılma algoritmasının geliştirilmesinin karmaşık hale geldiği yer burasıdır (Haykin, 2009). Denklem 3.11'e göre yerel değişim ölçüsü olan $\delta_j(n)$ j. gizli nöron için Denklem 3.12'de ki gibi belirlenir.

$$\delta_j(n) = -\frac{\partial \varepsilon(n)}{\partial y_j(n)} \frac{\partial y_j(n)}{\partial v_j(n)} = -\frac{\partial \varepsilon(n)}{\partial y_j(n)} \varphi_j'(v_i(n)) \tag{3.12}$$

Denklemin ikinci satırında Denklem 3.7 kullanılmıştır. Kısmi türev olan $\frac{\partial \varepsilon(n)}{\partial y_j(n)}$ hesaplamak için çıkış nöronun *k* olarak simgelendiği Denklem 3.13 kullanılmalıdır.

$$\varepsilon(n) = 1/2 \sum_{k \epsilon C} e_k^2\,(n) \tag{3.13}$$

Denklem 3.13'ün diferansiyeli alındığında sinyal fonksiyonu olan $y_j(n)$ Denklem 3.14'te ki şekilde ifade edilmektedir.

$$\frac{\partial \varepsilon(n)}{\partial y_j(n)} = \sum_k e_k \frac{\partial e_k(n)}{\partial y_j(n)} \tag{3.14}$$

Daha sonra kısmı türev $\frac{\partial e_k(n)}{\partial y_j(n)}$ için zincir kuralı uygulanır ve Denklem 3.14 yeniden yazılarak Denklem 3.15 elde edilir.

$$\frac{\partial \varepsilon(n)}{\partial y_j(n)} = \sum_k e_k(n) \frac{\partial e_k(n)}{\partial v_k(n)} \frac{\partial v_k(n)}{\partial y_j(n)} \tag{3.15}$$

Şekil 3.4'ten hareketle *k* çıkış düğümü olacak şekilde,

$$\begin{aligned} e_k(n) = d_k(n) - y_k(n) = d_k(n) - \varphi_k'\big(v_k(n)\big) = \frac{\partial e_k(n)}{\partial v_k(n)} \\ = -\varphi_k'(v_k(n)) \end{aligned} \tag{3.16}$$

Denklem 3.16'dan hareketle *k* için ayrılan lokal alanın belirteci olan $v_k(n)$ Denklem 3.17'de ki gibi hesaplanmaktadır.

$$v_k(n) = \sum_{j=0}^{m} w_{kj}(n) y_j(n) \tag{3.17}$$

Burada kullanılan *m* üst indisi *k* nöronuna uygulanan sapmalar hariç toplam girdi nöronu sayısıdır. Burada yine sinaptik ağırlık olarak tanımlanan $w_{kj}(n)$, *k.* nörona uygulanan sapmaya eşitlenmiş olmakta ve karşılık gelen girdi değeri de +1'e eşitlenmiştir. Denklem 3.17'nin diferansiyeli alındığında aşağıdaki Denklem 3.18 elde edilmektedir.

$$\frac{\partial v_k(n)}{\partial y_j(n)} = w_{kj}(n) \tag{3.18}$$

Denklem 3.18, 3.16 ve 3.15 birlikte kullanıldığında istenilen kısmi türev Denklem 3.19'da ki gibi elde edilmektedir.

$$\begin{aligned} \frac{\partial \varepsilon(n)}{\partial y_j(n)} = -\sum_k e_k(n) \varphi_k'\big(v_k(n)\big) w_{kj}(n) \\ = -\sum_k \delta_k(n) w_{kj}(n) \end{aligned} \tag{3.19}$$

Son olarak Denklem 3.12, Denklem 3.19'un içinde kullanıldığında lokal değişim noktası olan $\delta_j(n)$ için geri yayılma algoritma formülü bulunmuş olup Denklem 3.20'de ki şekliyle gösterilir.

$$\delta_j(n) = \varphi'_j(v_i(n)) \sum_k \delta_k(n) w_{kj}(n) \tag{3.20}$$

Denklem 3.12'de yer alan $\varphi'_j(v_i(n))$ dış faktörü, lokal değişim ölçüsü $\delta_j(n)$ nin sadece j. gizli nöron ile ilişkili aktivasyon fonksiyonuna bağlıdır. Bu hesaplamadaki diğer faktörlerden biri olan $\delta_k(n)$ j. nörona bağlanmış olan tüm nöronlar için $e_k(n)$ ile simgelenen hata sinyalleri hakkında bilgi gerektirir. İkinci faktör olan $w_{kj}(n)$ bu bağlantılarla ilişkili sinaptik ağırlıklardan oluşur.

Geri yayılım algoritmasını özetlemek gerekirse ilk olarak i. nöronu j. nörona bağlayan sinaptik ağırlıklara $\Delta w_{ji}(n)$ düzeltmesi delta kuralı ile uygulanır Denklem 3.21'de kuralın özet hali mevcuttur.

$$\begin{pmatrix} Ağırlıkların \\ Düzenlenmesi \\ \Delta w_{ji}(n) \end{pmatrix} = \begin{pmatrix} Öğrenme\ oranı \\ Parametresi \\ \eta \end{pmatrix} * \begin{pmatrix} Yerel \\ Gradyan \\ \delta_j(n) \end{pmatrix} * \begin{pmatrix} j.Nöronun \\ Giriş\ Sinyali \\ y_j(n) \end{pmatrix} \tag{3.21}$$

İkinci olarak, yerel gradyan $\delta_j(n)$ j. nöronun bir çıkış düğümü mü yoksa gizli bir düğüm mü olduğuna bağlıdır:

1. Eğer j. nöron bir çıkış düğümünde ise $\delta_j(n)$ her ikisi de nöron j ile ilişkili olan $\varphi'_j(v_i(n))$ türevine ve $e_j(n)$ hata sinyaline bağlıdır.
2. Eğer j. nöron gizli bir düğümdeyse $\delta_j(n)$ $\varphi'_j(v_i(n))$ türevine ve j. nörona bağlı olan bir sonraki gizli veya çıkış katmanındaki nöronlar için hesaplanan $\delta(s)$lerin toplamına eşittir.

3.1.1.4 Çok katmanlı algılayıcı ve algoritma parametreleri

Minsky ve Papert'in çalışmalarından yola çıkılarak, o sırada yapay sinir ağlarında büyük bir hayal kırıklığına neden olan ve bu alanlardaki sınırlamaların sonucunda, yapılan araştırmaların sayısı düştüğü için, tek katmanlı yapay sinir ağının ötesine geçmek bir gereklilik haline gelmişti. Graupe (2013), tek katmanlı bir YSA'nın dışbükey açık bir bölgede veya dışbükey kapalı bir bölgede yer alan noktaların sınıflandırma problemlerini çözebileceğini (temsil edebileceğini) göstermiştir.

Dışbükey bölge, o bölgedeki herhangi iki noktanın, o bölgede tam olarak yer alan düz bir çizgiyle birleştirilebildiği bölgedir. 1969'da çıkışına (y) erişilebilen nöronlar dışında ağırlık ayarlama yöntemi yoktu. Daha sonra, 2-tabakalı bir YSA'nın, Çizelge 3.1'de ki XOR problemi dahil olmak üzere dışbükey olmayan problemleri de çözebileceği gösterilmiştir. Üç veya daha fazla katmana genişletme, YSA tarafından temsil edilebilecek ve böylece çözülebilecek problem sınıflarını, esasen, sınırsız olarak genişletir. Bununla birlikte, 1960'larda ve 1970'lerde çok katmanlı bir sinir ağının ağırlıklarını ayarlamak için güçlü bir araç yoktu. Her ne kadar Madaline için çok katmanlı eğitim zaten bir dereceye kadar kullanılmış olsa da yavaştı ve genel çok katmanlı problem analizi için yeterince titiz değildi (Graupe, 2013).

Bu algoritma, "ileri beslemeli yapay sinir ağı" grubuna aittir. Giriş katmanı, ağırlıklar ve aktivasyon fonksiyonu olmak üzere toplam üç katmandan oluşur. Model, doğrusal olmayan etkinleştirme işlevini kullanarak çıktı verebilmek için girdileri ağırlıkları ile kullanır. Çok katmanlı algılayıcı, geri yayılım tekniğini kullanarak eğitim için denetimli öğrenmeyi kullanır. Çoklu katmanlar ve doğrusal olmayan aktivasyon fonksiyonu ÇKA'yı doğrusal bir algılayıcıdan (DA) ayırır. Karmaşık regresyon ve sınıflandırma problemlerini çözme yeteneğine sahiptirler (Giri ve diğ, 2019).

Çizelge 3.1 : XOR doğruluk tablosu.

Koşul	X_1	X_2	Z
A	0	0	0
B	1	0	1
C	0	1	1
D	1	1	0

Çok katmanlı algılayıcı parametrelerine bakıldığında, optimal parametre seçimi için bir yöntem olmamasından kaynaklı literatürde araştırmacılar problem çözümlerinde çoğunlukla deneme yanılma yöntemi ile parametreleri saptamışlardır. Çizelge 3.2'de bu parametreler ve tanımları ayrıntıları ile verilmektedir.

Çizelge 3.2 : Çok katmanlı algılayıcı parametreleri, Giri ve diğ. (2019)'dan uyarlanmıştır.

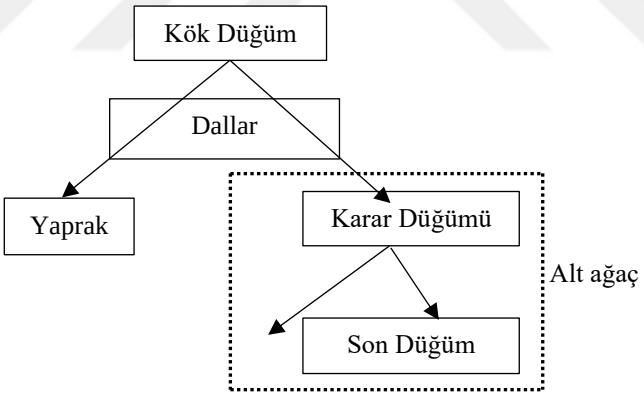
ÇKA Parametreleri	Tanımı
Gizli Nöron Sayısı	Gizli katmandaki optimum nöron sayısını hesaplayacak bir formül yoktur. Yine de Baily ve Thoms veya Kartz gibi yazarlar tarafından bazı kurallar ileri sürülmüştür (Giri ve diğ, 2019). En iyi gizli nöron yelpazesinin seçimi, test setinde iyi performans gösteren ağı seçerek deneme yanılma yöntemine göre elde edilir.
Öğrenme Oranı	Bu parametre nörondan hataya ağırlığın etkisine bağlı olarak ağırlık değişikliklerinin boyutunu belirleyen ve öğrenme hızını ağ eğitim süresini hızlandıran bir sabittir.
Momentum Oranı	Bir önceki bağlantı ağırlığı değişim miktarının belli bir oranının yeni değişim miktarına eklenmesidir.
Giriş ve Çıkış Katmanları	Bu parametre, giriş ve çıkış katmanlarının nöronları arasında bir bağlantı olup olmadığını tanımlar.
Çapraz Nöronlar	Nöronların toplamına sabit bir element sağlayarak öğrenme sürecine katılan ek bir nöron. Bu özel nöron, giriş ve gizli katmanlardan diğer tüm nöronlara bağlanır.
Toplu İş Boyutu	Toplu tahmin gerçekleştiriliyorsa işlenecek tercih edilen örnek sayısı. Daha fazla veya daha az örnek sağlanabilir, ancak bu, uygulamalara tercih edilen bir toplu iş boyutu belirtme şansı verir.
Eğitim Süresi	Yapılacak eğitim sayısı. Doğrulama kümesi sıfırdan farklıysa, ağı erken sonlandırabilir

3.1.2 Rastgele orman algoritması

3.1.2.1 Karar ağaçlarının çalışma prensibi

Karar ağacı, sınıflandırma, karar teorisi, kümeleme ve tahmin işlevleri için kullanılan ağaç benzeri bir yapıdır (Browne ve Berry, 2006). Bir karar ağacı kategorik ve nümerik değişkenler için kullanılabilir. Kategorik değişkenlere uygulandığında sınıflandırma ağacı, nümerik değişkenlere uygulandığında ise regresyon ağacı olarak nitelendirilirler. Bu ağaçlar oluşum yapılarına göre değişebilirler. Örneğin bazı durumlarda ağaçlar yukarıdan aşağıya, bazı durumlarda ise soldan sağa doğru oluşturulurlar.

Karar çok aşamalı kriterlere ve değişkenlere dayandığında bir karar ağacı kullanılır. Bir karar ağacı, diğer öngörücü modellerin çıktısına kıyasla anlaşılması ve uygulanması daha kolay olan resimsel bir çıktıya sahip olduğu için karar verme aracı olarak çok etkilidir (Vo.T.H, 2017). Bir ağaç düğümler, dallar ve yapraklar olmak üzere toplam üç bölümden oluşmaktadır. Düğümler bir veya daha fazla dalın çıktığı noktalardır. Tipik bir ağaç Şekil 3.5'te ki gibi görünmektedir.



Şekil 3.5 : Temel bileşenleri ile bir karar ağacının gösterimi, Vo.T.H, (2017).

Bir karar ağacı bir kök düğümle başlar, karar düğümlerine ve nihayetinde karar kurallarının uygulandığı son düğümlere ilerler. Son düğümü hariç tüm düğümler bir değişkeni temsil eder ve dallar bu değişkenin farklı değerlerini temsil eder. Son düğüm, bu rota için nihai kararı veya değeri temsil eder (Vo.T.H, 2017).

3.1.2.2 Entropi

Bir düğümün safsızlığı veya heterojenliği entropi adı verilen terim kullanılarak ölçülebilir (Vo. T. H, P, 2017). Bir değişkenin entropisi belirsizlik derecesinin ölçüsü olarak tanımlanır (Elaidi ve diğ, 2018). Temel dayanak noktası, daha homojen veya saf düğümlerin temsil edilmesi için daha az bilgi gerektirmesidir. Bilgilere bir örnek, iletilmesi gereken şifreli bir sinyal olabilir. Mesajın bileşenleri daha homojen ise, daha az bit tüketecektir. Daha az bilgi gerektiren yapılandırma her zaman tercih edilir. Sistemdeki belirsizlik arttığında entropi değeri artar, azalırsa aynı oranda entropi değerinde de azalış gözlemlenir.

Değerleri 0 ve 1 arasında değişen entropinin genel formülü Denklem 3.22'de ki gibidir.

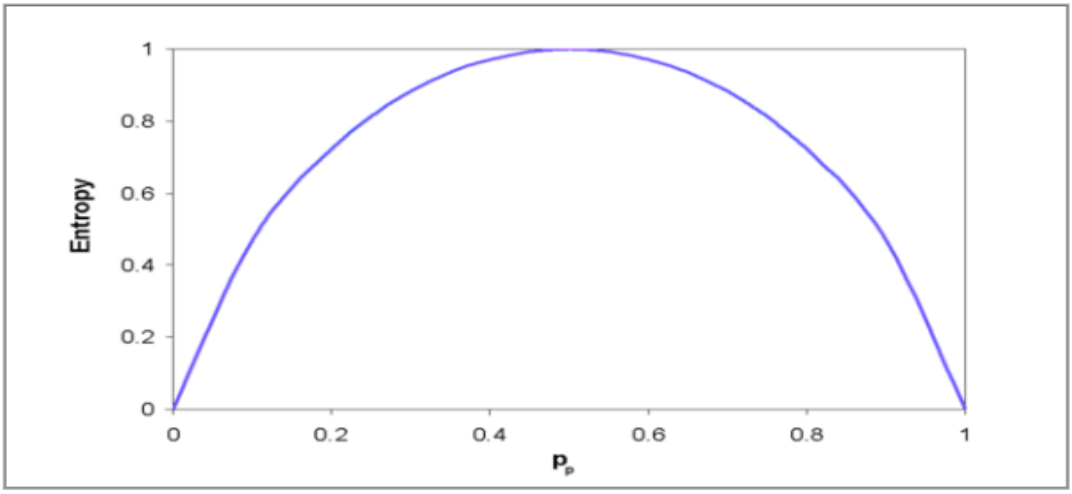
$$Entr(S) = -\sum_{i=1}^{k} p_i \log_2(p_i) \tag{3.22}$$

Denklem 3.22'de,

S: C_1, C_2, C_3……, C_k sınıflarına bölünmüş biz dizi örnek,

p_i: C_i sınıfına ait olan *S* kümesinden bir nesne koleksiyonun olasılığıdır.

Entropi 0 ile log2*c* arasında değişir; *c*, hedef değişkende bulunan kategori sayısıdır. Şekil 3.6'da görüleceği üzere, gözlemler mükemmel homojen olduğunda entropi 0 olurken, gözlemler mükemmel şekilde heterojen olduğunda log2*c* olacaktır (Vo. T. H, P, 2017). Tüm p_i'ler eşit olduğunda entropi maksimum değere (log2*c*) ulaşır. Entropi p_i değeri 1 olduğunda minimum (0) değerine ulaşır.



Şekil 3.6 : p_i değerleri için karşılık gelen entropi değerleri, Vo.T.H (2017).

3.1.2.3 Bir karar ağacının oluşum süreci

Bir karar ağacının oluşturulması kabaca iki adımdan oluşur.

1. X_1, X_2, X_3……,X_p değişken değerlerini simgeleyen tahminleyici uzay R_1, R_2, ….., R_j adet alana bölünür.
2. R_j bölgesine düşen her bir gözlem değeri için aynı tahmin yapılır. Bu tahmin de Rj bölgesindeki gözlenen değerlere karşılık gelen sonuçların ortalamasının alınmasıdır.

Örneğin, 1. Adımda R_1 ve R_2 olmak üzere iki bölge elde edildiği ve birinci bölgedeki eğitim gözlemlerine karşılık gelen değerlerin ortalamasının 10, ikinci bölgedeki eğitim gözlemlerine karşılık değerlerin ortalamasının 20 olduğu varsayılsın. Daha sonra belirli bir gözlem için $X = x$, $x \in R_1$ ise 10 değeri tahmin edilir ve $x \in R_2$ ise 20 değerini tahmin edilir. 1. adımda anlatılan R_1,…., R_J bölgelerinin nasıl oluşturulduğundan bahsetmek gerekirse, teorik olarak, bölgeler herhangi bir şekle sahip olabilir. Bununla birlikte, Denklem 3.23'te ki gibi öngörücü modelin basitliği ve yorum kolaylığı için belirleyici alanı yüksek boyutlu dikdörtgenlere veya kutulara bölünür.

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{Rj})^2 \tag{3.23}$$

Amaç RSS'yi en aza indiren R_1,…., R_J kutularını bulmaktır.

Burada $\hat{y}_{Rj}$ j. kutu içindeki eğitim gözlemleri için ortalama yanıttır. Ne yazık ki, özellik alanının tüm olası bölümlerini J kutularına ayırmak hesaplama açısından mümkün değildir. Bu nedenle, özyinelemeli ikili bölünme olarak bilinen yukarıdan aşağıya, 'greedy' ismiyle bilinen bir yaklaşım benimsenmiştir. Özyinelemeli ikili bölme yaklaşımı yukarıdan aşağıyadır, çünkü ağacın tepesinden başlar (bu noktada tüm gözlemler tek bir bölgeye aittir) ve ardından art arda bölünmelerle tahmin alanını ayırır; her bir bölünme ağaç üzerinde iki yeni dal ile gösterilir. Bu ismin verilmesinin sebebi, ağaç oluşturma sürecinin her adımında, geleceğe bakmak ve gelecekteki bir adımda daha iyi bir ağaca yol açacak bir bölünme seçmek yerine, her adımda en iyi bölünme yapılmasıdır. Özyinelemeli ikili bölme yapmak için, önce X_j belirleyicisi ve kesme noktası s seçilir, böylece öngörücü alanı $\{X \mid X_j < s\}$ ve $\{X \mid X_j\}\}$ bölgelerine bölmek RSS'de mümkün olan en büyük azalmaya yol açmaktadır.

Yani X_1'den X_R'ye tüm tahminleyiciler ve kesme noktası s'nin tüm değerleri ve ardından son ağacın en düşük RSS'ye sahip olacağı şekilde belirleyici ve kesme noktası Denklem 3.24'te ki gibi seçilir.

$$R_1(j,s) = \{X|X_j < s\} \; ve \; R_2(j,s) = \{X|X_j \geq s\} \tag{3.24}$$

Denklem 3.24'ü minimize eden j ve s değerleri aranır.

$$R_1(j,s) = \{X|X_j < s\} \; ve \; R_2(j,s) = \{X|X_j \geq s\} \tag{3.25}$$

Denklem 3.25'te, $\hat{y}_{R1}$ $R_1(j, s)$ 'de ki eğitim gözlemleri için ortalama yanıttır, $\hat{y}_{R2}$ $R_2(j, s)$' deki eğitim gözlemleri için ortalama yanıt olmaktadır.

Denklem 3.25 değerini en aza indiren j ve s değerleri, özellikle p değeri çok büyük olmadığında oldukça hızlı bir şekilde bulunabilir. Daha sonra, elde edilen bölgelerin her birinde RSS'yi en aza indirgemek için verileri daha da bölmek için en iyi öngörücüyü ve en iyi kesme noktasını aranarak işlem tekrar eder.

3.1.2.4 Rastgele orman algoritması çalışma prensibi

Rastgele Orman, büyük değişkenlik gösteren bir veri setini bağımsız sınıflandırıcıların her modele göre daha doğru bir sınıflandırma üretmek için çoğunluk oyu kullanılarak toplandığı "bagging" yöntemine dayanan bir topluluk yöntemidir (Fratello ve Tagliaferri, 2018). Rastgele Ormanlarda kullanılan temel öngörücü yapı karar ağacı modellemesi mantığı ile çalışır. Verilerin eğitim aşamasında kullanıcı isteği doğrultusunda hem sınıflandırma hem de regresyon problemlerinde başarılı bir şekilde kullanılabilir. Sağladığı en büyük avantaj bireysel sınıflandırıcılardaki aşırı öğrenme (ezberleme) problemini ortadan kaldırmasıdır.

Rastgele ormanlar, her bir ağaç bağımsız olarak örneklendiği ve ormandaki tüm ağaçlar için aynı dağılıma sahip rastgele bir vektörün değerlerine bağlı olacak şekilde ağaç öngörücülerinin bir kombinasyonudur (Breiman, 2001). Ormanlar için genelleme hatası, ormandaki ağaçların sayısı arttıkça bir sınıra yaklaşmaktadır. Bir ağaç sınıflandırıcıları ormanının genelleme hatası, ormandaki tek tek ağaçların gücüne ve aralarındaki korelasyona bağlıdır. Dahili tahminler hatayı, gücü ve korelasyonu izler ve bunlar bölünmede kullanılan özelliklerin sayısının artmasına yanıt vermek için kullanılır. İç tahminler de değişken önemi ölçmek için kullanılır. Bu fikirler regresyon için de geçerlidir. Breiman'ın fikirleri Amit ve Geman'ın geometrik özellik seçimi, Ho'nun rasgele alt uzay yöntemi ve Dietterich'in rasgele bölünmüş seçim yaklaşımı

üzerine yaptığı ilk çalışmalardan büyük ölçüde etkilendi (Breiman, 2001). Çeşitli ampirik çalışmalarla vurgulandığı gibi, rastgele ormanlar, vektör makinelerinin etkinliğini artırma ve destekleme gibi popüler yöntemlere rakip olarak ortaya çıkmıştır. Hızlı ve kolay uygulanırlar, son derece hassas tahminler üretirler ve çok fazla veriyi aynı anda işleyerek, çok sayıda girdi değişkenini işleyebilirler. Aslında, mevcut en doğru genel amaçlı öğrenme tekniklerinden biri olarak kabul edilirler.

Yapı olarak, rastgele bir orman, birçok regresyon ağaçlarından oluşmaktadırlar. {r_n (x, Q_m, D_n), $m \geq 1$}, burada Q_1, Q_2, bir koleksiyonundan oluşan bir tahminleyicidir. Bu rasgele ağaçlar, regresyon tahminini oluşturmak için birleştirilir (Biau, 2012).

$$\bar{r}_n(X, D_n) = E_Q[r_n(X, Q, D_n)] \tag{3.26}$$

Denklem 3.26'da E_Q, rastgele parametreye ilişkin beklentiyi, X'e ve D_n veri setine şartlı olarak gösterir. Rastgeleleştirme değişkeni Q, ayrı ayrı ağaçları inşa ederken, bölünecek koordinatın ve bölünmenin konumu gibi birbirini izleyen kesimlerin nasıl yapıldığını belirlemek için kullanılır. Önerilen modelde Q değişkeni X ve eğitim seti Dn'den bağımsızdır (Fratello ve Tagliaferri, 2018). Her bir rastgele ağaç aşağıdaki şekilde inşa edilir. Ağacın tüm düğümleri, ağacın yapımının her aşamasında, ağacın yaprakları ile ilişkili hücrelerin toplanması (yani, dış düğümler) $[0,1]^{d'}$nin bir bölümünü oluşturacak şekilde dikdörtgen hücrelerle ilişkilidir. Ağacın kökü $[0,1]^{d}$'dir. Aşağıdaki prosedür daha sonra $log_2 k_n$ kez tekrarlanır, burada $\log_2$, 2 tabanındaki logaritma ve $k_n \geq 2$, kullanıcı tarafından önceden sabitlenmiş ve muhtemelen n'ye bağlı olarak deterministik bir parametredir (Biau, 2012).

1. Her bir düğümde, $p_{nj} \in (0,1)$ olma olasılığı bulunan j. özellik ile $X = (X^{(1)},.., X^{(d)})$ koordinatı seçilir.
2. Her düğümde, koordinat seçildikten sonra, bölünme seçilen tarafın orta noktasındadır.

Her bir rassal ağaç r_n (X, Q) karşılık gelen vektörleri Xi' nin X ile rastgele bölümün aynı hücresine düştüğü tüm Y_i üzerindeki ortalamayı verir Her bir ağaç Denklem 3.27'de ki formüle göre oluşturulur.

$$\bar{r}_n(X, Q) = \frac{\sum_{i=1}^{n} Y_i 1\left[X_{i \in A_n(X,Q)}\right]}{\sum_{i=1}^{n} 1\left[X_{i \in A_n(X,Q)}\right]} 1_{E_n(X,Q)} \tag{3.27}$$

Denklem 3.27'de ki $E_n(X, Q)$ olayı Denklem 3.28'de ki şekilde tanımlanır.

$$E_n(X, Q) = \left[\sum_{i=1}^{n} 1_{[X_{i \in A_n}(X,Q)]} \neq 0\right] \tag{3.28}$$

Son olarak Q parametresine ilişkin beklentiyi alarak, rastgele orman regresyon tahmini Denklem 3.29'da ki formu alır.

$$\bar{r_n}(X) = E_Q[r_n(X, Q)] = E_Q\left[\frac{\sum_{i=1}^{n} Y_i 1\left[X_{i \in A_n(X,Q)}\right]}{\sum_{i=1}^{n} 1\left[X_{i \in A_n(X,Q)}\right]} 1_{E_n(X,Q)}\right] \tag{3.29}$$

Başka bir deyişle, rastgele bir orman inşa ederken, ağaçtaki her bir bölünmede, algoritmanın mevcut öngörücülerin çoğunluğu karar verme aşamasında büyük bir önem taşıyan faktörler olmaz. Veri setinde orta derecede güçlü bir dizi diğer öngörücüyle birlikte çok güçlü bir öngörücü olduğu varsayıldığında, ağaçların bir araya getirilip toplanma (bagging) aşamasında, ağaçların çoğu veya tamamı bu güçlü belirleyiciyi üst bölünmede kullanacaktır. Dolayısıyla bir araya getirilmiş ağaçlardan yapılan tahminler birbirleriyle oldukça ilişkili olacaktır.

3.2 Sınıflandırma ve Kümeleme

Veri dolu bir dünyada yaşıyoruz. Her gün insanlar her türlü ölçüm ve gözlemden gelen farklı veri türleriyle uğraşırlar. Veriler, canlı bir türün özelliklerini tanımlar, doğal bir olayın özelliklerini tasvir eder, bilimsel bir deneyin sonuçlarını özetler ve çalışan bir makine sisteminin dinamiklerini kaydeder. Daha da önemlisi, veriler ileri analiz, muhakeme, kararlar ve nihayetinde her türlü nesne ve olayın anlaşılması için bir temel oluşturmaktadır. Sayısız veri analizi faaliyetinin en önemlilerinden biri, verileri bir dizi kategori veya kümede sınıflandırmak veya gruplandırmaktır. Aynı grupta sınıflandırılan veri nesneleri, bazı kriterlere göre benzer özellikler göstermelidir. Aslında, insanların en ilkel faaliyetlerinden biri olarak sınıflandırma, insani gelişimin uzun tarihinde önemli ve vazgeçilmez bir rol oynamaktadır. Yeni bir nesneyi öğrenmek veya yeni bir fenomeni anlamak için, insanlar her zaman tanımlayıcı özellikleri tanımlamaya ve bu özellikleri, bazı standartlara göre yakınlık olarak genelleştirilmiş olan benzerlik veya farklılıklarına dayanarak bilinen nesneler veya

fenomenlerle daha fazla karşılaştırmaya çalışırlar. Örnek olarak, tamamen doğal nesneler temel olarak üç gruba ayrılır: hayvan, bitki ve mineral. Biyolojik taksonomiye göre, tüm hayvanlar genel olarak özgül olmak üzere sınıf, düzen, aile, cins ve tür kategorilerine ayrılır. Böylece, kaplanlar, aslanlar, kurtlar, köpekler, atlar, koyunlar, kediler, fareler vb. Aslında, adlandırma ve sınıflandırma esasen eşanlamlıdır. Eldeki bu tür sınıflandırma bilgileriyle, ait olduğu kategoriye göre belirli bir nesnenin özelliklerini çıkarabiliriz. Örneğin, kolayca yerde yatan bir mühür gördüğümüzde, yüzdüğünü görmeden iyi bir yüzücü olduğunu hemen biliyoruz (Xu ve Wunsch, 2009). Temel olarak, sınıflandırma sistemleri sırasıyla sınırlı sayıda denetimli sınıftan veya denetimsiz kategoriden birine yeni veri nesneleri atamalarına bağlı olarak denetlenir veya denetlenmez (Xu ve Wunsch, 2009).

Denetimli sınıflandırmada, $x \in \Re d$ olarak gösterilen bir dizi girdi veri vektöründen, burada girdi alanı boyutsallığı, $y \in 1,....,C$ olarak temsil edilen sonlu bir dizi ayrı sınıf etiketine eşleme, burada C, toplam sınıf tipi sayısı, bazı matematiksel fonksiyon $y = y(x, w)$ açısından modellenmiştir; burada w, ayarlanabilir parametrelerin bir vektörüdür. Bu parametrelerin değerleri, giriş-çıkış örneklerinin sonlu veri seti üzerinde ampirik bir risk fonksiyonunu en aza indirgemek olan bir öğrenme algoritması tarafından belirlenir.

(x_i, y_i), $i = 1,..., N$, burada N, mevcut temsili veri kümesinin sınırlı niceliğidir (Xu ve Wunsch, 2009). Kümelemenin amacı, aynı olasılık dağılımından üretilen gözlemlenmemiş örneklerin doğru bir karakterizasyonu sağlamak yerine, sonlu, etiketlenmemiş bir veri kümesini sonlu ve ayrık bir "doğal" gizli veri yapıları kümesine ayırmaktır (Xu ve Wunsch, 2009).

Kümeleme algoritmaları veri nesnelerini (kalıplar, varlıklar, örnekler, gözlemler, birimler) belirli sayıda kümeye (gruplar, alt kümeler veya kategoriler) ayırır. Xu ve Wunsch 2009 yılında yaptıkları çalışmada küme tanımını aşağıdaki şekilde tanımlarlar;

"Küme, birbirine benzeyen bir dizi varlıktır ve farklı kümelerden gelen varlıklar birbirine benzemez."

Bir küme, "kümedeki herhangi iki nokta arasındaki mesafe, kümedeki herhangi bir nokta ile içinde olmayan herhangi bir nokta arasındaki mesafeden daha az olacak şekilde, test alanındaki noktaların toplamıdır."

3.2.1 *K*-ortalamalar algoritması ve çalışma prensibi

K-ortalamalar algoritması en iyi bilinen ve en popüler kümeleme algoritmalarından biridir (Xu ve Wunsch, 2009). *K*-ortalamalar algoritması denklemdeki kare toplamı hata ölçütünü en aza indirerek verilerin en uygun şekilde bölünmesi anlamına gelir.

$$x_j \in C_1, eğer \left\|x_j - m_l\right\| < \left\|x_j - m_i\right\|;\ j = 1, \dots, N, i \neq l\ ve\ i = 1, .., K; \tag{3.30}$$

K-ortalamalar algoritması Denklem 3.30'da ki hata karelerinin toplamını en aza indirerek tırmanma algoritmalarının kategorisine bağlı olan bir optimizasyon süreci ile çalışır. *K*-ortalamalar temel kümeleme prosedürü aşağıdaki gibi özetlenmiştir (Xu ve Wunsch, 2009):

1. Girilen veriler doğrultusunda *k* adet küme rassal bir şekilde belirlenir ve her bir kümenin kendi merkezleri bulunmaktadır. Küme matrisi $M = [m_1, \dots, m_k]$ şeklinde oluşur.
2. Veri kümesindeki her nesneyi en yakın kümeye uzaklık ilişkileri gözetilerek atanır.
3. Daha sonra küme matrisleri yeniden ayarlanıp yeni merkezler oluşturulur. Yeni merkezler Denklem 3.31'de ki formül aracılığı ile belirlenir.

$$m_i = \frac{1}{N_i} \sum_{x_j \in C_i} x_j \tag{3.31}$$

4. Her küme için yapılabilecek başka bir değişim kalmayana kadar 2. Ve 3. Adımlar tekrarlanır.

3.2.2 *X*-ortalamalar algoritması ve çalışma prensibi

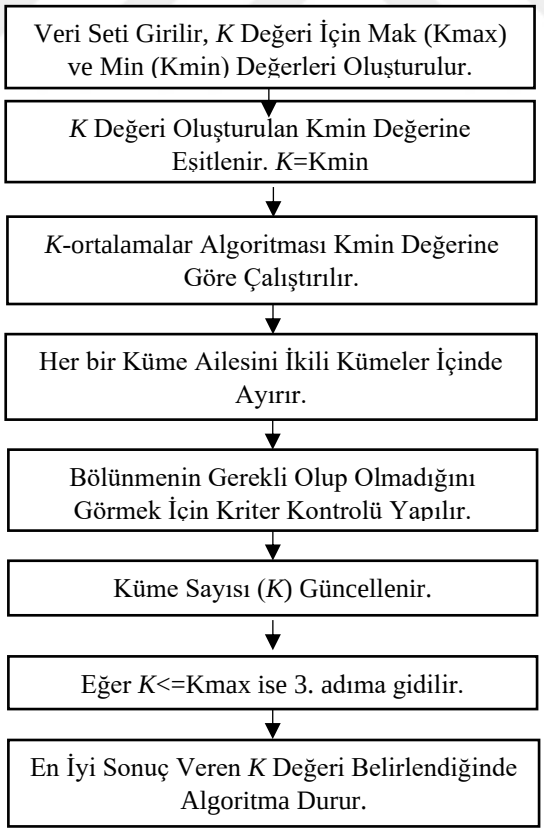
X-ortalamalar kümeleme *K*-ortalamalar kümelemenin ana dezavantajlarından birini çözmek için getirilmiştir (Shahbaba ve Beheshti, 2012) Bu yöntemde, *K*'nın gerçek değeri denetimsiz bir şekilde belirlenir ve sadece veri setine dayanır. Şekil 3.7'de *X*-ortalamalar algoritmasının çalışma aşamaları sırasıyla verilmiştir. *K*-ortalamalar iyi bilinen güçlü bir denetimsiz kümeleme yöntemidir. Etiketlenmemiş verileri belirli bir mesafe tanımıyla *k* kümelerine kümeler. *K*-ortalamalar için küme numarası (yani *k*) önceden tanımlanmıştır ve küme boyunca sabit kalır. *X*-ortalamalar sabit sayıda küme içermeyen *K*-ortalamaların genişletilmiş bir versiyonudur. Geleneksel *K*-ortalamalar birleştikten sonra, Bayes bilgi kriterine (BIC) göre bir ağırlık merkezi tahmini yapılır. BIC, bir model seçimi kriteridir; BIC değeri daha yüksek olan model daha iyi kabul

edilir. Kümelerden herhangi biri bölünmüşse, k + 1. ağırlık merkezi ile başka bir *K*-ortalamalar gerçekleştirilecektir. Kümelenecek veri kümesinin $X = \{x_1, x_2..., x_N\}$ olduğunu, *N*'nin veri kümesindeki örnek sayısını, *K* kümelerin sayısını açıkladığını ve μ_k, *K*. kümenin ağırlık merkezini her *k* adımında tanımladığı varsayıldığında, ikili fonksiyon r_{nk}, x_n'nin *k*. kümesine atanıp atanmadığını göstermek için de tanımlanır (Wu ve diğ, 2014):

$$r_{nk} = \begin{cases} 1, eğer\ x_n\ k.kümeye\ atanırsa, \\ 0, aksi\ halde \end{cases} \tag{3.32}$$

Denklem 3.32 elemanların belirlenen kümelere atanıp atanmadıklarının kontrolünü yapmak için kullanılır. *K*-ortalamalar algoritmasının her bir basamağının amaç fonksiyonu Denklem 3.33'te ki şekilde tanımlanır, aşağıdaki formülde $d\ (x_n, \mu_k)$:

$$J(r, \mu) = \sum_{n}^{N} \sum_{k}^{K} r_{nk} d(x_n, \mu_k) \tag{3.33}$$



Şekil 3.7 : *X*-ortalamalar algoritması adımları, Shahbaba ve Beheshti, (2012).

3.3 Makine Öğrenmesi Yöntemlerinde Kullanılacak Veri Normalizasyonu İçin Teknikler

Normalleştirme ham verilerin üzerinde, her bir özelliğin eşit bir katkısı olacak şekilde yeniden ölçeklendiren veya dönüştüren bir işlemdir. Makine öğrenme algoritmalarının öğrenme sürecini, yani baskın özelliklerin ve aykırı değerlerin varlığını engelleyen iki ana veri sorunuyla ilgilenir. Ham (normalleştirilmemiş) verilerden elde edilen istatistiksel önlemlere dayanarak verileri belirli bir aralıkta normalleştirmek için birçok yöntem önerilmiştir (Singh, 2019). Verilerin önceden işlenmesi, makine öğrenme algoritmaları üzerindeki verileri değerlendirmeden önce iyi bir sınıflandırma performansı elde etmek için önemli bir adımdır. Verilerin ayrıklaştırılması, verilerden aykırı değerlerin kaldırılması, verilerin çeşitli kaynaklardan entegrasyonu, eksik verilerle ilgilenilmesi ve verilerin karşılaştırılabilir dinamik aralıklara dönüştürülmesi gibi birtakım süreçlerden geçirilerek veri ön işlemesi gerçekleştirilmiş olur. Veri normalizasyonu, birbirlerinden farklı birimlerle ifade edilen sayısal değerlerin arasındaki bağlantıyı kurması adına özelliklerin ortak bir aralıkta dönüşümünü içeren önemli bir ön işleme aşamasıdır. Temel amaç ayrımcı örüntü sınıflarında sayısal katkısı daha yüksek olan bu özelliklerin yanlılığını en aza indirmektir. Özelliklerin nispi önemi bilinmiyorsa, verilerdeki özellikler, bilinmeyen bir örneğin çıktı sınıfını tahmin ederken eşit derecede önemli hale getirilmiştir (Singh, 2019). İstatistiksel öğrenme yöntemleri için çok yararlıdır çünkü verilerdeki tüm özellikler öğrenme sürecine eşit katkıda bulunur. *K*-ortalamalar, yapay sinir ağları ve karar destek makineleri gibi çeşitli makine öğrenme algoritmaları için doğru tahmin modelleri oluşturulması için veri normalleştirmesi çok önemli bir adım olmaktadır.

Aşağıdaki şekilde temsil edilen f özelliklerine ve N örneklerine sahip bir veri kümesi düşünülecek olduğunda Denklem 3.34 ortaya çıkar:

$$D = \{x_i, n, y_n | \; i \in f \; ve \; n \in N\}; \tag{3.34}$$

Burada *x*, öğrenme algoritması ile öğrenilecek verileri temsil eder ve *y*, karşılık gelen sınıf etiketidir. Normalizasyon için çeşitli yöntemler önerilmiştir. Bu yöntemler, ham verilerin belirli istatistiksel özelliklerinin verileri normalleştirmek için nasıl kullanıldığına göre kategorilere ayrılır. Yöntemler aşağıdaki gibi tarif edilir (Singh, 2019).

3.3.1 Ortalama ve standart sapmaya dayalı normalleştirme yöntemleri

Bu yöntemlerde, verileri normalleştirmek için ham verilerin istatistiksel ortalaması ve standart sapması kullanılır. Verileri bu yaklaşımlar doğrultusunda normalleştiren birçok farklı yöntem bulunmaktadır. Bunlar aşağıda açıklamaları ile formüle edilmiştir.

3.3.1.1 *Z*-değeri normalizasyonu

Ortalama ve standart sapma ölçümleri, elde edilen özelliklerin sıfır ortalaması ve birim varyansı olacak şekilde verileri yeniden ölçeklendirmek için kullanılır (Singh, 2019).

Verilerin her bir örneği, $x_{i,n}$ Denklem 3.35'te ki gibi $x'_{i,n}$'e dönüştürülür:

$$x'_{i,n} = \frac{x_{i,n} - \mu_i}{\sigma_i} \tag{3.35}$$

Burada μ ve σ sırasıyla i. özelliğin ortalama ve standart sapmasını gösterir.

3.3.1.2 Merkezi ortalama yöntemi ile normalizasyon

Bir özelliğin ortalamasını o özelliğin her bir örneğinden çıkararak dalgalanmayı verilerden kaldırır. Yöntem Denklem 3.36'da ki gibi tanımlanır:

$$x'_{i,n} = \frac{x_{i,n} - \mu_i}{\sigma_i} \tag{3.36}$$

3.3.2 Pareto ölçeklendirme yöntemi ile normalizasyon

Bu yöntem, ölçeklendirme faktörünün standart sapmanın kare kökü olduğu bir skor yöntemine benzer (Singh, 2019).

Bu nedenle, yeni özellikler normalleştirilmemiş özelliklerin standart sapmasına eşit bir varyansa sahiptir. Formülasyonu Denklem 3.37'de ki gibidir.

$$x'_{i,n} = \frac{x_{i,n} - \mu_i}{\sqrt{\sigma_i}} \tag{3.37}$$

Yöntem, verilerdeki gürültünün katkısını en aza indirirken, daha düşük yoğunluktaki özelliklerin temsilini geliştirir. Ayrıca, skor yönteminin aksine birim varyansının sınırlamasını kaldırır ve verilerin yapısını kısmen sağlam tutar (Singh, 2019).

3.3.3 Değişken kararlılık ölçekleme yöntemi ile normalizasyon

Yöntem, ölçeklendirme faktörü olarak Varyasyon Katsayısı (CV) ekleyerek *z* skoru normalizasyonunu genişletir.

Varyasyon katsayısı, verinin ortalamasının standart sapmasına Denklem 3.38'de ki gibi tanımlanan oranı olarak verilir (Singh, 2019):

$$x'_{i,n} = \frac{x_{i,n} - \mu_i}{\sigma_i} * \frac{\mu_i}{\sigma_i} \tag{3.38}$$

Varyasyon katsayısı, küçük standart sapmaya sahip özelliklere daha yüksek, büyük standart sapmaya sahip özelliklere daha az önem vermektedir.

3.3.4 Güç dönüşümü

Bu yöntem, heteroskedastisitenin (değişen varyans) etkilerini azaltarak verileri homoscedastisiteye (eş varyanslık) dönüştürür ve ortalamasının kökü ile orantılı olarak standart sapmaya sahip olan verilere uygulanır. Ham veriler, kare kökü hesaplanarak dönüştürülür ve daha sonra ortalama ortalanmış yöntem kullanılarak yeniden ölçeklendirilir (Singh, 2019). Formülleştirilmiş hali Denklem 3.39'da ki gibidir:

$$x'_{i,n} = p_{i,n} - \mu_i^p, p_{i,n} = \sqrt{\hat{x}_{i,n}} \tag{3.39}$$

Yöntem, negatif değerleri gerçek değerlere dönüştürmek için kullanılamaz. Bu nedenle, bu özelliklerin değerleri, verileri normalleştirmeden önce Denklem 3.40'ta ki gibi kaydırılır:

$$x'_{i,n} = x_{i,n} - \min(x_i) \tag{3.40}$$

Burada min, *d*. özelliğin minimum değerini gösterir. Veriler, minimum değer sıfır ve maksimum değerle çakışacak şekilde yeniden konumlandırılır, özelliğin iki uç değeri arasındaki farka eşittir. Güç dönüşümünün ana dezavantajı, katkı maddesi olmak için veriler üzerinde çarpma etkisi yapamamasıdır. Çarpma etkisi, verilerin standart sapması olduğunda ortaya çıkar.

Bu yöntemler, verilerden aykırı değerlerin etkisini azaltmaya yardımcı olur, ancak *z*-skor yöntemi dışında baskın özellikler sorununun tamamen üstesinden gelmez. Bununla birlikte, yukarıda bahsedilen tüm yöntemlerde verilerin ölçeklendirilmesi

veya aynı sayısal aralığa dönüştürülmemesi nedeniyle ortalama ve standart sapma ölçümleri zamanla değişebilir (Singh, 2019).

3.3.5 Minimum – maksimum değer tabanlı normalleştirme yöntemleri

Normalleştirilmemiş verilerin minimum ve / veya maksimum değerleri yeniden ölçeklendirme için kullanılır. Bu yöntemler şunları içerir:

3.3.5.1 Min – Maks normalleştirme (MMN):

Yöntem, normalleştirilmemiş verileri doğrusal olarak önceden tanımlanmış bir alt ve üst sınırlarla ölçeklendirir (Singh, 2019). Veriler genellikle 0 ila 1 veya -1 ila 1 aralığında yeniden ölçeklendirilir. Denklem 3.41 aşağıdaki gibi verilir:

$$x'_{i,n} = \frac{x_{i,n} - \min(x_i)}{mak(x_i) - \min(x_i)} (nMak - nMin) + nMin \tag{3.41}$$

Burada min ve max, sırasıyla *i*. özelliğinin minimum ve maksimum değerini belirtir. Verileri yeniden ölçeklendirmek için alt ve üst sınırlar sırasıyla nMin ve nMax ile gösterilir (Singh, 2019).

3.3.5.2 Maks normalleştirme (MN):

Bu yöntem, her özelliğin −1 ila +1 aralığında yeniden ölçeklendirildiği, min- maks normalizasyonun bir varyantıdır. Her özelliği maksimum değerine bölerek verileri yeniden ölçeklendirir ve Denklem 3.42'de ki gibi tanımlanır (Singh, 2019):

$$x'_{i,n} = \frac{x_{i,n}}{mak(|x_i|)} \tag{3.42}$$

Bu normalleştirme yöntemleri, verilerin ortalama ve standart sapmasına dayanan normalleştirme yöntemlerinin aksine, bu değerler zamanla değişebileceğinden orijinal girdi verileri arasındaki ilişkilerin korunması için yararlıdır. Bu veri normalleşmesinin en büyük dezavantajı, normalleştirilmemiş verilerde mevcut olan aşırı değerlerin yanı sıra aşırı değerlere olan duyarlılığıdır (Singh, 2019). Diğer bir dezavantaj, test verilerinin bazı değerleri normalleştirilmemiş verilerin aralığının dışında kaldığında ortaya çıkan " sınır dışı " sorunudur.

3.4 Performans Değerlendirme Kriterleri

3.4.1 Regresyon çalışmasında kullanılan performans değerlendirme kriterleri

Genel olarak regresyon çalışmasının başarılı olup olmadığının anlaşılabilmesi için bazı istatistiksel hata ölçütlerine bakılmaktadır. Bunlar, ortalama mutlak hata (MAE), ortalama kare tahmin hatası (MSE), ortalama tahmin hatasının karesinin karekökü (RMSE), ortalama mutlak yüzde hatası (MAPE) gibi değerlendirme kriterleridir (Kwon, 2011).

Uygulama kapsamında göz önünde bulundurulacak değerlendirme kriterleri aşağıda açıklanmıştır. Denklem 3.43 MAE değerinin formülü, Denklem 3.44 RMSE değeri için kullanılan formül ve Denklem 3.45 korelasyon katsayısı hesabı için kullanılan formül olmaktadır.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \tag{3.43}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)}{n}} \tag{3.44}$$

$$R^2 = c^T R_{xx}^{-1} c \tag{3.45}$$

Denklem 3.46'da korelasyon katsayısının karesi x_n bağımsız değişkenleri ve hedef değişken olan y bağımlı değişkeni arasındaki $r_{x_n y}$ korelasyonlarının bir vektörü olan $c = \left(r_{x_1 y}, r_{x_2 y}, \dots\dots, r_{x_N y}\right)$ vektörü ile bağımsız değişkenler arasındaki korelasyon matrisi olan R_{xx} kullanılarak hesaplanır (URL-1, 2020). Bu formüldeki c^T değeri c vektörünün tersi şeklinde belirtilmiştir.

3.4.2 Kümeleme çalışmasında kullanılan performans değerlendirme kriteri

Rand istatistik değeri

Bu değer temel olarak bir veri setindeki değerlerin kümeleme işlemi sonrasında veri gruplarına ayrıldığında gözlem çiftlerinin nereye yerleştirildiğini karşılaştırarak yapılan kümelemenin etkinlik derecesini ölçmek için kullanılır (Steinley ve Brusco, 2018).

N adet verinin bir bölümü $P_1 \cup P_2 \cup \ldots \ldots \cup P_R = P \; ve \; P_i \cap P_j = \emptyset$ olarak tanımlanan *R* alt kümesi ile tanımlanır. Sırasıyla *R* ve *C* alt kümeleri ile *P* ve *Q* olmak üzere iki bölüm verildiğinde, *P* ve *Q* arasındaki grup örtüşmesini göstermek için iki yönlü bir olasılık tablosu oluşturulabilir, n_{rc}, *P* bölümünün *r*. alt kümesinde sınıflandırılan nesne sayısını temsil eder ve *Q* bölümünün *c*. alt kümesinde sınıflandırılan nesne sayısını temsil eder. Rand $\binom{N}{2}$ adet olası nesne çiftlerinden ortaya çıkabilecek dört farklı temel çift tipinin olduğunu göstermektedir (Steinley ve Brusco, 2018):

1. Bir çiftteki nesneler *P* ile aynı sınıfa ve *Q* ile aynı sınıfa yerleştirilir (bu çiftlerin toplam sayısı *a* ile gösterilir);
2. Bir çiftteki nesneler *P* ile aynı sınıfa ve *Q*'da ki farklı sınıflara yerleştirilir (bu çiftlerin toplam sayısı *b* ile gösterilir);
3. Bir çiftteki nesneler *P*'de farklı sınıflara ve *Q*'da aynı sınıfa yerleştirilir (bu çiftlerin toplam sayısı *c* ile gösterilir);
4. Bir çiftteki nesneler *P*'de farklı sınıflara ve *Q*'da farklı sınıflara yerleştirilir (bu çiftlerin toplam sayısı *d* ile gösterilir).

Bu yaklaşım formül haline getirildiğinde;

$$RI = \frac{\binom{N}{2} + \sum_{r=1}^{R} n_{rc}^2 - \frac{1}{2}\left(\sum_{r=1}^{R} n_r^2 + \sum_{c=1}^{C} n_c^2\right)}{\binom{N}{2}} \tag{3.46}$$

Denklem 3.46'da n_r ve n_c sırasıyla *P* ve *Q* bölümlerinin *r*. ve *c*. alt kümelerindeki gözlem sayılarıdır.

$$RI = \frac{TP + TN}{TP + FP + FN + TN} \tag{3.47}$$

Denklem 3.46’nın daha kolay algılanabilmesi için geliştirilmiş bir gösterim daha mevcuttur.

Denklem 3.47’de belirtilen değerler aşağıda açıklamaları ile belirtilmiştir.

TP (True Positive): Birbirine benzer nitelikte olan verileri aynı kümede yer alması.

TN (True Negative): Birbirine benzer nitelikte olmayan ürünlerin farklı kümelerde yer alması.

FP (False Positive): Birbirine benzer nitelikte olan ürünlerin aynı kümede yer alması.

FN (False Negative): Birbirine benzer nitelikte olan ürünlerin farklı kümelerde yer alması.

4. UYGULAMA

4.1 Problem Tanımı ve Metodoloji

Çalışmada spor ürünleri satan, uluslararası en büyük markalardan birinin verisi kullanılmıştır. Firma gizlilik ilkesi gereğince isminin kullanılmamasını istediği için çalışma kapsamında firma ismi belirtilmemektedir. Firmanın belirli bir yılına ait bir sezonluk (ilkbahar-yaz sezonu) verileri ele alınarak sezon içi yeni ürün talebi tahmin edilmeye çalışılmıştır. Talep tahmini yapılırken makine öğrenmesi yöntemlerine başvurulmuştur. Bunun nedeni ise özellikle tekstil sektöründe talep tahmininin klasik yollarla yapılmasını engelleyen ürün çeşitliliği ve ürünlerdeki sirkülasyonun çok fazla olmasıdır. Günümüz dünyasında tekstil sektörü hızlı moda terimi ile karşı karşıyadır. Hızlı moda teriminin ortaya çıkışı, pazardaki rekabeti güçlendirecek çok fazla sayıda ve iyi bir müşteri potansiyeline sahip üretim yapan firma sayısı ve bunun getirdiği rekabet ortamı dolayısıyla özellikle ürün çeşitliliği konusunda müşteri beklentilerini karşılayabilme adına hızlı bir tedarik ağının gelişmesini gerekli kılar. Sonuç olarak, stok tutma birimi (SKU) başına satış hacmi çok düşüktür ve aynı ürün grubundaki SKU'lara olan talep önemli ölçüde değişebilir. Böylece, toplam talep miktarı kesin olarak tahmin edilebilse bile; bu talebin sunulan birçok ürüne nasıl dağıtılacağını tahmin etmek çok zordur.

Üretim planlama ve sipariş karmaşıklığı da buna bağlı olarak artmaktadır. Hızlı ve doğru bir tedarik ağının kurulabilmesi konusundaki ilk aşama talep tahmininin doğru ve eksiksiz yapılmasıdır. Bu şartlar altında başarılı bir tahmin çalışması yapabilmek için öğrenme yeteneğine sahip olan ve bu yeteneğiyle ileriye dönük tahminlerde başarılı olan makine öğrenme yöntemlerinin son on yıl için yapılan literatür araştırması sonucunda oldukça etkili olduğu gözlemlenmiştir.

Firma sipariş süreci incelendiğinde, her sezon başı ve sezon içi belirli dönemlerde ürün kataloğu yayınlandığında belirli ülkelerdeki perakende mağazalarını yöneten birimlerin pazarlama departmanları ‘global marketing meeting’ adı verilen GMM toplantısına katılarak satışa sunulabilecek ürünleri ve bu ürünler içerisinden kendi

ülkeleri için hangi ürünleri seçeceklerine karar vermektedirler. Buradan hareketle her ütünün her ülkede satılmadığı sonucuna varılabilir.
Ürün seçimleri yapıldıktan sonra seçilen ürünlerin numuneleri her ülkeye gönderilir. Sonraki süreçte de numune gönderilen ülkeler ürünlerini bir fuar ya da belirlenen pilot mağazalarda müşterilerine gösterilip test edilir ve müşterilerden aldıkları tepkilerden yola çıkarak firmaya kaçar adet sipariş etmek istedikleri bilgisini vermektedirler. Firma hızlı moda ilkesini benimsediğinden belirtilen süreç her üç veya altı haftada bir tekrarlanır. Yani firma her üç veya altı haftada bir ürünlerinde yeniliğe giderek üretim ve satışlarını gerçekleştirmektedir.
Dolayısı ile üç/altı haftada satılan ürünlere bakılarak yeni sipariş edilecek ürünlerin kaçar adet satılacağı tahmininde makine öğrenmesi yöntemleri kullanılmıştır.

Ancak firmadan elde edilen veri seti 2004-2007 yılları arasındaki sezon satış verileri olmak üzere güncel veriler değildir. Uygulama sonunda yukarıda anlatılan mantıkla hareket ederek bu mantığın firmadan elde edilen veri setleri ile uygulanabilir olduğu gözlemlenmiştir. Şekil 4.2 ve Şekil 4.3'te süreç akış diyagramları belirtilmiştir.

Uygulamada kullanılan program, makine öğrenme konusunda geliştirilmiş, hızlı ve güvenli veri işleme ve kullanıcı dostu ara yüze sahip olduğu için Yeni Zelanda'da bulunan Waikato üniversitesi tarafından geliştirilmiş Weka (Waikato Environment for Knowledge Analysis) programıdır. Programın ara yüzü Şekil 4.1'de gösterilmiştir.
Weka programı, son teknoloji makine öğrenme algoritmaları ve veri önişleme araçlarının bir koleksiyonudur (Witten ve diğ, 2011). Girdi verilerinin hazırlanması, öğrenme şemalarının istatistiksel olarak değerlendirilmesi ve girdi verilerinin ve öğrenmenin sonucunun görselleştirilmesi dahil tüm deneysel veri madenciliği süreci için kapsamlı destek sağlar.
Çeşitli öğrenme algoritmalarının yanı sıra, çok çeşitli önişleme araçları içerir. Bu farklı ve kapsamlı araç setine, kullanıcıların farklı yöntemleri karşılaştırabilmeleri ve mevcut sorun için en uygun olanları belirleyebilmeleri için ortak bir ara yüz üzerinden erişildiği bir platformdur (Witten ve diğ, 2011).

Şekil 4.1 : Weka programı arayüzü, versiyon 3.9.4.

Sezon verilerini kullanarak yapılan talep tahmin çalışmasında kullanılan yöntemler bir yapay sinir ağı tekniği olan çok katmanlı algılayıcı, bir karar ağacı algoritması olan rassal orman algoritması ve kümeleme yöntemlerinden *K*-ortalamalar ve *X*-ortalamalar algoritmaları kullanılmıştır.

İlk aşamada veriler %90 eğitim ve %10 test verisi olarak tamamen rassal bir şekilde Weka programı ile ikiye bölünerek ele alındı. Daha sonra eğitim veri seti kullanılarak çok katmanlı algılayıcı ve rassal orman algoritması için parametre değerleri deneme yanılma yöntemi ile elde edildi. Deneme yanılma yöntemi ile elde edilen en iyi performans değerlendirme kriterlerinin elde edildiği parametreler optimal parametreler olarak belirlendi ve her iki algoritma ile bu parametreler kullanılarak regresyon çalışması yapıldı. Elde edilen sonuca göre çok katmanlı algılayıcı tahmini daha iyi performans değerleri ile ölçtü. Bu ölçümün tesadüfi mi yoksa tutarlı bir ölçüm olup olmadığı sorusunun çözümünde de birbirinden tamamen bağımsız 10 deneme sonuçlarına istatistiksel bir kontrol aracı olarak kullanılabilen *t*-testi yapılmış ve tüm denemelerde çok katmanlı algılayıcı performans değerleri rassal orman algoritmasına göre daha iyi bulunmuştur.

Devamında ise veri setindeki birbirine benzer özellik gösteren ürünleri kümeleyerek her bir küme içerisindeki elemanları ayrı ayrı değerlendirip tahmin çalışması yapma fikri üzerinde durulmuştur. Bunun nedeni; veri setinde bir sezon dahilindeki üretilen tüm ürünler listelenmektedir. Yani, ayakkabıdan pantolon, tişört, aksesuarlar (raket, top vs.) çok çeşitli bir grup üzerinde değerlendirme yapılmıştır. Ancak bu ürünleri tek seferde ele alıp tahmin yapmaya çalışmaktansa kümelere ayırıp her bir küme içinde tahmin çalışması gerçekleştirildiğinde sonuçların iyileşip iyileşmeyeceği değerlendirilmiştir. Seçilen iki kümeleme algoritması veri setine uygulandıktan sonra performans değerlendirmesi yapılırken rand endeks değeri hesaplanmıştır. Bu endeks kümeleme performansını ölçmek için sıklıkla kullanılan bir değerdir. Temelinde

benzer özellik gösteren ürünlerin aynı kümede yer almasını gerekli kılan bir yaklaşıma sahiptir. 0-1 arasında değerlenerek 1'e yakın sonuçlar tatmin edici olmaktadır.

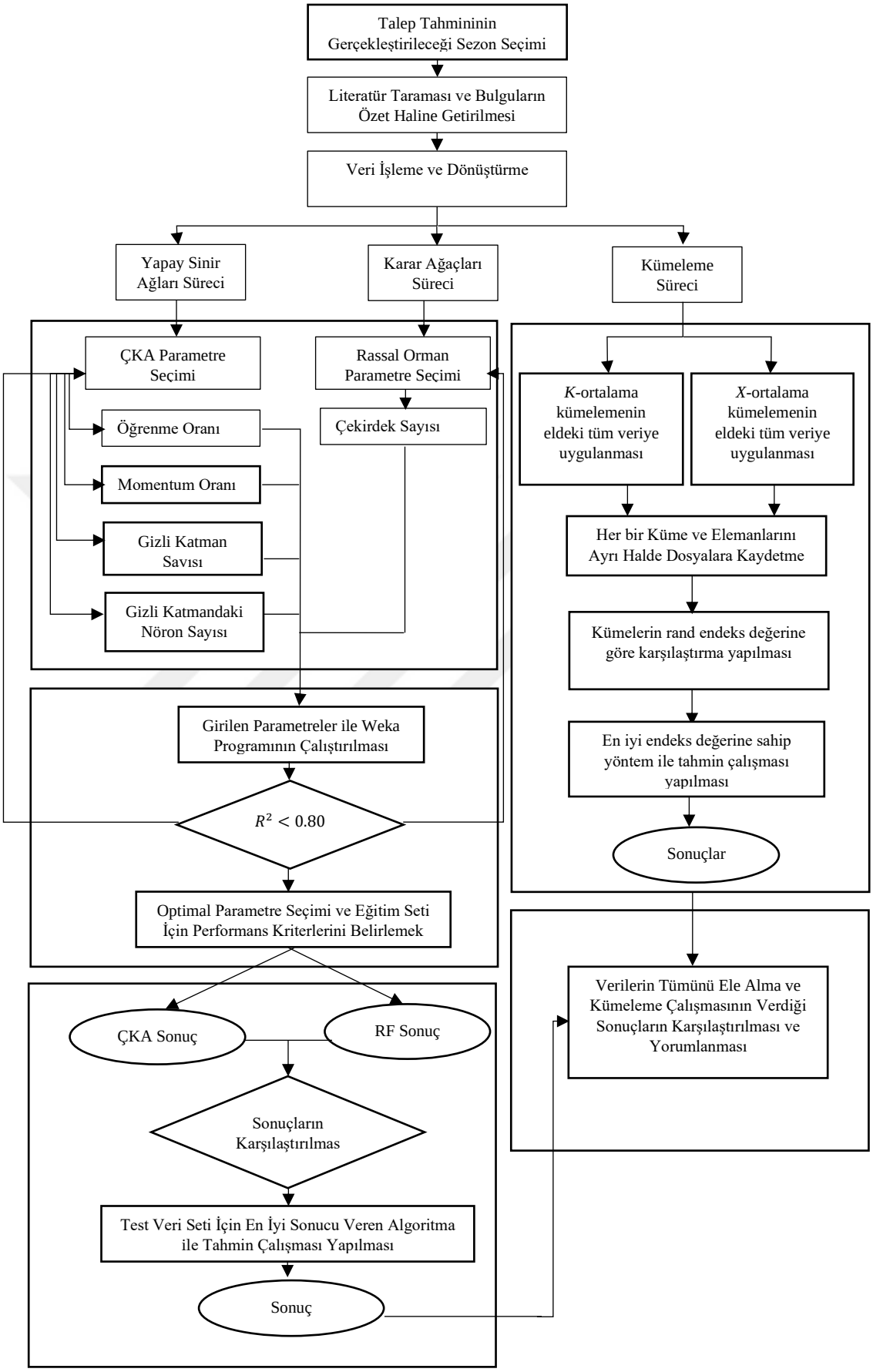
Uygulamada kullanılan kümeleme yaklaşımlarının rand endeks değerleri hesaplandığında *K*-ortalamalar kümeleme daha iyi bir rand endeks değerine sahip olduğu görülmüştür dolayısıyla tahmin çalışması bu yöntem ile yapılmıştır. Çalışma yapılırken bir önceki süreci en iyileyen algoritma seçilen çok katmanlı algılayıcı kullanılmıştır.

Elde edilen değerler kümeleme yapılmadığı haldeki değerler ile karşılaştırışmış ve kümelemenin tahmin performansını iyileştirmediği gözlemlenmiştir. İkinci bir uygulama olarak sezon içi talep değil de belirli bir ürüne ait geçmiş sezon satış değerleri kullanılarak gelecek sezondaki gelebilecek sipariş değeri tahmin edilmeye çalışılmıştır. Bu veri talebi mevsimsel özellik gösteren ceket ürününe aittir. Veri seti işleme konusuna gelindiğinde tüm sezonlarda ortak olarak bulunan karar değişkenleri ve değerleri ayrı bir dosyada toplanmış olup veri dönüştürme işlemleri yapılmıştır.

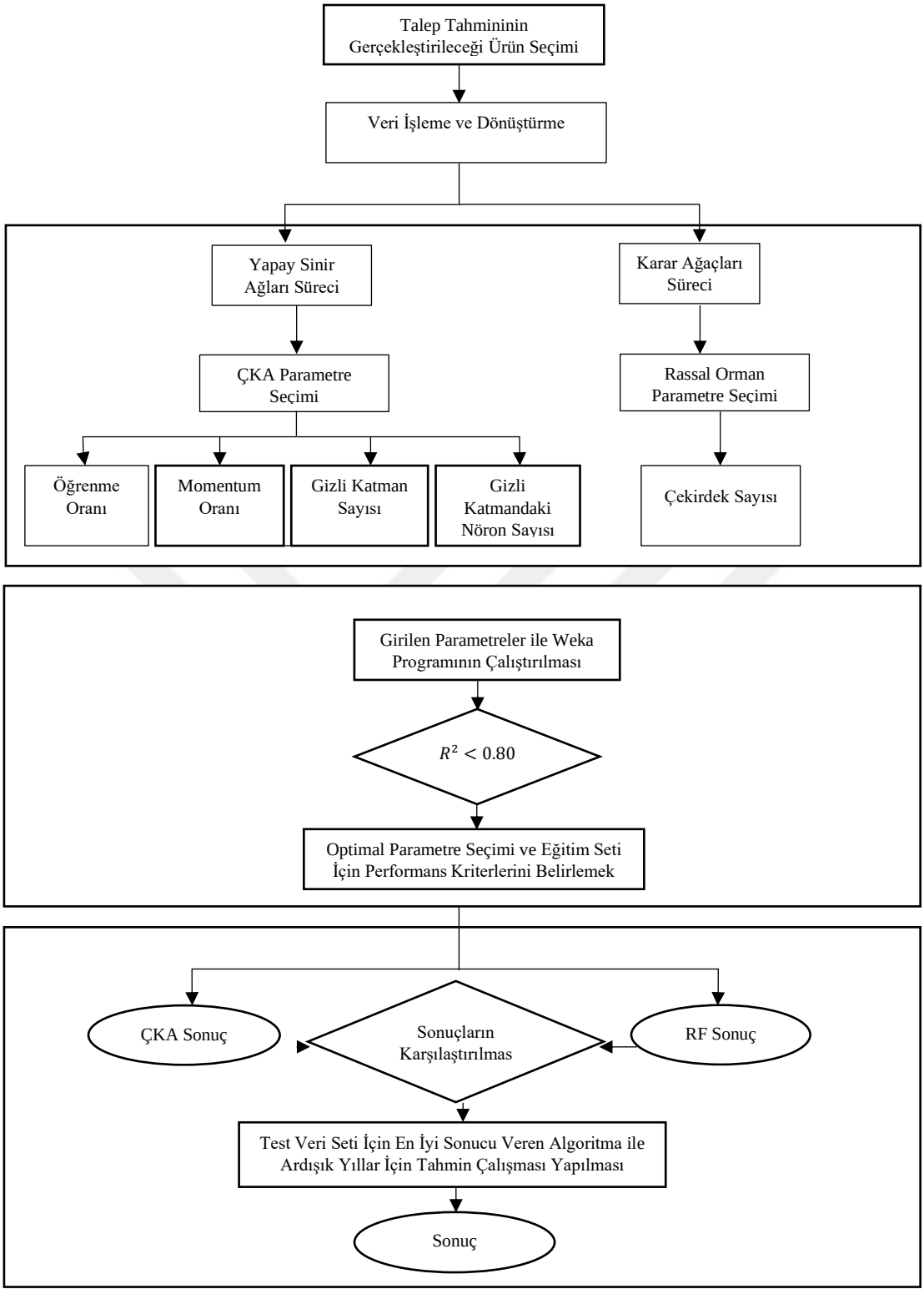
Bir sezonluk veri setine nazaran eldeki değişken sayısı azalmıştır. 2004 ilkbahar yaz ve sonbahar kış sezonlarından başlanarak 2005 yılı ve 2006 yılı verileri kullanılarak 2007 yılı sonbahar kış sezonundaki ceket talebi sayısı tahmin edilmeye çalışılmıştır. İlk etapta talep tahmininin ne kadar başarılı bir şekilde yapıldığının anlaşılabilmesi için 2004 ve 2005 yılının verileri girilerek 2006 yılı ilkbahar yaz ve sonbahar kış sezonundaki ceket talebi tahmin edilmiş olmakta ve gerçek değerleri ile karşılaştırılmaktadır. Bu tahmin değerlerinin hesaplanabilmesi için kullanılan algoritmalar çok katmanlı algılayıcı ve rassal orman algoritmalarıdır. Birinci uygulamadan hareketle yine algoritmaların ikisi de eğitim seti üzerinde denenmiş ve sonucu eniyileyen algoritma ve parametreler belirlendikten sonra bunlarla Weka programında tahmin çalışması yapılmıştır.

Ancak bu uygulamada çok katmanlı algılayıcı değil de rassal orman algoritması başarılı şekilde sonuç vermiştir. Bu sonuca kanıt niteliğinde ilk çalışmada yapıldığı gibi 10 adet birbirinden bağımsız deneme yapılmış ve bu denemelerden elde edilen performans değerlerine *t*-testi uygulanarak yapılan karşılaştırma sonucunda rassal orman algoritması çok katmanlı algılayıcıya göre belirgin şekilde daha iyi bir sonuç vermiştir.

Buradan hareketle, veri setine uygun algoritmayı seçmenin ne kadar önemli olduğu gözlemlenmiştir. Yani her bir problemi çözebilen tek bir makine öğrenme yöntemi diye bir şey olmadığı anlaşılmıştır.



Şekil 4.2 : Uygulama süreç akış diyagramı (bir sezonluk veri).



Şekil 4.3 : Uygulama süreç akış diyagramı (sezonlar arası).

4.2 Veri Setleri

4.2.1 Değişkenlerin tanımlanması

Uygulamada kullanılan veri kümesi şirketin 2004-2007 yılları arasındaki satış verilerini içermektedir. Veri setinde yer alan değişken sayısı gerekli olanlar dışındaki değişkenler ve boş hücreler silindikten sonra bölüm, pazarlama bölümü, ürün grubu, ürün tipi, spor kodu, renk, cinsiyet, ürün alt grubu, ürün kırılımı, tema, ürün sıralaması, ürün kataloğu, ürün bedeni, ürün isteyen firmanın talep edildiği ülke, perakende ayı, ürün üretim ve gönderim maliyeti, ürün satış fiyatı, sipariş miktarı gibi biri bağımsız değişken (sipariş miktarı) olmak üzere toplam 18 adet değişken içermektedir. 3549 adet ürün bu kriterler altında değerlendirmeye tabi tutulmuştur.

Uygulama kapsamında kullanılan karar değişkenleri açıklamaları ile Çizelge 4.1'de ki gibidir. Şekil 4.4'te veri setinden belli bir bölüm ekran görüntüsü şeklinde gösterilmiştir.

Çizelge 4.1 : Değişkenler ve açıklamaları.

Değişkenler	Tanımı	Veri Tipi	Kategori Sayısı	Kategoriler
Bölüm	Üretilen ürünlerin hitap ettiği kategoriyi göstermek için kullanılır.	Nominal	3	Aksesuar, Ayakkabı Ürünleri, Tekstil Ürünleri
Pazarlama Bölümü	Üretilen ürünün performans ürünü olup olmamasına göre değerlendirme yapan bir değişkendir.	Nominal	2	Performans Ürünleri, Genel Ürünler
Spor Kodu	Üretilen ürünün hangi spor dalı için üretiliyor olmasına göre isim alır.	Nominal	13	Yüzme, Tenis, Basketbol, Futbol, Koşu vb.
Ürün Grubu	İlk değişken olan bölüm değişkenindeki kategorileri daha ayrıntılı inceler. Örneğin; üretilen ürünü aksesuar bölümünde ikiye ayırır donanımsal aksesuarlar (toplar, raketler vb.) ve tekstil ürünü aksesuarları (kolyeler, saç bantları vs.) olmak üzere.	Nominal	22	Ayakkabı, tişört, şapka, ceket, pantolon, şort, etek, elbise vb.
Ürün Tipi	Toplu tahmin gerçekleştiriliyorsa işlenecek tercih edilen örnek sayısı. Daha fazla veya daha az örnek sağlanabilir, ancak bu, uygulamalara tercih edilen bir toplu iş boyutu belirtme şansı verir.	Nominal	73	Yüksek topuk, spor ayakkabı, sandal, terlik, omuz çantası, sırt çantası vb.
Ürün Alt Grubu	Ürün grubunun bir alt kademesi olarak üretilen ürünleri çoğunlukla cinsiyet veya ürün tipi gözetilerek inceler.	Nominal	63	Tişört, plaj voleybolu, kemer, bermuda, bikini alt, bikini üstü, taşıma çantası, beşik ayakkabı, elbise, eldiven, gömlek vb.
Renk	Üretilen ürünün rengini tanımlayan değişkendir.	Nominal	79	Sun, afblue, airblu, altıtu, alu, apgre, argblu, artgre, autoba, b.bırd vb.
Cinsiyet	Üretilen ürünleri cinsiyet ve yaşlarına göre ayıran bir değişkendir.	Nominal	7	Erkek çocukları, kız çocukları, yeni doğanlar, çocuklar-unisex, kadınlar, erkekler ve unisex

Çizelge 4.1 (devam) : Değişkenler ve açıklamaları.

Değişkenler	Tanımı	Veri Tipi	Kategori Sayısı	Kategoriler
Ürün Kırılımı	Spot kodunun bir alt kırılımı olarak veri setinde yer almaktadır.	Nominal	35	Basketbol SS 2006, BU Acc. – çantalar, BU Acc. – çoraplar, BU Acc. şapkalar, futbol lisanslı giysiler, 2006 dünya kupası, yürüyüş, mayo aksesuarları vb.
Ürün Teması	Ürün teması, ürün hedeflerine bağlı bir dizi özelliktir. Üretilecek ürün hakkında genel bir fikir oluşturur.	Nominal	120	Süperstar, eşofman takımları, x-trail koşu, phaidon, erkek t-mac, erkek x-edge ayakkabı, güç, koşu parkuru vb.
Ürün Sıralaması	Bu değişken üretilen ürünleri üretim maliyetlerine ve değerlerine göre kategorize eder.	Nominal	4	R2, R3, R4, R5
Ürünlerin Müşteriye İletilme Kanalları	Üretilen ürünlerin hangi kanallar aracılığıyla müşteriye ulaştırılacağını belirleyen değişkendir.	Nominal	3	Reklamlarda yer alan ürünler, mağazada görsellerinin kullanılacağı ürünler, marketing çalışmasının yapılmadığı tekrarlayan ürünler
Ürün Bedeni	Üretilen ürünlerin beden aralıklarını gösteren bir değişkendir.	Nominal	29	3.5, 4, 4.5, 5, 5.5……11,5, XS, S, M, L, XL vb.
Ürünlerin Talep Edildiği Ülke	Ürünlere talep gelmesi halinde satılacağı satış merkezlerinin bulunduğu ülkeleri gösteren bir değişkendir.	Nominal	15	Almanya, İsviçre, Kamboçya, Çin, Endonezya, Pakistan, Tayland, Türkiye, Vietnam, Yunanistan, İtalya, Malezya, Filipinler, Tunus, Ukrayna
Sipariş Ayı	Ürünlerin talep edildiği ayları gösteren değişkendir.	Nominal	12	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
Üretim Maliyeti	Firmanın, talep gelen ürünleri üretim ve teslim etme maliyetini içeren belirlediği fiyattır.	Nümerik	-	Firma gizlilik ilkeleri gereğince belirtilmemiştir.
Ürün Satış Fiyatı	Üretilen ürünler için belirlenen satıl fiyatıdır.	Nümerik	-	Firma gizlilik ilkeleri gereğince belirtilmemiştir.
Talep Miktarı	Ürünlere gelen talep miktarıdır. Araştırma kapsamında bağımsız değişkeni oluşturmaktadır.	Nümerik	-	Firma gizlilik ilkeleri gereğince belirtilmemiştir.

	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Key	Division	Marketing Divisi	Product Group	Product Type	Sportscode	Colour	Gender	Subgroup	Breake	Theme	Diff. R	Comm
2	7499632	FOOTWEAR	Sport Performance	SHOES	SHOES - LOW (NO-FB)	TENNIS	DKMARI	Men	Shoes	Tennis	Men's Response Court	R4	article con
3	7495212	FOOTWEAR	Sport Performance	SHOES	SHOES - LOW (NO-FB)	RUNNING	MEDLEAD	Women	Shoes	Running	Women's Technical	R4	Model cor
4	7499452	FOOTWEAR	Sport Performance	SHOES	SHOES - LOW (NO-FB)	RUNNING	BLACK	Men	Shoes	Running	Men's Versatile	R5	article con
5	3525822	FOOTWEAR	Sport Performance	SHOES	SHOES - LOW (NO-FB)	OUTDOOR	DARKSLATE	Men	Shoes	Outdoor	Men's Versatile	R2	collection
6	8076552	FOOTWEAR	Sport Performance	SHOES	FOOTBALL SHOE-HARD	FOOTBALL/SOCCER	MTSILVER	Men	Shoes	Football	Men's X-Edge - Footwear	R5	collection
7	7494532	FOOTWEAR	Sport Performance	SHOES	SHOES - LOW (NO-FB)	RUNNING	MTSILVER	Men	Shoes	Running	Men's Technical Trail	R3	collection
8	4132843	FOOTWEAR	Sport Performance	SHOES	SHOES - LOW (NO-FB)	RUNNING	BLACK	Men	Shoes	Running	Men's Versatile	R3	collection
9	4514422	FOOTWEAR	Sport Heritage	SHOES	SHOES - LOW (NO-FB)	LIFESTYLE	WHITE	Kids	Kids shoe	LIFESTYL	Running Basics	R4	collection
10	7489083	FOOTWEAR	Sport Performance	SHOES	SHOES - LOW (NO-FB)	INDOOR	DARONX	Men	Shoes	Indoor	Men's Handball	R4	collection
11	4628233	FOOTWEAR	Sport Performance	SHOES	FOOTBALL SHOE-FIRM	FOOTBALL/SOCCER	WHITE	Kids	Shoes	Football	Kids Elite	R5	article con
12	8079413	FOOTWEAR	Sport Performance	SHOES	FOOTBALL SHOE-FIRM	FOOTBALL/SOCCER	WHITE	Men	Shoes	Football	Men's Classics	R5	collection
13	8025574	ACCESSORIES	Sport Performance	BAGS	CROSSOVER BACKPACI	TRAINING	NEONPINK	Unisex	Mobile/I-pod F	BU Acc. - I	Feel adidas	R4	collection
14	8025564	ACCESSORIES	Sport Performance	BAGS	CROSSOVER BACKPACI	TRAINING	ORANCOUNT	Unisex	Mobile/I-pod F	BU Acc. - I	Feel adidas	R4	article con
15	4627393	FOOTWEAR	Sport Heritage	SHOES	SHOES - LOW (NO-FB)	LIFESTYLE	DKNAVY	Kids	Kids shoe	LIFESTYL	SPORT GOOFY	R3	collection
16	8077783	FOOTWEAR	Sport Performance	SHOES	FOOTBALL SHOE-TURF	FOOTBALL/SOCCER	MTSILVER	Men	Shoes	Football	Football Lifestyle	R3	collection
17	4514623	FOOTWEAR	Sport Heritage	SHOES	SHOES - LOW (NO-FB)	LIFESTYLE	INFRARED	Men	Men shoe	LIFESTYL	FOOTBALL	R3	Model cor
18	7492403	FOOTWEAR	Sport Performance	SHOES	SHOES - LOW (NO-FB)	TRAINING	SUN	Kids	Shoes	Kids	Kids Only Explore	R3	article con
19	9487603	ACCESSORIES	Sport Performance	BALLS	BALL (HAND-STITCHED)	FOOTBALL/SOCCER	WHITE	Men	Footballs	Soccer Ha	Elite/WC 2006 balls	R4	article con
20	4132933	FOOTWEAR	Sport Performance	SHOES	FOOTBALL SHOE-HARD	FOOTBALL/SOCCER	WHITE	Kids	Shoes	Football	Kids Elite	R5	collection
21	4132935	FOOTWEAR	Sport Performance	SHOES	FOOTBALL SHOE-HARD	FOOTBALL/SOCCER	WHITE	Kids	Shoes	Football	Kids Elite	R5	collection
22	4478593	FOOTWEAR	Sport Heritage	SHOES	SHOES - LOW (NO-FB)	LIFESTYLE	WHITE	Women	Women shoe	LIFESTYL	Running Track	R2	collection
23	4515383	FOOTWEAR	Sport Performance	SHOES	FOOTBALL SHOE-HARD	FOOTBALL/SOCCER	BLACK	Kids	Shoes	Football	Kids Elite	R5	collection
24	4520723	FOOTWEAR	Sport Performance	SHOES	FOOTBALL SHOE-TURF	FOOTBALL/SOCCER	WHITE	Kids	Shoes	Football	Kids Modern	R5	article con
25	4629331	FOOTWEAR	Sport Performance	SHOES	FOOTBALL SHOE-FIRM	FOOTBALL/SOCCER	BLACK	Kids	Shoes	Football	Kids Elite	R5	article con
26	7487972	FOOTWEAR	Sport Heritage	SHOES	SHOES - LOW (NO-FB)	LIFESTYLE	BLACK	Men	Men shoe	LIFESTYL	Running Track	R2	collection
27	7491352	FOOTWEAR	Sport Performance	SHOES	SHOES - LOW (NO-FB)	RUNNING	BLACK	Women	Shoes	Running	Women's Technical	R5	collection
28	7495453	FOOTWEAR	Sport Performance	SHOES	FOOTBALL SHOE-FIRM	FOOTBALL/SOCCER	WHITE	Kids	Shoes	Football	Kids Modern	R5	Model cor
29	7498232	FOOTWEAR	Sport Performance	SHOES	FOOTBALL SHOE-TURF	FOOTBALL/SOCCER	BLACK	Kids	Shoes	Football	Kids Classics	R5	Model cor
30	7498235	FOOTWEAR	Sport Performance	SHOES	FOOTBALL SHOE-TURF	FOOTBALL/SOCCER	BLACK	Kids	Shoes	Football	Kids Classics	R5	Model cor
31	7498245	FOOTWEAR	Sport Performance	SHOES	FOOTBALL SHOE-TURF	FOOTBALL/SOCCER	WHITE	Kids	Shoes	Football	Kids Classics	R5	Model cor
32	8025584	ACCESSORIES	Sport Performance	BAGS	CROSSOVER BACKPACI	TRAINING	ZENITH	Unisex	Mobile/I-pod F	BU Acc. - I	Feel adidas	R4	collection

Şekil 4.4 : Veri setini oluşturan değişkenlerden bazılarının gösterimi.

1	Sportscode	Colour	Gender	Subgroup	Breako	Theme	Diff. Ra	Commu	NMM S	Origin1
2	TENNIS	DKMARI	Men	Shoes	Tennis	Men's Response Court	R4	article com	13	CHINA
3	RUNNING	MEDLEAD	Women	Shoes	Running	Women's Technical	R4	Model com	9	CHINA
4	RUNNING	BLACK	Men	Shoes	Running	Men's Versatile	R5	article com	9	CHINA
5	OUTDOOR	DARKSLATE	Men	Shoes	Outdoor	Men's Versatile	R2	collection c	15	CHINA
6	FOOTBALL/SOCCER	MTSILVER	Men	Shoes	Football	Men's X-Edge - Footwear	R5	collection c	14	INDONESIA
7	RUNNING	MTSILVER	Men	Shoes	Running	Men's Technical Trail	R3	collection c	14	VIETNAM
8	RUNNING	BLACK	Men	Shoes	Running	Men's Versatile	R3	collection c	14	CHINA
9	LIFESTYLE	WHITE	Kids	Kids shoe	LIFESTYLE	Running Basics	R4	collection c	15	VIETNAM
10	INDOOR	DARONX	Men	Shoes	Indoor	Men's Handball	R4	collection c	15	CHINA
11	FOOTBALL/SOCCER	WHITE	Kids	Shoes	Football	Kids Elite	R5	article com	17	VIETNAM
12	FOOTBALL/SOCCER	WHITE	Men	Shoes	Football	Men's Classics	R5	collection c	17	INDONESIA
13	TRAINING	NEONPINK	Unisex	Mobile/I-pod Po	BU Acc. - B	Feel adidas	R4	collection c	22	CHINA
14	TRAINING	ORANCOUNT	Unisex	Mobile/I-pod Po	BU Acc. - B	Feel adidas	R4	article com	22	CHINA
15	LIFESTYLE	DKNAVY	Kids	Kids shoe	LIFESTYLE	SPORT GOOFY	R3	collection c	22	INDONESIA
16	FOOTBALL/SOCCER	MTSILVER	Men	Shoes	Football	Football Lifestyle	R3	collection c	22	INDONESIA
17	LIFESTYLE	INFRARED	Men	Men shoe	LIFESTYLE	FOOTBALL	R3	Model com	17	INDONESIA
18	TRAINING	SUN	Kids	Shoes	Kids	Kids Only Explore	R3	article com	22	CHINA
19	FOOTBALL/SOCCER	WHITE	Men	Footballs	Soccer Har	Elite/WC 2006 balls	R4	article com	22	Pakistan
20	FOOTBALL/SOCCER	WHITE	Kids	Shoes	Football	Kids Elite	R5	collection c	22	CHINA
21	FOOTBALL/SOCCER	WHITE	Kids	Shoes	Football	Kids Elite	R5	collection c	22	CHINA
22	LIFESTYLE	WHITE	Women	Women shoe	LIFESTYLE	Running Track	R2	collection c	22	VIETNAM
23	FOOTBALL/SOCCER	BLACK	Kids	Shoes	Football	Kids Elite	R5	collection c	22	VIETNAM
24	FOOTBALL/SOCCER	WHITE	Kids	Shoes	Football	Kids Modern	R5	article com	22	VIETNAM
25	FOOTBALL/SOCCER	BLACK	Kids	Shoes	Football	Kids Elite	R5	article com	22	VIETNAM
26	LIFESTYLE	BLACK	Men	Men shoe	LIFESTYLE	Running Track	R2	collection c	18	CHINA
27	RUNNING	BLACK	Women	Shoes	Running	Women's Technical	R5	collection c	22	INDONESIA
28	FOOTBALL/SOCCER	WHITE	Kids	Shoes	Football	Kids Modern	R5	Model com	22	VIETNAM
29	FOOTBALL/SOCCER	BLACK	Kids	Shoes	Football	Kids Classics	R5	Model com	22	VIETNAM

Şekil 4.4 (devam) : Veri setini oluşturan değişkenlerden bazılarının gösterimi.

4.2.2 Veri Dönüştürme

Veri dönüştürme işlemlerinin tamamı Excel programı üzerinde yapılmıştır. Veri dönüşümü yapılırken veri setinin içinde uygulama kapsamında kullanılmayacak olan değişkenler ve boş hücreler göz ardı edilerek silinmiştir. Böylelikle ilk etapta 8864 adet olan ürün sayısı 3548'e düşürülmüştür. Sonrasında, nominal değerleri nümerik değerlere çevirme amacı ile bu değişkenler ikili değişkenlere dönüştürülmüştür. Bunu yapabilmek için her bir değişken için Excel programında özet tablolar oluşturarak değişkenlerin ait olduğu kategoriler sırasıyla numaralandırılmıştır. Numaralandırma işlemi bittikten sonra Excel uygulamasında DEC2BIN(sayı;[basamak]) komutunu kullanarak ikili sayılar elde edilmiştir. Şekil 4.5'te bu uygulamanın alt grup değişkeni ile yapılan bir örneği gösterilmiştir. Şekil 4.6'da dönüşüm sonucunda programa girilmeye hazır hale getirilmiş veri seti bulunmaktadır. Bu yöntemin sonucu olarak bir değişken ikili sayılar cinsinden birden fazla sayı basamakları ile ifade edilebilen bir sayı biçimine dönüştürülmüştür. Ancak dönüştürülen verileri ikili değişkenler olarak ifade edebilme amacı ile her bir değişken içerdiği basamak sayısı kadar alt değişkenlere ayrılmıştır. Bütün nominal değişkenlere bu dönüşüm uygulandıktan sonra beden değişkeni için ise ayrı bir yönteme başvurulmuştur. Beden değişkenine bakıldığında, bu değişken aralıklar cinsinden ifade edildiği görülmüştür. Bu aralıkları firmanın kendi sitesinden alınan cinsiyet, yaş ve ürün kategorisi gözetilerek oluşturulmuş beden tablolarından yararlanılarak verilen aralıkta kaç tane beden olduğu hesaplanmış ve bu bulunan değerler ikili değişken türüne dönüştürülmüştür. Bunun yanı sıra üretim ve gönderi maliyeti değişkeni hem Euro hem de USD cinsinden para birimleri ile ifade edilmiştir. Bu değişkene uygulanan dönüşüm işlemi için ise ilgili olduğu yılın döviz kurları göz önünde bulundurularak Euro USD dönüşümü yapılmıştır. Bu dönüşüm yapıldıktan sonra üretim ve gönderi maliyeti değişkeni artık sadece USD cinsinden tanımlanan bir değişkene dönüşmüş olmaktadır. Nümerik değişkenlere bakıldığında ise sipariş miktarı adet cinsinden, satış fiyatı, üretim ve gönderi maliyetinin de birbirlerinden farklı para birimleri cinsinden olduğu görülmektedir. Regresyon çalışması yaparken bu üç nümerik değişkenin de birbirleri cinsinden ifade edilmesi gerekliliği göz önüne alındığında nümerik değişkenler için de bir normalleştirme çalışması yapılmalıdır. Veri normalleştirme yöntemi olarak yukarıda bahsedilen Min-Maks yöntemi seçilmiştir.

AO10 =DEC2BIN(AN10;6)

	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW
1																	
2																	
3					Count of Key					Count of Key					Count of Key		
4	Toplam		No	Code	Subgroup	Toplam		No	Code	Breakout	Toplam		No	Code	Theme	Toplam	
5	98		1	0000001	Accessories	22		1	000001	Basketball	137		1	0000001	3SA Clima Dynamic Classic	64	
6	33		2	0000010	Backpack	65		2	000010	Basketballs SS 2006	36		2	0000010	3-Stripes	101	
7	121		3	0000011	Basic T-shirt	291		3	000011	Bodywear/Intimates	18		3	0000011	Accessories	16	
8	241		4	0000100	Basketballs	36		4	000100	BU Acc. - Bags	306		4	0000100	Active Intimates	36	
9	1854		5	0000101	Beach Volleyballs	4		5	000101	BU Acc. - Bottles	4		5	0000101	BASKETBALL	124	
10	447		6	0000110	Belt	4		6	000110	BU Acc. - Socks	48		6	0000110	Beach	14	
11	754		7	0000111	Bermuda	7		7	000111	BU Acc. Headwear	122		7	0000111	Beach Men - 3 Stripes	16	
12	3548		8	0001000	Bikini	47		8	001000	BU Accessories LICENSED	7		8	0001000	Beach Men - adidas wave	44	
13			9	0001001	Boxer	6		9	001001	BU Accessories/Hardware - In Line	7		9	0001001	Beach Men - Beach Basics	29	
14			10	0001010	Cap	156		10	001010	Evolution	82		10	0001010	Beach Men - Beach Soccer/WC	18	
15			11	0001011	Carrybag	9		11	001011	Football	217		11	0001011	Beach Women - Basic+Contemporary	50	
16			12	0001100	Crib shoe	29		12	001100	Football Generic Apparel	76		12	0001100	Boys 3SA Clima	36	
17			13	0001101	Dress	5		13	001101	Football Licensed Apparel	2		13	0001101	Boys Stripes	11	
18			14	0001110	FB accessories	5		14	001110	Football Licensed Apparel - World Cup 2006 OLP	61		14	0001110	Boys Tracksuits	18	
19			15	0001111	Footballs	114		15	001111	Indoor	21		15	0001111	Boys Training and Separates	18	
20			16	0010000	GK Gloves	15		16	010000	Indoor Hardware SS06	12		16	0010000	COASTAL	67	
21			17	0010001	Goggles	5		17	010001	Kids	155		17	0010001	CORE BASICS	72	
22			18	0010010	Gym Bag	12		18	010010	Lic. Acc Football Licensed	31		18	0010010	Corporate	54	
23			19	0010011	Handballs	2		19	010011	LIFESTYLE	644		19	0010011	COURT CLASSICS	31	
24			20	0010100	Headband	3		20	010100	NY Yankees	6		20	0010100	Dassler	11	
25			21	0010101	Infant shoe	49		21	010101	Outdoor	47		21	0010101	DB Predator	18	
26			22	0010110	Jacket	15		22	010110	Running	256		22	0010110	Dynamic	17	
27			23	0010111	Jersey	45		23	010111	SB/SC Kids	129		23	0010111	Dynamic Boys	5	
28			24	0011000	Kids shoe	51		24	011000	Slides	156		24	0011000	Dynamic Girls	8	
29			25	0011001	L/S Tee	20		25	011001	Soccer Hardware SS 2006	154		25	0011001	Dynamic Pulse Men	10	
30			26	0011010	Make-up bag	1		26	011010	Sport Basics Men	312		26	0011010	Dynamic Pulse Women	5	

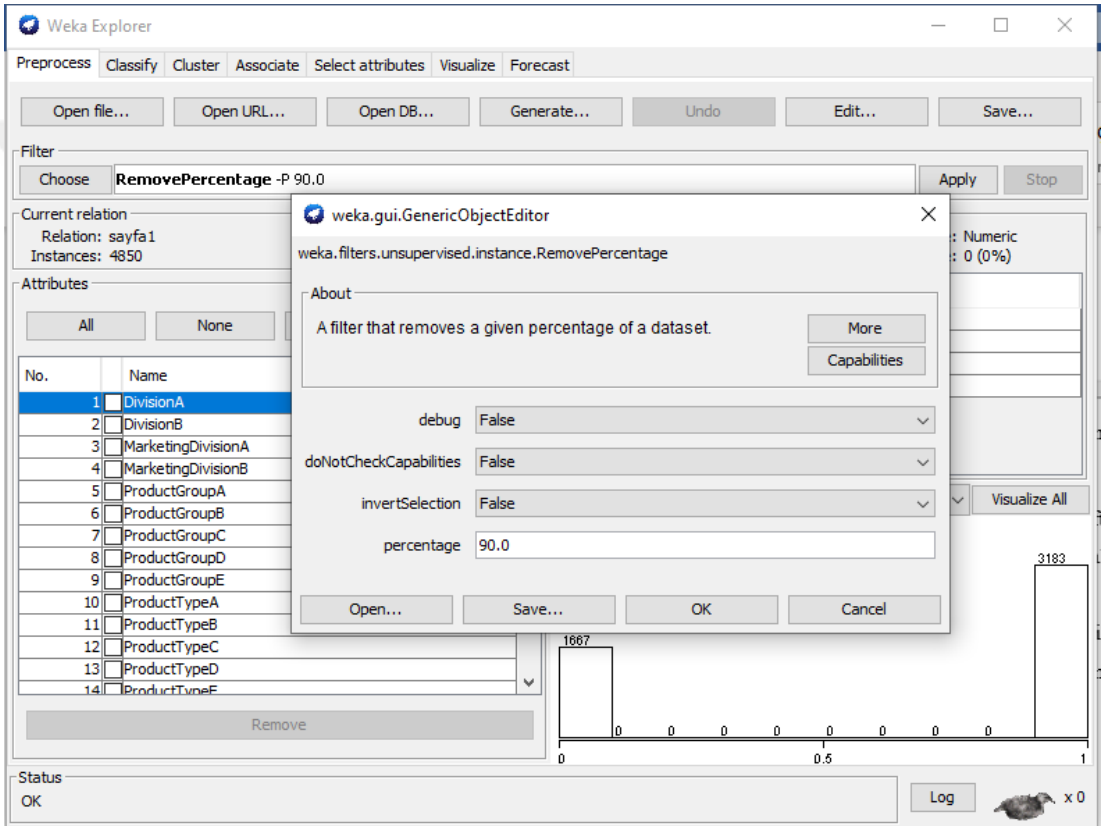
Şekil 4.5 : Veri setinde nominal değerlerin ikili değerlere dönüştürülmesi.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
DivisionA	DivisionB	MarketingDivisionA	MarketingDivisionB	ProductGroupA	ProductGroupB	ProductGroupC	ProductGroupD	ProductGroupE	ProductTypeA	ProductTypeB	ProductTyp	ProductTyp	ProductTyp	Producti
0	1	1	0	0	1	1	0	0	0	1	1	0	0	
0	1	1	0	0	1	1	0	0	0	1	1	0	0	
0	1	1	0	0	1	1	0	0	0	1	1	0	0	
0	1	1	0	0	1	1	0	0	0	1	1	0	0	
0	1	1	0	0	1	1	0	0	0	0	1	0	0	
0	1	1	0	0	1	1	0	0	0	1	1	0	0	
0	1	1	0	0	1	1	0	0	0	1	1	0	0	
0	1	0	1	0	1	1	0	0	0	1	1	0	0	
0	1	1	0	0	1	1	0	0	0	1	1	0	0	
0	1	1	0	0	1	1	0	0	0	0	1	0	0	
0	1	1	0	0	1	1	0	0	0	0	1	0	0	
1	1	1	0	0	0	0	1	0	0	0	0	1	1	
1	1	1	0	0	0	0	1	0	0	0	0	1	1	
0	1	0	1	0	1	1	0	0	0	1	1	0	0	
0	1	1	0	0	1	1	0	0	0	0	1	0	1	
0	1	0	1	0	1	1	0	0	0	1	1	0	0	
0	1	1	0	0	1	1	0	0	0	1	1	0	0	
1	1	1	0	0	0	0	1	1	0	0	0	0	1	
0	1	1	0	0	1	1	0	0	0	0	1	0	0	
0	1	1	0	0	1	1	0	0	0	0	1	0	0	
0	1	0	1	0	1	1	0	0	0	1	1	0	0	
0	1	1	0	0	1	1	0	0	0	0	1	0	0	
0	1	1	0	0	1	1	0	0	0	0	1	0	1	
0	1	1	0	0	1	1	0	0	0	0	1	0	0	
0	1	0	1	0	1	1	0	0	0	1	1	0	0	
0	1	1	0	0	1	1	0	0	0	1	1	0	0	
0	1	1	0	0	1	1	0	0	0	0	1	0	0	
0	1	1	0	0	1	1	0	0	0	0	1	0	1	
0	1	1	0	0	1	1	0	0	0	0	1	0	1	

Şekil 4.6 : Dönüşüm sonucunda programa girilmesi hazır hale getirilmiş veri seti.

4.2.3 Eğitim ve test veri kümelerinin belirlenmesi

Veriler %90 eğitim ve %10 test verisi olacak şekilde ikiye ayrılarak incelenmiştir. Bu ayrım kullanıcı tarafından değil, Weka programının filtre kısmındaki Unsupervised sekmesindeki Remove Percentage seçeneği kullanılarak yapılmıştır. Bu süreç Şekil 4.7'de açıkça belirtilmektedir. Bu özellik sayesinde de 3548 adet veri 3194 adet eğitim ve 354 adet test verisi olacak şekilde ikiye ayrılmıştır. Bir sonraki bölümde de eğitim veri setine çok katmanlı algılayıcı ve rassal orman algoritması uygulanıp sonuçlar karşılaştırılarak en iyi yöntem bulunmuştur.



Şekil 4.7 : Veri setinin eğitim ve test veri seti olarak bölünmesi.

4.3 Bir Sezonluk Veri ile Regresyon ile Satış Tahmini Sonuçları

4.3.1 Çok katmanlı algılayıcı algoritması için kullanılan parametreler

Çok katmanlı algılayıcı parametreleri için Bölüm 3.1.5'te bahsedildiği üzere optimal bir seçim yapılabilmesi için bir yöntem yoktur. Bunun için kullanılacak parametrelere deneme yanılma yönteminin verdiği sonuçlara göre karar verilmiştir.

Çizelge 4.2 : Çok katmanlı ağ mimarisi.

ÇKA Mimarisi	
Öğrenme Oranı	0.05, 0.1, 0.5, 0.8
Gizli Katman Sayısı	1, 2
Gizli Katmandaki Nöron Sayısı	50, 60, 70
Momentum Oranı	0.7
Normalizasyon	Tüm nümerik değişkenler 0-1 arasında Min-Max kuralı ile normalleştirilmiştir.
Eğitim Kuralı	Geri yayılma algoritması

Çizelge 4.3'te çok katmanlı algılayıcı için deneme yanılma yöntemi ile elde edilen performans değerleri verilmiştir. Deneme yanılma yönteminin kullanılmasının sebebi yukarıdaki bölümlerde anlatıldığı gibi parametre seçimi için optimum bir yöntemin bulunmayışından kaynaklanmaktadır. Yapılan tüm denemelerin sonunda algoritmanın performans değerlerini en iyileyen parametreler, öğrenme oranı 0.05, momentum oranı 0.7 ve gizli katmandaki nöron sayısı 70 olacak şekilde belirlenmiştir. Değerlendirme kriterleri olarak korelasyon katsayısı, MAE ve RMSE değerleri baz alınmıştır.

Çizelge 4.3 : Çok katmanlı algılayıcı parametre seçimi.

Öğrenme Oranı	Momentum Oranı	Gizli Katman Sayısı	Gizli Katmandaki Nöron Sayısı	Korelasyon Katsayısı (R^2)	MAE (%)	RMSE (%)
0.05	0.7	1	50	0.938	0.0108	0.0142
0.05	0.7	1	60	0.9554	0.0093	0.012
0.05	**0.7**	**1**	**70**	**0.9873**	**0.0087**	**0.0094**
0.05	0.7	1	100	0.9675	0.0085	0.0107
0.1	0.7	1	50	0.8567	0.0169	0.0223
0.1	0.7	1	60	0.9352	0.0249	0.0241
0.1	0.7	1	70	0.9408	0.019	0.022
0.1	0.7	1	100	0.9629	0.009	0.0114
0.5	0.7	1	50	0.8415	0.0212	0.0266
0.5	0.7	1	60	0.8849	0.0179	0.022
0.5	0.7	1	70	0.9274	0.0125	0.0159
0.5	0.7	1	100	0.9645	0.0082	0.0106
0.8	0.7	1	50	0.8723	0.0151	0.0201
0.8	0.7	1	60	0.749	0.0235	0.0298
0.8	0.7	1	70	0.9374	0.0126	0.0159
0.8	0.7	1	100	0.963	0.0083	0.0109
0.05	0.7	2	40, 50	0.9339	0.0116	0.0152
0.05	0.7	2	50, 40	0.9655	0.0096	0.0118
0.1	0.7	2	40, 50	0.9435	0.0137	0.0167
0.1	0.7	2	50, 40	0.9265	0.013	0.0165
0.5	0.7	2	40, 50	N/A	N/A	N/A
0.5	0.7	2	50, 40	0.0406	0.304	0.412
0.8	0.7	2	40, 50	0	0.3	0.46
0.8	0.7	2	50, 40	N/A	N/A	N/A
0.05	0.3	1	50	0.9522	0.0088	0.012
0.05	0.3	1	60	0.9762	0.0087	0.0109
0.05	0.3	1	70	0.9655	0.0079	0.0104
0.1	0.3	1	50	0.9477	0.0131	0.0164
0.1	0.3	1	60	0.9604	0.008	0.0109
0.1	0.3	1	70	0.9647	0.0085	0.0108
0.5	0.3	1	50	0.9380	0.0119	0.015
0.5	0.3	1	60	0.9574	0.0086	0.0114
0.5	0.3	1	70	0.9523	0.0166	0.0196
0.8	0.3	1	50	0.9172	0.0133	0.017
0.8	0.3	1	60	0.9255	0.0152	0.0187
0.8	0.3	1	70	0.9647	0.0085	0.0108
0.05	0.3	2	40, 50	0.9563	0.0099	0.0125
0.05	0.3	2	50, 40	0.9559	0.0118	0.0146
0.1	0.3	2	40, 50	0.9487	0.0099	0.0125
0.1	0.3	2	50, 40	0.9553	0.0091	0.0117
0.5	0.3	2	40, 50	0.5667	0.0397	0.0463
0.5	0.3	2	50, 40	0.7811	0.0235	0.0299
0.8	0.3	2	40, 50	0.0079	0.3333	0.463
0.8	0.3	2	50, 40	0.1236	0.0323	0.0405

Çizelge 4.4 : ÇKA ile test veri seti kullanılarak yapılan tahmin sonuçları.

Division	Product Type	Sport Code	Colour	Theme	Tahmin Değeri	Gerçek Değer
Footwear	Shoes Low	Lifestyle	Black	Football basics	1080	1150
Textiles	Tshirt (Shortsleeve)	Sport Basics	White	Essentials Basics	1103	1151
Textiles	Tracksuit	Training	Dark Navy	Shoes	1154	1152
Footwear	Shoes Low	Lifestyle	Black	Shoes	1085	1155
Textiles	Graphic T (Short Sleeve)	Training	Lightonix	Clima Dynamic Classic	1241	1300
Footwear	Football shoe-turf	Football/Soccer	Cardin	Kid’s Elite	1296	1305
Footwear	Football shoe-turf	Football/Soccer	White	Men’s Elite	1267	1305
Textiles	Swim shorts	Swim	Sun	Beach men 3 stripes	1251	1308
Textiles	Cap	Training	Putty	Corporate	1299	1309
Footwear	Shoes Low	Lifestyle	White	Motors	1738	1805
Textiles	Pants	Training	Black	Essentials Basic	1806	1818
Footwear	Shoes Low	Lifestyle	Espresso	Coastal	1760	1825
Textiles	Tshirt (Shortsleeve)	Running	White	3-Stripes	1830	1826
Footwear	Shoes Low	Lifestyle	White	Running Basics	1852	1980
Footwear	Football shoe-turf	Football/Soccer	White	Kids Elite	1845	1840
Textiles	Cap	Training	Dark Navy	Corporate	1790	1862

Çizelge 4.4’te belirlenen en iyi parametreler ile test veri seti üzerinde yapılan tahmin çalışması sonuçları verilmiştir. Bu tahminin performans değerleri 0.9962 korelasyon katsayısı, 0.0069 MAE değeri ve 0.0087 RMSE değerleridir ve oldukça tatmin edici bir sonuçtur.

4.3.2 Rassal orman algoritması için kullanılan parametreler

Rassal orman algoritması için Weka programında kullanıcıya sadece seed (çekirdek) sayısı belirleme opsiyonu sunulmuştur. Programda bu değere bir değer sağlanırsa, model yerleştirilmeden önce Weka'nın rastgele sayı üretecini başlatmak için kullanılır, aksi halde, rasgele sayı üreteci rastgele program tarafından belirlenen bir zamandan başlatılır. Bu parametre, istenirse model birleştirme işlemini hassas bir şekilde kullanıcı tarafından kontrol edilebilmesini sağlar. Rasgele ormanlar rasgele veri seçimlerine dayandığından, varsayılan davranış, model her çalıştırıldığında farklı bir ormanın oluşturulmasıdır. Bunun önüne geçmek için rasgele sayı üretecinin çekirdek sayısı belirlenebilir. Bu sayede, giriş verilerinin veya diğer parametrelerin değiştirmediği varsayılarak, bu aracın her çalıştırılmasıyla aynı rasgele seçim sırasının gerçekleştirilmesine neden olur. Çizelge 4.5'te görüldüğü üzere çekirdek sayısının belirlenmesi konusunda belirlenen sayılar değerlendirme kriterleri göz önüne alındığında çok büyük fark yaratmamıştır.

Çizelge 4.5 : Rassal orman algoritması parametreleri.

Seed (Çekirdek) Sayısı	Korelasyon Katsayısı (R^2)	Eğitim MAE (%)	RMSE (%)
20	0.9782	0.01	0.0126
40	0.9785	0.01	0.0126
70	0.978	0.0101	0.0125
90	0.9785	0.01	0.0126
100	**0.979**	**0.01**	**0.0125**

Çizelge 4.6 : RF ile test veri seti kullanılarak yapılan tahmin sonuçları.

Division	Product Type	Sport Code	Colour	Theme	Tahmin Değeri	Gerçek Değer
Footwear	Shoes Low	Lifestyle	Black	Football basics	1065	1066
Textiles	Tshirt (Shortsleeve)	Sport Basics	White	Essentials Basics	1096	1096
Textiles	Tracksuit	Training	Dark Navy	Shoes	1093	1100
Footwear	Shoes Low	Lifestyle	Black	Shoes	1109	1120
Textiles	Graphic T (Short Sleeve)	Training	Lighton ix	Clima Dynamic Classic	1138	1140
Footwear	Football shoe-turf	Football/Soccer	Cardin	Kid's Elite	1212	1220
Footwear	Football shoe-turf	Football/Soccer	White	Men's Elite	1224	1230
Textiles	Swim shorts	Swim	Sun	Beach men 3 stripes	1376	1382
Textiles	Cap	Training	Putty	Corporate	1445	1450
Footwear	Shoes Low	Lifestyle	White	Motors	1497	1500
Textiles	Pants	Training	Black	Essentials Basic	1502	1505
Footwear	Shoes Low	Lifestyle	Espress o	Coastal	1554	1560
Textiles	Tshirt (Shortsleeve)	Running	White	3-Stripes	1600	1609
Footwear	Shoes Low	Lifestyle	White	Running Basics	1760	1765
Footwear	Football shoe-turf	Football/Soccer	White	Kids Elite	2296	2305
Textiles	Cap	Training	Dark Navy	Corporate	1769	1802

Çizelge 4.6'da algoritma için belirlenen parametreler ile test veri seti üzerinde yapılan tahmin çalışması sonuçları verilmiştir. Bu sonuçlara göre bir o ek olarak gelecek ürünün bazı değişken bilgileri ve miktarı tespit edilebilir. Bu tahminin performans değerleri 0.9778 korelasyon katsayısı, 0.0107 MAE değeri ve 0.0104 RMSE değerleridir ve oldukça tatmin edici bir sonuçtur.

4.3.3 İki yöntem sonuçlarının karşılaştırılması ve iyi olanın seçimi

İki yöntemin de eğitim veri seti üzerinde bir kez çalıştırılmasının sonucunda çok katmanlı algılayıcının daha iyi sonuç verdiği gözlemlenmiştir. Ancak bunun tesadüfi bir sonuç olup olmadığı sorusunu gidermek amacı ile iki algoritma da Weka tarafından rastgele seçim ile oluşturulan %90 ve %10'luk 10 adet veri setleri ile çalıştırılıp bulunan performans değerlerine *t*-testi yapılarak en iyi algoritma belirlenmeye çalışılmıştır. Çizelge 4.7'de her iki algoritma ile performans değerleri gözetilerek karşılaştırılan toplam 10 denemenin regresyon sonuçları mevcuttur.

Çizelge 4.7 : ÇKA ve RF ile yapılan regresyon çalışmalarının sonuçları.

Deneme	Korelasyon Katsayısı (R^2)		MAE		RMSE	
	Çok Katmanlı Algılayıcı	Rassal Orman Algoritması	Çok Katmanlı Algılayıcı	Rassal Orman Algoritması	Çok Katmanlı Algılayıcı	Rassal Orman Algoritması
1	0.97	0.969	0.007	0.0105	0.0099	0.0126
2	0.975	0.965	0.01	0.014	0.0129	0.0132
3	0.976	0.9701	0.008	0.01	0.0104	0.0123
4	0.9741	0.9687	0.0057	0.0088	0.0122	0.0129
5	0.9787	0.9697	0.0076	0.0102	0.01	0.0124
6	0.9792	0.9761	0.0036	0.0099	0.0047	0.0075
7	0.9743	0.967	0.0095	0.0061	0.0122	0.0123
8	0.9721	0.9701	0.0072	0.0109	0.0115	0.0127
9	0.9689	0.9579	0.0066	0.008	0.0089	0.0135
10	0.9745	0.9687	0.0084	0.011	0.0102	0.0124

Çizelge 4.8 : Korelasyon katsayısı temel alınarak yapılan *t*-testi sonuçları.

Hipotez	Performans Kriteri	T değeri	P(T<=t)	$T_{0.05}$
H_0: $\rho_{ÇKA} \leq \rho_{RF}$ H_1: $\rho_{ÇKA} > \rho_{RF}$	Korelasyon Katsayısı	2.54	0.00015	1.83

Çizelge 4.8'de H_o hipotezi çok katmanlı algılayıcı korelasyon katsayısının rassal orman algoritmasının korelasyon katsayısından küçük ve eşit olması üzerine, H_1 hipotezi ise çok katmanlı algılayıcının korelasyon katsayısının rassal orman algoritmasının korelasyon katsayısından daha büyük olacağı üzerine kurulmuştur. Yapılan *t*-testinde anlamlılık düzeyi 0,05 olarak kullanıcı tarafından belirlenmiştir. Her iki algoritma korelasyon katsayısı kriterine göre birbirleri ile karşılaştırılmıştır. Karşılaştırmanın anlamlılık düzeyi 0.00015 olmakla birlikte ilk adımda belirlenen anlamlılık düzeyinden daha küçüktür ve *t* değeri de tablo değeri ile kıyaslanmış ve 2.54 değeri ile tablo değerinden daha yüksek bir değer elde edildiği için H_o hipotezi reddedilir. Bunun anlamı da yapılan her deneme için çok katmanlı algılayıcının rassal orman algoritmasına göre daha iyi sonuçlar verdiğinin bir kanıtı niteliğindedir.

4.4 Ürün Kümeleme ile Satış Tahmini Sonuçları

4.4.1 *K*-ortalamalar algoritması ile kümeleme

Kümeleme çalışmasına duyulan gereklilik Bölüm 3.3.1'de açıklanmıştır. Bu süreçte izlenecek yöntem birbirine benzer özellik gösteren verileri aynı kümelerde gruplandıracak bir sistem geliştirmektir. Burada göz önünde bulundurulması gereken en önemli konu *K*-ortalamalar algoritmasının optimum küme sayısını belirlemek için herhangi bir yöntemi bulunmamasından dolayı program her çalıştırıldığında farklı küme değerleri verebilmesinden dolayı ortaya çıkan tutarsızlık olmaktadır.

K-ortalamalar algoritmasında kullanılan küme sayısı sağlıklı bir performans karşılaştırılması yapılabilmesi adına, küme sayısını veri sayısı ile orantılı bir biçimde belirleyebilen bir alt yapı ile çalışan *X*-ortalamalar algoritmasında belirlenen küme sayısı ile aynı tutulmuştur. Buna göre 3548 adet veri ve 69 adet kriter için belirlenen küme sayısı *X*-ortalamalar algoritmasına göre 4 olarak belirlenmiştir. 4 adet küme için 3548 adet verinin %16'sı birinci kümeye, %24'ü ikinci kümeye, %31'i üçüncü kümeye ve geri kalanı da son kümeye atanmıştır. Bir sonraki aşama olarak her bir küme için bir önceki çalışmada tespit edilen en iyi parametreler ile çok katmanlı algılayıcı algoritması kullanılacaktır. Bunun için her küme Excel programında ayrı ayrı dosyalanır ve Weka'ya tekrar veri dönüşümü işlemleri uygulanarak yeniden girilir. Çizelge 4.9'da kümeler içerisinde yer alan ürünlerin ayrıntılı gösterimi belirtilmiştir.

Çizelge 4.9 : *K*-ortalamalar algoritmasına göre her bir kümeye atanan ürünler.

Ürünler	Küme 1	Küme 2	Küme 3	Küme 4
Apparel Accessories	4	12	0	8
Bags	29	33	0	**307**
Balls	0	**162**	0	0
Hardware Accessories	5	7	12	12
Headwear	6	18	0	**137**
Jackets	16	10	0	12
Jerseys	0	**44**	0	1
Other Shirts	5	**28**	0	2
Pants	12	**171**	0	2
Polo Shirts	1	**56**	0	0
Protection Gear	0	**36**	0	0
Shoes	**489**	0	**1070**	0
Shorts	4	**101**	0	8

Çizelge 4.9 (devam) : *K*-ortalamalar algoritmasına göre her bir kümeye atanan ürünler.

Ürünler	Küme 1	Küme 2	Küme 3	Küme 4
Skirts / Dresses	4	8	2	3
Socks	0	1	0	**48**
Suits	0	4	0	**94**
Sweaters	0	0	0	2
Sweatshirts	0	4	0	7
Swimwear	2	**138**	0	4
Tights	0	0	0	1
Tops	3	24	0	18
T-shirts	2	3	0	**356**
Genel Toplam	582	860	1084	1022

4.4.2 *X*-ortalamalar algoritması ile kümeleme

K-ortalamalar algoritmasının verdiği sonuçlar ile karşılaştırılması için *X*-ortalamalar yönteminin sebebi yukarıdaki bölümde açıklandığı üzere küme sayısının belirlenmesi şekli olmaktadır. Belirtildiği gibi *X*-ortalamalar algoritması küme sayısının belirlenmesi için bir kullanıcıya ya da karar vericiye ihtiyaç duymamaktadır, bu sebeple *K*-ortalamalar algoritmasından kullanım açısından daha elverişlidir. Çizelge 4.10'da *X*-ortalamalar algoritması ile yapılan kümeleme çalışmasının sonuçları verilmiştir. Küme içindeki verilerin ayrıntılı olarak görüntülenebilmesi için eldeki veriler bir Excel dosyasına aktarılmıştır, orada da gerekli veri dönüşümü işlemleri yapılarak bir sonraki süreçte tekrar Weka programına ayrı ayrı dört dosya halinde girişleri yapılarak Çizelge 4.10'da gösterilmiştir.

Çizelge 4.10 : *X*-ortalamalar algoritmasına göre her bir kümeye atanan ürünler.

Ürünler	Küme 1	Küme 2	Küme 3	Küme 4
Apparel	0	1	11	12
Accessories				
Bags	0	17	**310**	42
Balls	0	0	0	**162**
Hardware	6	5	11	14
Accessories				
Headwear	0	3	**140**	18
Jackets	0	8	20	10
Jerseys	0	0	1	44
Other Shirts	0	5	2	28
Pants	0	11	2	**172**

Çizelge 4.10 (devam) : *X*-ortalamalar algoritmasına göre her bir kümeye atanan ürünler.

Ürünler	Küme 1	Küme 2	Küme 3	Küme 4
Polo Shirts	0	1	1	**55**
Protection Gear	0	0	0	**36**
Shoes	**1070**	**489**	0	0
Shorts	**0**	4	8	**101**
Skirts / Dresses	2	2	5	8
Socks	0	0	**48**	1
Suits	0	0	**94**	4
Sweaters	0	0	2	0
Sweatshirts	0	0	7	4
Swimwear	0	0	4	**140**
Tights	0	0	1	0
Tops	0	2	19	24
T-shirts	0	0	**358**	3
Genel Toplam	1078	548	1044	878

4.4.3 Kümeleme algoritmaları performans değerlendirmesi ve en iyi olanın seçimi

K-ortalamalar ve *X*-ortalamalar algoritmaları aynı küme sayıları ile çalıştırılmıştır ve elde edilen sonuçlar yukarıdaki çizelgelerde belirtilmiştir. En iyi kümeleme algoritmasının seçimi için rand endeksi değerleri hesaplanmış olup Çizelge 4.11'de hesaplama değerleri gösterilmiştir. Rand endeks değeri 0 ile 1 arasında bir değer alır. Endeksin 0 olması verilerin kümelere atanırken tamamen rastlantısal olarak atıldığını küme elemanları arasında bir benzerlik ya da bağlantı bulunmadığını gösterir. Değerin 1 ve 1'e yakın olması ise verilerin kümelere dağıtılmasında başarılı olduğu, benzer ürünleri benzer ürünlerle eşleştirdiği anlamına gelmektedir. Bundan dolayı her iki algoritmadan hangisinin rand endeks değeri daha yüksek ise regresyon çalışmasına o algoritma ile devam edilir. Bu bilgiler ışığında tercih edilen algoritma *K*-ortalamalar algoritmasıdır.

Çizelge 4.11 : Her iki kümeleme için rand endeksi değerleri.

Kümeleme Yaklaşımı	Toplam İkili Sayısı	Gerçek Pozitif	Gerçek Negatif	Hatalı Pozitif	Hatalı Negatif	Rand Endeksi
K-ortalamalar	6292378	866505	4092292	780653	552928	**0.7881**
X-ortalamalar	6292378	738986	3952101	920844	680447	0.7455

4.4.4 En iyi kümeleme algoritması ile belirlenen kümelere uygulanan regresyon sonuçları

Çizelge 4.12'de görüldüğü gibi kümeleme çalışmasının sonucunda da tatmin edici sonuçlar elde edilmiştir. Gerek korelasyon katsayısı gerek MAE ve RMSE değerleri göz önüne alındığında hata oranının memnun edici düzeyde düşük olduğu görülmektedir. Ancak tüm verinin birlikte ele alındığı bir önceki uygulamanın performans değerleri kümeleme performans değerlerinden daha iyi bir sonuç vermiştir. Buradan da kümelemenin tahmin performansı iyileştirilmediği sonucuna varılmıştır.

Çizelge 4.12 : Her bir küme elemanı için bulunan regresyon sonuçları.

İncelenen Veri Kümesi	Korelasyon Katsayısı (R^2)	MAE (%)	RMSE (%)	Küme İçerisindeki Ürün Sayısı
Küme 1	0.9823	0.0105	0.0134	582
Küme 2	0.9796	0.0146	0.0163	861
Küme 3	0.9784	0.014	0.0185	1084
Küme 4	0.9703	0.0284	0.0215	1022
Tek Sezon Tüm Veri	**0.9962**	**0.0069**	**0.0087**	

4.5 Seçilen Bir Ürün İçin Sezonlar Arası Regresyon Çalışması ile Tahminleme

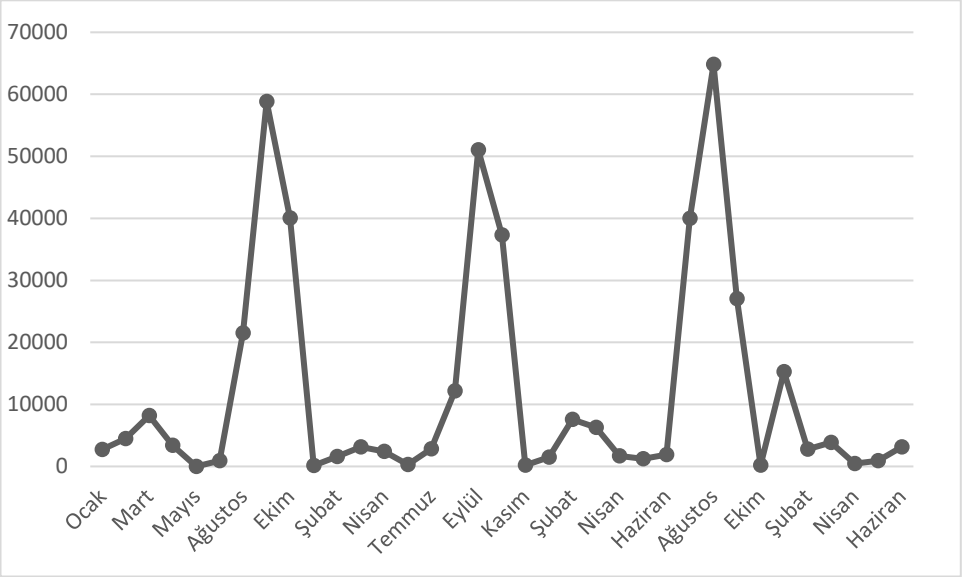
Bu kısımda önceden belirlenen bir ürünün gelecek sezondaki talebi tahmin edilmeye çalışılmıştır. Burada da kullanılacak olan yöntem çok katmanlı algılayıcı ve rassal orman algoritmalarından en iyi performansı veren algoritmaya göre belirlenmiştir. Sezonlar arası regresyon analizinde kullanılacak olan ürün, ürün grubu kategorilerinden seçilen ceket ürünüdür. Birinci aşamada, veri setinin anlamlı hale getirilebilmesi açısından bazı kriterler elenmiştir. Bunun sebebi veri setinde her bir sezondaki kriterler birbiriyle aynı olmamaktadır.

Bundan dolayı tüm sezonlar için ortak olan kriterler belirlenerek veri setinde sadeleştirmeye gidilmiştir. En son hal olarak veri setinde kullanılacak kriterler; sezon, pazarlama bölümü, ürünün ilgili olduğu spor kodu, cinsiyet, ürün alt grubu, perakende ayı, gönderim maliyeti, satış fiyatı ve sipariş miktarı olarak belirlenmiştir. Veri dönüştürme işlemleri yukarıda anlatılan süreçteki gibi yapılmış olup normalizasyon

işlemi ile son bulmuştur. 2004 ve 2005 yıllarındaki dört sezon ele alınarak 2006 yılındaki ilkbahar yaz sezonundaki ceket miktarı tahmin edilmeye çalışılmıştır.

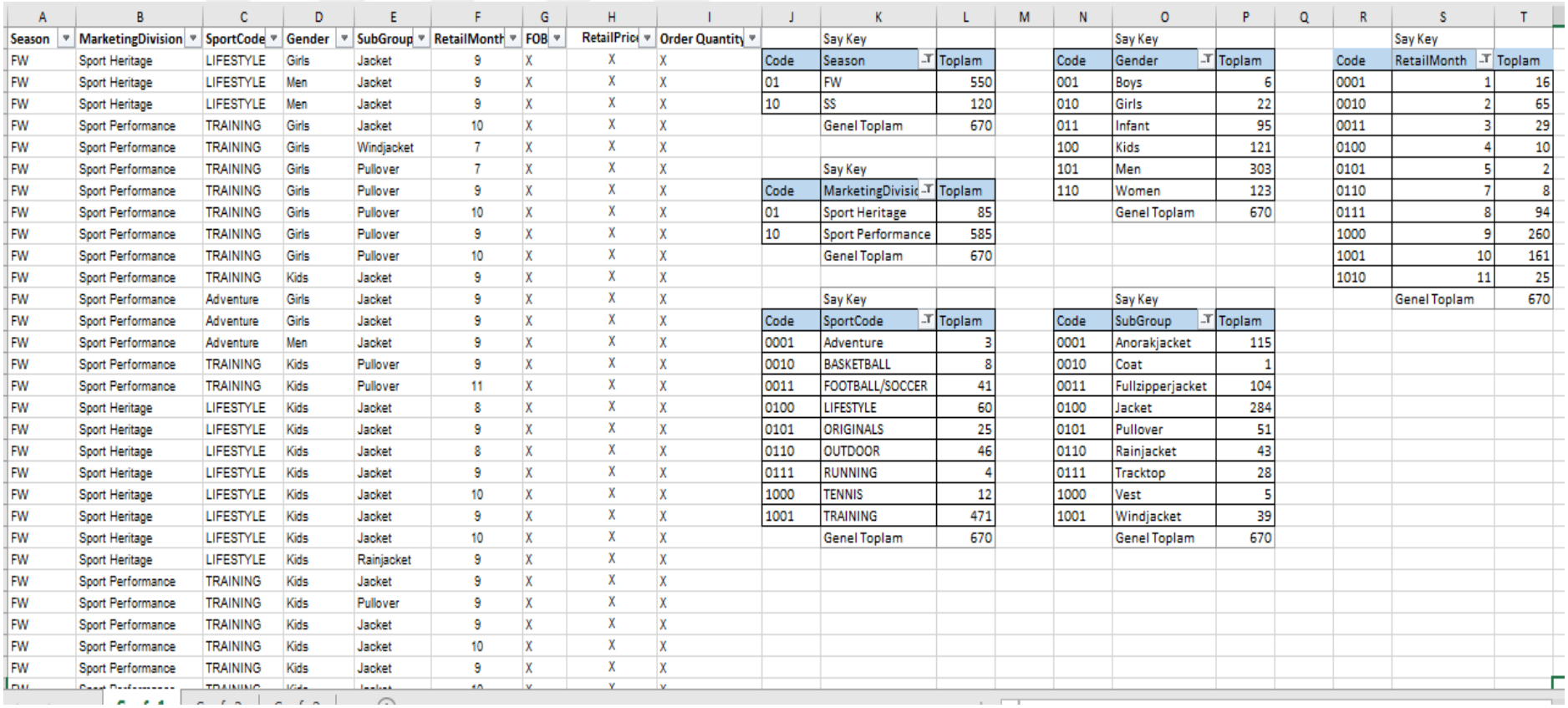
Bulunan tahmin sonucu 2006 yılı ilkbahar yaz sezonunda gerçekleştirilen ceket üretim miktarları ile karşılaştırılmış ve sonucun tutarlı olduğu gözlemlendiği için 2007 sonbahar kış sezonundaki ceket talep miktarı tahmin edilmeye amacıyla sırasıyla 2005 ve 2006 yıllarındaki sezon verileri kullanılmıştır. Şekil 4.8'de 2004-2007 yılları arası gerçekleşen ceket satışlarını göstermektedir.

Ceket satışlarının aylık grafiği incelendiğinde satışların mevsimsel özellik gösterdiği açıkça görülmektedir. Satışlar her yıl eylül-ocak ayları arası ivme kazanmaktadır.



Şekil 4.8 : Ceket ürününün aylık bazdaki talep grafiği.

Şekil 4.9'da veri seti dönüşümü işlemi belirtilmiştir. Bu aşamada her yıl için ortak olan bağımsız değişkenler belirtilmiş ve yeni bir değişken olarak da sezon adlı bir değişken oluşturulmuştur. Belirlenen değişkenlere bir önceki uygulamada yapıldığı gibi nominal nümerik dönüşümleri yapılmıştır. Bütün veri dönüşüm işlemleri Excel programı üzerinde gerçekleştirilmiştir. Firma gizlilik ilkeleri gereğince miktar ve birimi para olan değerler gösterilmemiştir.



Season	MarketingDivision	SportCode	Gender	SubGroup	RetailMonth	FOB	RetailPric	Order Quantity
FW	Sport Heritage	LIFESTYLE	Girls	Jacket	9	X	X	X
FW	Sport Heritage	LIFESTYLE	Men	Jacket	9	X	X	X
FW	Sport Heritage	LIFESTYLE	Men	Jacket	9	X	X	X
FW	Sport Performance	TRAINING	Girls	Jacket	10	X	X	X
FW	Sport Performance	TRAINING	Girls	Windjacket	7	X	X	X
FW	Sport Performance	TRAINING	Girls	Pullover	7	X	X	X
FW	Sport Performance	TRAINING	Girls	Pullover	9	X	X	X
FW	Sport Performance	TRAINING	Girls	Pullover	10	X	X	X
FW	Sport Performance	TRAINING	Girls	Pullover	9	X	X	X
FW	Sport Performance	TRAINING	Girls	Pullover	10	X	X	X
FW	Sport Performance	TRAINING	Kids	Jacket	9	X	X	X
FW	Sport Performance	Adventure	Girls	Jacket	9	X	X	X
FW	Sport Performance	Adventure	Girls	Jacket	9	X	X	X
FW	Sport Performance	Adventure	Men	Jacket	9	X	X	X
FW	Sport Performance	TRAINING	Kids	Pullover	9	X	X	X
FW	Sport Performance	TRAINING	Kids	Pullover	11	X	X	X
FW	Sport Heritage	LIFESTYLE	Kids	Jacket	8	X	X	X
FW	Sport Heritage	LIFESTYLE	Kids	Jacket	9	X	X	X
FW	Sport Heritage	LIFESTYLE	Kids	Jacket	8	X	X	X
FW	Sport Heritage	LIFESTYLE	Kids	Jacket	9	X	X	X
FW	Sport Heritage	LIFESTYLE	Kids	Jacket	10	X	X	X
FW	Sport Heritage	LIFESTYLE	Kids	Jacket	9	X	X	X
FW	Sport Heritage	LIFESTYLE	Kids	Jacket	10	X	X	X
FW	Sport Heritage	LIFESTYLE	Kids	Rainjacket	9	X	X	X
FW	Sport Performance	TRAINING	Kids	Jacket	9	X	X	X
FW	Sport Performance	TRAINING	Kids	Pullover	9	X	X	X
FW	Sport Performance	TRAINING	Kids	Jacket	9	X	X	X
FW	Sport Performance	TRAINING	Kids	Jacket	10	X	X	X
FW	Sport Performance	TRAINING	Kids	Jacket	9	X	X	X

Say Key

Code	Season	Toplam
01	FW	550
10	SS	120
	Genel Toplam	670

Say Key

Code	MarketingDivisic	Toplam
01	Sport Heritage	85
10	Sport Performance	585
	Genel Toplam	670

Say Key

Code	SportCode	Toplam
0001	Adventure	3
0010	BASKETBALL	8
0011	FOOTBALL/SOCCER	41
0100	LIFESTYLE	60
0101	ORIGINALS	25
0110	OUTDOOR	46
0111	RUNNING	4
1000	TENNIS	12
1001	TRAINING	471
	Genel Toplam	670

Say Key

Code	Gender	Toplam
001	Boys	6
010	Girls	22
011	Infant	95
100	Kids	121
101	Men	303
110	Women	123
	Genel Toplam	670

Say Key

Code	SubGroup	Toplam
0001	Anorakjacket	115
0010	Coat	1
0011	Fullzipperjacket	104
0100	Jacket	284
0101	Pullover	51
0110	Rainjacket	43
0111	Tracktop	28
1000	Vest	5
1001	Windjacket	39
	Genel Toplam	670

Say Key

Code	RetailMonth	Toplam
0001	1	16
0010	2	65
0011	3	29
0100	4	10
0101	5	2
0110	7	8
0111	8	94
1000	9	260
1001	10	161
1010	11	25
Genel Toplam		670

Şekil 4.9 : Ceket ürün verisi için veri setinde yapılan sadeleştirme.

4.5.1 Çok katmanlı algılayıcı parametre seçimi

Çizelge 4.13'te görüldüğü üzere en iyi değeri veren parametre değerleri 0.05 öğrenme oranı, 0.3 momentum değeri, 20 adet gizli katman sayısı olan çok katmanlı algılayıcıdır. Ancak korelasyon sayısına bakıldığında 0.80 değerinden daha küçük bir değer olduğu için başarısı tatmin edici bulunmamıştır.

Çizelge 4.13 : Ceket veri seti için çok katmanlı algılayıcı parametreleri.

Öğrenme Oranı	Momentum Oranı	Gizli Katman Sayısı	Gizli Katmandaki Nöron Sayısı	Korelasyon Katsayısı (R^2)	Eğitim MAE (%)	RMSE (%)
0.05	0.3	1	10	0.7373	0.0609	0.0916
0.05	**0.3**	**1**	**20**	**0.7465**	**0.0594**	**0.0903**
0.1	0.3	1	10	0.7465	0.0609	0.0929
0.1	0.3	1	20	0.7557	0.0594	0.0905
0.05	0.3	2	5, 10	0.6982	0.0644	0.0978
0.05	0.3	2	10, 5	0.7004	0.0639	0.0985
0.1	0.3	2	5, 10	0.7091	0.0643	0.0999
0.1	0.3	2	10, 5	0.7177	0.0651	0.0992
0.05	0.7	1	10	0.7515	0.0608	0.0893
0.05	0.7	1	20	0.7516	0.0622	0.0899
0.1	0.7	1	10	0.714	0.067	0.0956
0.1	0.7	1	20	0.7301	0.0655	0.0934
0.05	0.7	2	5, 10	0.7027	0.0696	0.0977
0.05	0.7	2	10, 5	0.7205	0.062	0.0938
0.1	0.7	2	5, 10	0.6864	0.0685	0.0984
0.1	0.7	2	10, 5	0.7125	0.0637	0.0949

4.5.2 Rassal orman algoritması parametre seçimi

Bölüm 4.4.2'de açıklandığı üzere rassal orman algoritmasında kullanıcı kararına bırakılan yalnızca seed (çekirdek) parametresi olmaktadır. Parametre seçimi için yapılan denemeler sonucunda seed sayısı 40 olarak belirlenmiştir. Yapılan 5 adet deneme ve deneme sonuçları Çizelge 4.14'te ki gibidir.

Çizelge 4.14 : Ceket veri seti için rassal orman algoritma parametre seçimi.

Seed (Çekirdek) Sayısı	Korelasyon Katsayısı (R^2)	Eğitim MAE (%)	RMSE (%)
10	0.8947	0.0373	0.0613
20	0.8951	0.0368	0.0611
40	**0.8978**	**0.0366**	**0.0604**
100	0.8967	0.0367	0.0608

4.5.3 İki yöntem sonuçlarının karşılaştırılması ve iyi olanın seçimi

İki yöntemin de eğitim veri seti üzerinde bir kez çalıştırılmasının sonucunda Rassal orman algoritmasının daha iyi sonuç verdiği gözlemlenmiştir. Ancak bunun tesadüfi bir sonuç olup olmadığı sorusunu gidermek amacı ile iki algoritma da Weka tarafından rastgele seçim ile oluşturulan %90 ve %10'luk 10 adet veri setleri ile çalıştırılıp bulunan performans değerlerine *t*-testi yapılarak en iyi algoritma belirlenmeye çalışılmıştır. Çizelge 4.15'te her iki algoritma ile performans değerleri gözetilerek karşılaştırılan toplam 10 denemenin regresyon sonuçları mevcuttur.

Çizelge 4.15 : ÇKAve RF ile yapılan regresyon çalışmalarının sonuçları.

Deneme	Korelasyon Katsayısı (R^2)		MAE		RMSE	
	Çok Katmanlı Algılayıcı	Rassal Orman Algoritması	Çok Katmanlı Algılayıcı	Rassal Orman Algoritması	Çok Katmanlı Algılayıcı	Rassal Orman Algoritması
1	0.6082	0.8255	0.0887	0.0616	0.113	0.0913
2	0.6895	0.8259	0.0884	0.0598	0.1	0.09
3	0.6251	0.8227	0.0896	0.0598	0.1094	0.0903
4	0.6139	0.8365	0.0898	0.0575	0.109	0.0862
5	0.5948	0.827	0.0913	0.0596	0.1127	0.0899

Çizelge 4.16 (devam) : ÇKA ve RF ile yapılan regresyon çalışmalarının sonuçları.

Deneme	Korelasyon Katsayısı (R^2)		MAE		RMSE	
	Çok Katmanlı Algılayıcı	Rassal Orman Algoritması	Çok Katmanlı Algılayıcı	Rassal Orman Algoritması	Çok Katmanlı Algılayıcı	Rassal Orman Algoritması
6	0.5936	0.8328	0.0919	0.0575	0.1215	0.0849
7	0.5849	0.8281	0.0927	0.0592	0.1291	0.0879
8	0.5747	0.8271	0.094	0.0595	0.1303	0.0891
9	0.5929	0.8319	0.0932	0.0591	0.1286	0.088
10	0.6092	0.8284	0.0877	0.0593	0.1247	0.0886

Çizelge 4.17 : Korelasyon katsayısı temel alınarak yapılan *t*-testi sonuçları.

Hipotez	Performans Kriteri	T değeri	P(T<=t)	$T_{0.05}$
$H_{0:}\ \rho_{ÇKA\geq}\ \rho_{RF}$ H_1: $\rho_{ÇKA<}\ \rho_{RF}$	Korelasyon Katsayısı	21.556	0.000	1.83

Çizelge 4.16'da ki H_o hipotezi çok katmanlı algılayıcının korelasyon katsayısının rassal orman algoritmasının korelasyon katsayısından daha büyük olması üzerine, H_1 hipotezi ise çok katmanlı algılayıcı algoritmasının korelasyon katsayısının rassal orman algoritması korelasyon katsayısından daha büyük olacağı üzerine kurulmuştur. Yapılan *t*-testinde anlamlılık düzeyi 0,05 olarak kullanıcı tarafından belirlenmiştir. Her iki algoritma korelasyon katsayısı kriterine göre birbirleri ile karşılaştırılmıştır. Karşılaştırmanın anlamlılık düzeyi 0.000 olmakla birlikte ilk adımda belirlenen anlamlılık düzeyinden çok daha küçüktür ve *t* değeri de tablo değeri ile kıyaslanmış ve 21.556 değeri ile tablo değerinden daha yüksek bir değer elde edildiği için H_o hipotezi reddedilir. Bunun anlamı da yapılan her deneme için rassal orman algoritmasının çok katmanlı algılayıcıya göre daha iyi sonuçlar verdiğinin bir kanıtı niteliğindedir.

Bu bulgudan hareketle yıllar itibari ile ilkbahar yaz ve sonbahar kış sezonları için yapılan tahmin sonuçları Çizelge 4.17'de gösterilmiştir.

2006 ilkbahar yaz sezonu için yapılan çalışmadaki tahmin ile performans değerine bakıldığında birbirine yakın olduğu gözlemlenmiştir. Bu sonuçtan hareketle 2005 ve 2006 yılı verileri kullanılarak 2007 yılı ceket talep miktarı tahmin edilmiştir.

Çizelge 4.18 : Farklı sezonlar için tahmin değerleri ile gerçek değerlerin gösterimi.

Sezon	RMSE	MAE	Gerçek Değer	Tahmin Değeri
2006 SS	0.0401	0.0303	13884	13751
2006 FW	0.067	0.0514	106680	106740
2007 SS	0.0423	0.0396	15478	15504
2007 FW	0.0547	0.0423	132305	132336

5. SONUÇ VE ÖNERİLER

Bu araştırma kapsamında makine öğrenmesi yöntemleri bir sezon içindeki değişen talep tahminini öngörebilmek için ve belirlenmiş bir ürüne ait gelecekteki sezon satış miktarı değerini tahmin edebilmek için kullanılmıştır. Yapılan literatür araştırması sonucunda son yıllarda perakende sektöründe makine öğrenme yöntemlerine sıkça başvurulduğu görülmüştür. Bu çalışmalardan hareketle makine öğrenmesi yöntemleri genellikle karşılaştırmalı olarak değerlendirilmiş ve veri setine en iyi yanıt veren, veri setini en iyi öğrenebilen algoritma seçimi daha çok değerlendirilmiştir.

Bu çalışmada uygulama kapsamında kullanılan makine öğrenmesi yöntemlerinden çok katmanlı algılayıcının ve rassal orman algoritmasının teorik altyapısı oluşturulmuş daha sonra spor ürünleri satan uluslararası bir markadan alınan sezon satış verileri ile amaçlar gerçekleştirilmiştir. Elde edilen veri seti sadeleştirilip dönüştürüldükten sonra programa girişleri yapılmıştır. Uygulama da ağırlıklı olarak Weka programı kullanılmıştır. Veri dönüşüm sadeleştirme işlemleri Excel üzerinde gerçekleştirilmiştir.

Çalışmanın ilk adımı mevcut sezon verisi için yukarıda bahsedilen iki algoritma ile veri setini eğitip performans değerlendirme kriterlerine bakılarak (korelasyon katsayısı, ortalama mutlak hata, kök ortalama kare hatası) daha iyi sonuç veren yöntemle tahmin çalışması yapılmış ve sezon içi değişen talep durumunda hangi ürüne talep geleceğini ve bu talebin hangi özelliklerde ve ne kadar olabileceği tahmin edilmeye çalışılmıştır.

İkinci adım olarak, bir sezonluk veri ile çalışıldığı için bir ürünlerin benzer özellik gösterip göstermediğine bakılmaksızın tüm ürün bilgisini içine alan bir regresyon çalışması yapmanın uygun olup olmayacağı sorusu da kümeleme yöntemi denenerek giderilmiştir. Kümeleme yaklaşımında da iki yöntem (*K*-ortalamalar, *X*-ortalamalar algoritması) birbiri ile rand endeksi değeri baz alınarak kıyaslanmış ve daha iyi sonuç veren yöntem ile tahmin çalışması yapılmıştır. Burada ilk çalışmadan farklı olarak tüm kümeler ayrı ayrı veri setleri olarak değerlendirilmiş ve programa girişleri ayrı yapılmıştır. Daha iyi bir endeks değerine sahip olduğu için *K*-ortalamalar kümeleme

yaklaşımı benimsenmiş ve bu yaklaşımın oluşturduğu kümelere ayrı ayrı regresyon çalışmaları yapılmıştır. Çıkan sonuçlara bakıldığında hem korelasyon katsayısı olarak hem de ortalama mutlak hata ve kök ortalama kare hatası olarak tatmin edici sonuçlar elde edilmiştir. Ancak kümeleme yaklaşımının verdiği performans değerleri kümeleme olmadan yapılan çalışmadan elde edilen performans değerlerinden daha düşük olduğundan dolayı kümelemenin tahmini iyileştirmediği gözlemlenmiştir.

Çalışmanın devamında sadece bir sezonluk veri değil, önceden belirlenmiş bir ürün için geçmiş sezonlardaki bu ürüne ait satış miktarları dikkate alınarak ürün için gelecek sezondaki talep ya da sipariş miktarı tahmin edilmeye çalışılmıştır. Çalışmanın bu kısmında tüm sezonlardaki ortak kriterler belirlenmiş ve bundan dolayı veri setinde belirli sadeleştirmeler yapılmıştır. Sadeleştirilen ve dönüştürülen veri seti Weka programına girilmiş ve hareket noktası olan iki algoritma set üzerinde eğitim ve test olarak ölçülmüştür. Sonuçlar rassal orman algoritması tarafında tatmin edici bulunmuş ve devamında bu yöntemle tahmin çalışmaları yapılmıştır. Tahmin sonuçlarına bakıldığında ise sonuçlar memnun edicidir.

Değişkenliğin getirdiği riskin çok yüksek olduğu perakende sektöründeki en önemli problemlerden biri olan talep tahmini konusunda makine öğrenme yöntemlerinin ne kadar etkin bir biçimde sonuç verdiği bu çalışma kapsamında gözlemlenmiştir. Ancak farklı algoritmalar aynı veri seti üzerinde benzer etkiyi yaratmamaktadır ve bundan dolayı da veri seti için algoritma seçimi büyük önem taşımaktadır. Algoritmalar veri setinin içerdiği değişken tiplerine göre de farklılıklar göstermektedir. Bazı yöntemler kategorik değişkenler ile nümerik değişkenleri bir arada işleyebilirken bazıları ise sadece kategorik ya da sadece nümerik değerler i işleyebilmektedir. Bu sebeple dikkat edilecek en büyük noktalardan biri veri setine göre algoritmayı ve algoritma parametrelerini iyi belirleyebilmektir.

Yapılan çalışma sonucunda makine öğrenme yöntemlerinin talep tahmini konusundaki başarısı kanıtlanmış ve yönetim için verimli ve uygulanabilir bir yöntem olduğu gözlemlenmiştir. Özellikle perakende sektörü ele alındığında ürün sirkülasyonunun çok yüksek olması, mevsimsel etkiler vb. sebeplerle talepler çok hızlı değişmekte olduğundan bir sezonluk katalog verisine bakılarak sezon içindeki dalgalanan talebi yönetme problemi için uygulanabilir bir çalışma olmaktadır.

Bir ürün ele alınarak yapılan sezonlar arası regresyon analizi sonuçları da tatmin edici düzeyde bulunduğundan algoritmanın iyi eğitildiğini bu nedenle de tahmini tatmin edici düzeyde verdiği görülmektedir. İleriki aşamalarda uzmanlar tarafından belirlenebilecek başka bir ürün için de bu uygulama kullanılabilir düzeyde olmakta olup gerektiği durumlarda algoritmaların parametrelerinde değişiklik yapılarak iyi sonuçlar elde edilmeye çalışılabilir.

KAYNAKLAR

Aksoy Asli, Öztürk Nursel, & Sucky Eric. (2014). Demand forecasting for apparel manufacturers by using neuro-fuzzy techniques. *Journal of Modelling in Management*, 9(1), 18–35.

Bala, P. K. (2010). Decision tree based demand forecasts for improving inventory performance. 2010 *IEEE International Conference on Industrial Engineering and Engineering Management*, *Industrial Engineering and Engineering Management (IEEM), 2010 IEEE International Conference On*, 1926–1930.

Biau, G. (2012). Analysis of a Random Forests Model. *Journal of Machine Learning Research*, 13(4), 1063–1095.

Brahmadeep, & Thomassey, S. (2016). 8 - Intelligent demand forecasting systems for fast fashion. *Information Systems for the Fashion and Apparel Industry*, 145–161.

Breiman, L. (2001). Random Forests. *Machine Learning*. 45(1):5-32.

Browne, M., & Berry, M. W. (2006). Lecture Notes In Data Mining. *World Scientific*.

Chaudhuri, T. D., Ghosh, I., & Singh, P. (2017). Application of Machine Learning Tools in Predictive Modeling of Pairs Trade in Indian Stock Market. *IUP Journal of Applied Finance*, 23(1), 5–25.

Chen, I.-F., & Lu, C.-J. (2017). Sales forecasting by combining clustering and machine-learning techniques for computer retailing. *Neural Computing & Applications*, 28(9), 2633–2647.

De Amorim, R. C. (2012). Constrained clustering with Minkowski Weighted K-Means. *2012 IEEE 13th International Symposium on Computational Intelligence and Informatics (CINTI), Computational Intelligence and Informatics (CINTI), 2012 IEEE 13th International Symposium On*, 13–17.

Demiriz, A. (2014). Demand Forecasting based on Pairwise Item Associations. *Procedia Computer Science*, 36, 261–268.

Dietrich, D., EMC Education Services, Heller, R., & Yang, B. (2015). Data Science and Big Data Analytics : Discovering, Analyzing, *Visualizing and Presenting Data*. Wiley.

Fratello Michele, Tagliaferri Roberto (2019). Decision Trees and Random Forests. *Encyclopedia of Bioinformatics and Computational Biology* Vol 1, 374-383

Giri, C., Thomassey, S., Balkow, J., & Zeng, X. (2019). Forecasting New Apparel Sales Using Deep Learning and Nonlinear Neural Network Regression. 2019 *International Conference on Engineering, Science, and Industrial*

Applications (ICESI), Engineering, Science, and Industrial Applications (ICESI), 2019 *International Conference* On, 1–6.

Graupe, D. (2013). Principles Of Artificial Neural Networks (3rd Edition): Vol. 3rd edition. *World Scientific.*

Halima Elaidi, A., Zahra Benabbou, A., & Hassan Abbar, A. (2018). A comparative study of algorithms constructing decision trees: ID3 and C4.5. Learning and Optimization Algorithms: Theory and Applications

Haykin, S. (2009). Neural Networks and Learning Machines. https://cours.etsmtl.ca/sys843/REFS/Books/ebook_Haykin09.pdf sitesinden alındı.

He Huang, & Qiurui Liu. (2017). Intelligent Retail Forecasting System for New Clothing Products Considering Stock-out. *Fibres & Textiles in Eastern Europe*, 25(1), 10.

Huang, J.-C., & Pan, W.-T. (2010). Forecasting classification of operating performance of enterprises by probabilistic neural network. *Journal of Information & Optimization Sciences*, 31(2), 333.

Huber, J., Gossmann, A., & Stuckenschmidt, H. (2017). Cluster-based hierarchical demand forecasting for perishable goods. *Expert Systems With Applications*, 76, 140–151.

Hui, P. C. L., & Choi, T.-M. (2016). 5 - Using artificial neural networks to improve decision making in apparel supply chain systems. *Information Systems for the Fashion and Apparel Industry*, 97–107.

Junwei Xiao, Jianfeng Lu, & Xiangyu Li. (2017). Davies Bouldin Index based hierarchical initialization K-means. *Intelligent Data Analysis*, 21(6), 1327–1338.

Kang, Z., Zuo, Y., Huang, Z., Zhou, F., & Chen, P. (2017). Research on the Forecast of Shared Bicycle Rental Demand Based on Spark Machine Learning Framework. *2017 16th International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES), Distributed Computing and Applications to Business, Engineering and Science (DCABES), 2017 16th International Symposium on, DCABES*, 219–222.

Kumar, M., & Patel, N. R. (2010). Using clustering to improve sales forecasts in retail merchandising. *Annals of Operations Research*, 174(1), 33–46.

Kusrini, K. (2015). Grouping of Retail Items by Using K-Means Clustering. *Procedia Computer Science*, 72, 495–502.

Kwon, S. J. (2011). Artificial Neural Networks. *Nova Science Publishers*, Inc.

Liao, H., & Sun, W. (2010). Forecasting and Evaluating Water Quality of Chao Lake based on an Improved Decision Tree Method. *Procedia Environmental Sciences*, 2, 970–979.

Loureiro, A. L. D., Miguéis, V. L., & da Silva, L. F. M. (2018). Exploring the use of deep neural networks for sales forecasting in fashion retail. *Decision Support Systems*, 114, 81–93.

Ngai, E. W. T., Peng, S., Alexander, P., & Moon, K. K. L. (2014). Decision support and intelligent systems in the textile and apparel supply chain: An academic review of research articles. *Expert Systems With Applications*, 1, 81.

Nuaimi, E. A., Marzooqi, S. A., & Zaki, N. (2016). Predicting the decision for the provision of municipal services using data mining approaches. *2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC), Big Data and Smart City (ICBDSC), 2016 3rd MEC International Conference On*, 1–6.

Semenov, V. P., Chernokulsky, V. V., & Razmochaeva, N. V. (2017). Research of artificial intelligence in the retail management problems. *2017 IEEE II International Conference on Control in Technical Systems (CTS), Control in Technical Systems (CTS), 2017 IEEE II International Conference On*, 333–336.

Shahbaba, M., & Beheshti, S. (2012). Improving X-means clustering with MNDL. *2012 11th International Conference on Information Science, Signal Processing and Their Applications (ISSPA), Information Science, Signal Processing and Their Applications (ISSPA), 2012 11th International Conference On*, 1298–1302.

Sharma, S. K., Chakraborti, S., & Jha, T. (2019). Analysis of book sales prediction at Amazon marketplace in India: a machine learning approach. *Information Systems and E-Business Management*, 17(2–4), 261.

Singh, D., & Singh, B. (2019). Investigating the impact of data normalization on classification performance. *Applied Soft Computing Journal*, 105524.

Slimani, I., Farissi, I. E., & Al-Qualsadi, S. A. (2016). Configuration of daily demand predicting system based on neural networks. *2016 3rd International Conference on Logistics Operations Management (GOL), Logistics Operations Management (GOL), 2016 3rd International Conference On*, 1–5.

Steinley, D., & Brusco, M. J. (2018). A note on the expected value of the Rand index. *British Journal of Mathematical & Statistical Psychology*, 71(2), 287–299. https://doi.org/10.1111/bmsp.12116

Taraji, M., Haddad, P. R., Amos, R. I. J., Talebi, M., Szucs, R., Dolan, J. W., & Pohl, C. A. (2017). Error measures in quantitative structure-retention relationships studies. *Journal of Chromatography A*, 298.

Thomassey, S. (2010). Sales forecasts in clothing industry: The key success factor of the supply chain management. *International Journal of Production Economics*, 128(2), 470–483.

Vo. T. H, P. (2017). *Python: Data Analytics and Visualization*. Packt Publishing**.**

Witten, I. H., & Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques: Vol. 3rd ed. Ian H. Witten, Frank Eibe, Mark A. Hall. Morgan Kaufmann.

Xixin Wu, Zhiyong Wu, Jia Jia, Meng, H., Lianhong Cai, & Weifeng Li. (2014). Automatic speech data clustering with human perception based weighted distance. *The 9th International Symposium on Chinese*

Spoken Language Processing, Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium On, 216–220.

Xu, R., & Wunsch, D. C. (2009). *Clustering*. Wiley-IEEE Press.

You Lingxian, A., Kou Jiaqing, A., & Wang Shihuai, A. (2019). Online Retail Sales Prediction with Integrated Framework of K-mean and Neural Network. *E-Business, Management and Economics*, 115.

Yu, Y., Choi, T.-M., & Hui, C.-L. (2011). An intelligent fast sales forecasting model for fashion products. *Expert Systems With Applications*, 38(6), 7373–7379.

Zhu Ying, & Xiao Hanbin. (2010). Study on the Model of Demand Forecasting Based on Artificial Neural Network. *2010 Ninth International Symposium on Distributed Computing and Applications to Business, Engineering and Science, Distributed Computing and Applications to Business Engineering and Science (DCABES), 2010 Ninth International Symposium On*, 382–386.

Url-1 Multiple Correlation. (2020, Ocak 14). In *Wikipedia*. https://en.wikipedia.org/wiki/Multiple_correlation

ÖZGEÇMİŞ

Ad-Soyad **: Sinem ÖZTÜRK**

Doğum Tarihi ve Yeri **: 06.08.1993 - Bursa**

E-posta **:** sinzturk@gmail.com

ÖĞRENİM DURUMU:

- **Lisans** **:** 2017, Yıldız Teknik Üniversitesi, Makine Fakültesi, Endüstri Mühendisliği

MESLEKİ DENEYİM VE ÖDÜLLER:

- 2017-2019 Proje Mühendisi - ARÇELİK