İSTANBUL TEKNİK ÜNİVERSİTESİ ★ SOSYAL BİLİMLER ENSTİTÜSÜ

MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE NÜFUS TAHMİNİ: TÜRKİYE ÖRNEĞİ

YÜKSEK LİSANS TEZİ

Fatih Veli ŞAHİNARSLAN

İşletme Anabilim Dalı

İşletme Yüksek Lisans Programı

HAZİRAN 2019

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ SOSYAL BİLİMLER ENSTİTÜSÜ

MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE NÜFUS TAHMİNİ: TÜRKİYE ÖRNEĞİ

YÜKSEK LİSANS TEZİ

Fatih Veli ŞAHİNARSLAN
403161037

İşletme Anabilim Dalı

İşletme Yüksek Lisans Programı

Tez Danışmanı: Prof. Dr. Ferhan ÇEBİ

HAZİRAN 2019

İTÜ, Sosyal Bilimler Enstitüsü'nün 403161037 numaralı Yüksek Lisans Öğrencisi Fatih Veli ŞAHİNARSLAN, ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı "MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE NÜFUS TAHMİNİ: TÜRKİYE ÖRNEĞİ" başlıklı tezini aşağıda imzaları olan jüri önünde başarı ile sunmuştur.

Tez Danışmanı :	**Prof. Dr. Ferhan Çebi** İstanbul Teknik Üniversitesi	.....
Jüri Üyeleri :	**Prof. Dr. Şakir ESNAF** İstanbul Üniversitesi	.....
	Doç. Dr. Dilay ÇELEBİ İstanbul Teknik Üniversitesi	.....

Teslim Tarihi : 02 Mayıs 2019
Savunma Tarihi : 10 Haziran 2019

Dostlarıma,

ÖNSÖZ

Bu tez çalışmasında farklı makine öğrenmesi algoritmaları ile ülkelerin nüfus tahminini kolaylaştıracak bir model geliştirmek istenmiştir.

Öncelikle bu çalışmanın gerçekleşmesinde güler yüzü ve samimiyeti ile desteğini hiçbir zaman esirgemeyen ve gelecek meslek hayatımda bana yol gösterecek tecrübeler edinmemi sağlayan saygıdeğer danışman hocam; Prof. Dr. Ferhan ÇEBİ'ye, analizler süresince desteğini esirgemeyen Sayın Ahmet Tezcan Tekin'e ve Mehmet Emin Öztürk kardeşime sonsuz teşekkürlerimi sunarım. Baki selamlar.

Haziran 2019 Fatih Veli ŞAHİNARSLAN

İÇİNDEKİLER

Sayfa

KISALTMALAR

ARIMA : Autoregressive Integrated Moving Average
TDO : Toplam Doğurganlık Oranı
YÖDO : Yaşa Özel Doğurganlık Oranı
TÜİK : Türkiye İstatistik Kurumu
XGBoost : Ekstrem Gradyan Artırma
LightGBM : Light Gradyan Artırma
CCM : Kuşak Bileşenleri Yöntemi (Cohort Component Model)
PCA : Temel Bileşen Analizi (Principal Component Analysis)

ÇİZELGE LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE NÜFUS TAHMİNİ: TÜRKİYE ÖRNEĞİ

ÖZET

Ülkelerin geleceğine dair sosyal, ekonomik, politik ve çevresel kararların daha tutarlı alınabilmesi için ülke nüfusu hakkında bilgi sahibi olunmalıdır. Bu sebeple, büyük maliyetler ile nüfus sayımları yapılmakta ve farklı teknikler ile ülkenin gelecekteki nüfusu tahmin edilmektedir. Bu çalışmada farklı makine öğrenmesi algoritmaları ile nüfus tahmini yapılmıştır. Bunun için altı farklı makine öğrenmesi algoritması seçilmiştir; Light Gradyan Artırma (Light Gradient Boosting, LightGBM), Doğrusal Regresyon, Ridge Regresyon, Üstel Düzeltme yöntemlerinden biri olan Holt-Winters, Bütünleşik Otoregresif Hareketli Ortalama (Autoregressive Integrated Moving Average, ARIMA) ve Prophet tahmin modeli. 262 farklı ülkenin 1960-2017 yılları arasındaki 1595 farklı demografik göstergesi kullanılarak modeller eğitilmiştir. Kullanılan veri seti dünya genelinde demografik göstergeler yayınlayan Dünya Bankası'ndan (World Bank) temin edilmiştir. Makine öğrenmesi algoritmaları Python programlama dili ile Jupyter program arayüzü kullanılarak eğitilmiştir.

2017 yılı toplam nüfusu tahmin edilerek sonuçlar gerçek değerler ile karşılaştırılmıştır. Algoritmaların performansları karşılaştırıldığında, bütün modeller arasından ARIMA modeli en başarılı sonucu vermiştir. Doğrusal regresyon, Ridge regresyon ve Holt-Winters modelleri performans açısından ARIMA modeline yakın sonuçlar vermiştir. Sadece nüfus istatistikleri kullanılarak modeller eğitildiğinde Doğrusal regresyon ve Ridge regresyon modellerinin daha başarılı sonuçlar verdiği gözlemlenmiştir.

Ayrıca, eğitilen modeller ile 2017 yılı toplam Türkiye nüfusu tahmin edilmiştir. Kuşak bileşenleri yöntemi (Cohort Component Method) nüfus tahmini için ülkelerin istatistik kurumları tarafından sıkça kullanılan yöntemlerden biridir. Bu yöntem ile nüfus tahmini yapmak için ülkenin nüfus istatistikleri, doğum, ölüm ve göç oranları, doğumda beklenen yaşam süresi ve doğumdaki cinsiyet oranı verileri kullanılır. Makine öğrenmesi algoritmaları ile tahmin edilen 2017 yılı toplam Türkiye nüfusu, kuşak bileşenleri yöntemi ile de tahmin edilerek sonuçlar karşılaştırılmıştır. Ridge regresyon modeli Türkiye'nin nüfusunu altı bin kişi fark ile tahmin ederken, Kuşak bileşenleri yöntemi dört yüz bin kişi fark ile tahmin etmiştir. Elde edilen sonuçlara göre, makine öğrenmesi algoritmalarının nüfus tahmin başarısı kuşak bileşenleri yöntemine göre oldukça yüksektir. Makine öğrenmesi algoritmalarının nüfus tahmini üzerinde kullanılması ülkeler için önemli bir katkı sağlayacak ve karar alma sürecini iyileştirecektir.

MACHINE LEARNING ALGORITHMS TO FORECAST POPULATION: TURKEY EXAMPLE

SUMMARY

The population of the countries should be known to be more consistent with the social, economic, political and environmental decisions regarding the future of them. For this reason, censuses are made with great cost and the future population of the country is estimated with different techniques. In this study, different machine learning algorithms are used to forecast population. This study employed six different machine learning algorithms; Light Gradient Boosting, Linear Regression, Ridge Regression, Holt-Winters, Exponential Autoregressive Integrated Moving Average (ARIMA) and Prophet Prediction Model. Models were trained using 1595 different demographic indicators of 262 different countries between 1960 and 2017. The data set is obtained from the World Bank, which publishes demographic indicators worldwide. Machine learning algorithms are trained using the Jupyter program interface with Python programming language.

The total population of 2017 was estimated and the results were compared with the actual values. When the performance of algorithms was compared, the ARIMA model was the most successful among all models. Linear regression, Ridge regression, and Holt-Winters models were close to the ARIMA model in terms of performance. Linear regression and Ridge regression models showed more successful results when the only total populations data set is trained.

In addition, Turkey's total population of 2017 were forecasted with trained models. Cohort Component Method is one of the methods commonly used by countries' statistical institutions for population estimation. With this method, population statistics, birth, death and migration rates, life expectancy at birth and sex ratio at birth are used to estimate the population. The total population of Turkey in 2017 estimated by pre-trained machine learning algorithms were compared with the result predicted by Cohort Component Method. Ridge regression model estimated Turkey's population with six thousand people difference while Cohort Component Method predicted with 400 thousand people difference. Results showed that machine learning algorithms performed better than the demographic model. Therefore, using machine learning algorithms on population estimation will make a significant contribution to countries and enhance the decision-making process.

1. GİRİŞ

Teknolojinin gelişmesiyle beraber insanlar ve şirketler tarafından üretilen bilgiler büyük bir veri yığını (big data) oluşturmuştur. Bu veri yığını gün geçtikçe hızla büyümektedir. Örneğin, şu anda yaklaşık 1 trilyon internet sayfası bulunmaktadır. Youtube adlı video sağlayıcısına her bir saniyede bir saatlik video yüklenmektedir ve günde yaklaşık 10 yıllık içerik yüklenir. Walmart şirketi tarafından saat başı 1 milyon alım satım işlemi gerçekleşmektedir ve şirketin veri tabanında 2.5 petabayttan (1 petabayt = 1024 terabayt) fazla veri bulunmaktadır (Cukier, 2010). 2017 yılında internet ortamında yalnızca bir hafta içinde üretilen veri hacmi, 20. yüzyıl boyunca üretilen veri hacmine eşittir (Tekin ve diğ, 2018). Bu örnekler çoğaltılabilir ancak asıl mesele tüm bu veri yığınının içerisinden verimli algoritmalar ile (machine learning) anlamlı bilgiler çıkarmak ve kullanabilmektir (data mining).

İşletmeler yaptıkları tahminleri göz önüne alarak değerli kaynaklarını nasıl tahsis edeceklerine karar verirler. Kaynakların planlanması, risk faktörlerinin değerlendirilmesi, ihaleler ve benzeri konularda kararların alınması için büyük veriler üzerinden bir takım tahminler yapılması gerekir. Bu tahminlerin yapılmasında geleneksel matematik modelleri yetersiz kaldığından makine öğrenmesi algoritmaları imdada yetişmiştir. Makine öğrenmesi algoritmaları ile veri kısa sürede analiz edilerek tutarlı ve doğru tahminler yapılmasının önü açılmıştır (Mair ve diğ, 2000). Yapılan bir çalışmada, karar alma sürecinde büyük veri analizi uygulayan firmaların, basit bilgi sistemlerini kullanan firmalara göre daha tutarlı kararlar aldığı gözlemlenmiştir (Tekin ve diğ, 2019)

Yıllar boyunca çeşitli makine öğrenmesi algoritmaları bilinen veri üzerinden eğitilerek bilinmeyen verilerin tahmin edilmesinde kullanılmıştır. Wolpert ve Macready tarafından 1997 yılında ortaya atılan “No Free Launch” teoremine göre bütün problemleri tam verimlilik ile çözebilecek optimum bir algoritma bulunmamaktadır (Wolpert ve Macready, 1997). Herhangi bir algoritma bir problem için doğru ve kullanışlı sonuçlar verebilirken, farklı problemler için başarısız olabilir. Bu konuda

George Box makine öğrenmesi algoritmaları için şu ifadeyi kullanmıştır: “Bütün modeller yanlıştır, ancak bazı modeller kullanışlı olabilir.” (Box ve Draper, 1987).

Makine öğrenmesi algoritmalarının işlevselliği minimum veri seti ile yaptıkları tahminler üzerinden belli olur. Gerçek hayatta istenilen ölçüde veri bulunamayabilir. Araştırmacıların gözlemlediği veya tecrübe ettiği veriler kısıtlıdır. Yine de bu veri seti ile tahmin yürütmek isterler. Asıl mesele az veri seti ile gerçeğe yakın tahminler yapabilmektir (Bélisle ve diğ, 2015)

Bu zamana kadar yapılan çalışmalarda çok çeşitli tahmin teknikleri kullanılarak öngörüleme (forecasting) yapılmıştır. Bu teknikler mühendislik metodları, istatistiksel metodlar ve yapay zeka metodları olarak üç ana başlıkta toplanabilir. Bunlardan yapay zeka metodları makine öğrenmesi algoritmaları ile veri analizi üzerinde sıkça kullanılmaktadır. Bu metodlara örnek olarak Yapay Sinir Ağları (Artificial Neural Network), Sınıflandırma ve Regresyon Ağaçları (Classification and Regression Trees, CART), Destek Vektör Makineleri (Support Vector Machine), Lineer Regresyon, Üstel Düzeltme (Exponential Smoothing), Bütünleşik Otoregresif Hareketli Ortalama (Autoregressive integrated moving average, ARIMA) verilebilir (Chou ve Tran, 2018).

Bu çalışmada farklı makine öğrenmesi algoritmalarının nüfus projeksiyonu üzerindeki tahmin yetenekleri değerlendirilecektir. Nüfus ve sağlık istatistikleri veri seti olarak kullanılacaktır. Tahmin edilen nüfus verileri kuşak bileşenleri yöntemi ile elde edilen nüfus projeksiyonları ile karşılaştırılacaktır.

Ülkelerin nüfus dinamikleri ile gelişim düzeyi arasında kuvvetli bir ilişki vardır. Ülkeler eğitim, sağlık, yatırım, barınma gibi ulusal ihtiyaçları planlamak için nüfus istatistiklerine gereksinim duyarlar. Nüfus istatistikleri ile ülke hakkında yaş-cinsiyet dağılımı, işgücü, eğitim, doğum yeri gibi bilgilere; nüfusun yapısı hakkında ise doğum, ölüm ve göç oranları gibi verilere ulaşılabilir. Ülkeler, vatandaşlarının gelecekteki ihtiyaçlarını planlamak için nüfus projeksiyonu üretirler. Böylece ülkenin geleceğini etkileyecek olan sosyal, ekonomik, çevresel ve politik kararlar verilere dayandırılarak daha sağlıklı alınabilir. Bu sebeple üretilen nüfus projeksiyonlarının tutarlı olması çok önemlidir. Makine öğrenmesi algoritmalarının kullanımı, tahmin edilen nüfusun gerçeğe yakınlığını ve tutarlılığını artırmak için önemli bir katkı sağlayabilir.

Nüfus projeksiyonu üretmek için birçok farklı yöntem bulunmaktadır. Bunların arasından en çok kullanılan yöntem kuşak bileşenleri yöntemidir (Cohort Component Model). Kuşak bileşenleri yöntemi ile nüfus projeksiyonu üretmek için bir takım nüfus verilerine ihtiyaç duyulur. Bu veriler, her yaşa ve cinsiyete göre nüfus sayımları, toplam doğurganlık oranları (TDO), yaşa özel doğurganlık oranları (YÖDO), ölüm oranları, doğumda beklenen yaşam süresi, doğumdaki cinsiyet oranları ve göç oranlarıdır. Kuşak bileşenleri yöntemi ile gelecekteki herhangi bir yılın nüfusunu tahmin etmek için yalnızca bu veriler girdi olarak kullanılır. Ancak, toplam nüfusu etkileyen farklı değişkenler de bulunabilir. Ülkenin demografik göstergeleri nüfus üzerinde farklı etkilere sahiptir. Bu çalışmanın amacı makine öğrenmesi algoritmalarını kullanarak bir sonraki yılın nüfusunu tahmin edebilecek alternatif bir model geliştirmektir. Böylece, makine öğrenmesi algoritması kullanım alanlarının genişletilmesi ve üretilecek olan nüfus projeksiyonunun gerçeğe daha yakın olması hedeflenmektedir. Modelin performansı değerlendirildikten sonra Türkiye'nin verileri üzerinden 2017 yılı toplam Türkiye nüfusu tahmin edilecektir. 6 farklı makine öğrenmesi algoritması ile tahmin edilen 2017 yılı Türkiye nüfusu, farklı bir nüfus tahmin modeli olan ve dünya genelinde ülkelerin nüfus birimleri tarafından sıkça kullanılan kuşak bileşenleri yöntemi ile de tahmin edilerek sonuçlar karşılaştırılacaktır.

Çalışmanın geri kalan kısımlarında; literatürdeki çalışmalar, makine öğrenmesi ve nüfus projeksiyonu hakkında literatür özeti, metodoloji, veri toplama, verinin hazırlanması ve modelin kurulması aşamaları, uygulama, performansların karşılaştırılması, sonuç ve öneriler hakkında bölümler bulunmaktadır.

1.1 Literatür Taraması

Literatürde makine öğrenmesi algoritmaları ile tahmin uygulaması yapılan birçok farklı çalışma bulunmaktadır:

Bir çalışmada farklı makine öğrenmesi algoritmalarını kullanarak materyallerin fiziksel özellikleri üzerine tahminde bulunulmuştur. Sinir ağları, polinom interpolasyonu, Gauss (Gaussian process) ve karar ağacı yöntemlerini kullanarak mol hacmi, elektrik iletkenliği ve Martensite sıcaklığı ile ilgili veri setleri üzerinde tahminler yapılmış ve sonuçlar karşılaştırılmıştır. Tahmin sonuçlarına göre Gauss prosesi yavaş fakat daha tutarlı sonuç vermektedir (Bélisle ve diğ, 2015).

Biyolojik alanda farklı makine öğrenmesi algoritmalarını kullanarak hepatit B virüsünün tahmini, DLBCL adı verilen kötü huylu tümörün büyüme tahmini, kanserin önceden tahmin edilmesi, balıklarda ve omurgasızlardaki biyo yoğunlaşma faktörünün tahmin edilmesi gibi farklı alanlarda çalışmalar bulunmaktadır (Shipp ve diğ, 2002; Ye ve diğ, 2003; Cruz ve Wishart, 2006; Kourou ve diğ, 2015; Miller ve diğ, 2019).

Diğer bir çalışmada ise internet üzerinde reklam veren otellerin tıklanma oranları tahmin edilmiştir. Bu çalışmada Extrem Gradyan Artırma (XGBoosting) algoritması kullanılarak internet üzerinde reklam veren seyahat acentalarının bir sonraki gün alması beklenen tıklanma oranı (Click-Through-Rate, CTR) tahmin edilmiştir (Çakmak ve diğ, 2019).

Finans alanında da birçok tahmin çalışmaları bulunmaktadır. Finansal krizin önceden tahmin edilmesine yönelik yapılan bir çalışmada destek vektör makineleri, yapay sinir ağları ve k-en yakın komşu algoritması kullanılmıştır (Lin ve diğ, 2012). Farklı bir çalışmada ise destek vektör makineleri ile iflas durumunun tahmini yapılmıştır (Shin ve diğ, 2005; Yu ve diğ, 2014). Yine, destek vektör makineleri kullanılarak finansal zaman serileri tahmini yapılmıştır (Trafalis ve Ince, 2000; Kim, 2003). Borsanın trend hareketi ve hisse senedi fiyatlarının tahmini üzerine yapılan çalışmalar da bulunmaktadır (Huang ve diğ, 2005; Pai ve Lin, 2005; Kara ve diğ, 2011).

Eğitim alanındaki çalışmalardan birinde uzaktan eğitim veren üniversiteler için öğrenci performanslarının değerlendirilmesi üzerine model kurulmuştur. Bu modelde en başarılı algoritma Naive Bayes algoritması bulunmuştur. Bu model ile başarısız öğrencilerin önceden tespit edilmesi hedeflenmiştir (Kotsiantis ve diğ, 2004). Uzaktan eğitim alan öğrencilerin başarısını ölçmek için karar ağaçları algoritmasını kullanılan farklı bir çalışma yapılmıştır (Kalles ve Pierrakeas, 2006). Makine öğrenmesi algoritmaları ile akademik performansı önceden tahmin etmeye yönelik başka çalışmalar da bulunmaktadır (Minaei-Bidgoli ve diğ, 2003; Vandamme ve diğ, 2007).

Enerji alanında binaların enerji tüketimini tahmin etmeye yönelik farklı çalışmalar bulunmaktadır. Gerçek zamanlı enerji kullanımının doğrusal olmaması konutların enerji tüketimini tahmin etmeyi oldukça zorlaştırmaktadır. Bu zorluğun önüne geçmek ve enerji tüketimini öngörmek için çeşitli makine öğrenmesi algoritmaları kullanılmıştır. Bu çalışmalarda kullanılan makine öğrenmesi algoritmaları şöyle sıralanabilir; Yapay Sinir Ağları (Kandananond, 2011; Khandelwal ve diğ, 2015),

Destek Vektör Makineleri (Ahmad ve diğ, 2014), Lineer Regresyon (Yip ve diğ, 2017), Bütünleşik Otoregresif Hareketli Ortalama (ARIMA) (Khashei ve Bijari, 2011; Ahmad ve diğ, 2014; Khandelwal ve diğ, 2015), Mevsimsel Bütünleşik Otoregresif Hareketli Ortalama (SARIMA)(Chen ve Wang, 2007), Regresyon ve Sınıflandırma Ağaçları (Hu ve diğ, 2018).

Ayrıca, nüfus projeksiyonu üretmek için literatürde birçok farklı metodoloji kullanılmıştır. Bu metodolojilerin birbirlerinden farkı gelecekteki nüfusu tahmin etmek için kullandıkları değişkenler ve varsayımlardır. Matematiksel projeksiyonlar yöntemi genellikle geçmişteki değerlerin trendine göre ekstrapolasyon yaparak tahminlerde bulunmaktadır. Geçmiş ile gelecek arasındaki nüfus oranlarında yüksek korelasyon olduğu varsayımı yapılır. Ayrıca, matematiksel projeksiyonlarda nüfusun doğum, ölüm ve göç oranı gibi temel değişkenleri hesaba katılmaz. Bu tür varsayımlar, üretilen nüfus projeksiyonunun tutarlılığını etkileyeceği için mümkün mertebe matematiksel projeksiyonlardan kaçınılmalıdır (Boon ve Kok, 1995).

Kuşak bileşenleri yöntemi, nüfus tahmini için doğum, ölüm, göç gibi bir çok faktörü ayrı ayrı ele alır. Bu yöntem nüfus projeksiyonu üretmek için birçok araştırmacı tarafından kullanılmış ve bir ikon haline gelmiştir (Cannan, 1895; Bowley, 1924; Whelpton, 1928; Pritchett, 1891; Leslie, 1945; Dorn, 1950; Shorter ve diğ, 1995).

HIV virüsünün nüfus üzerindeki etkisini inceleyen ve Heuveline tarafından yapılan bir çalışmada kuşak bileşenleri yöntemi farklılaştırılmıştır. HIV virüsünden etkilenme oranı yüksek insanlar göz önüne alınarak nüfus sınıflara ayrılmış ve virüsün zaman içerisinde nüfusa olan etkisi incelenmiştir (Heuveline, 2003). Heuveline tarafından yapılan çalışma güncel veriler kullanarak Thomas ve Clark tarafından geliştirilmiştir (Thomas ve Clark, 2011).

Bu alanda yapılan çalışmaların çoğu kuşak bileşenleri yönteminin farklı uygulama alanlarında kullanılması ile ilgilidir. Bununla beraber, nüfus tahmini için kuşak bileşenleri yöntemine alternatif olarak geliştirilen bir takım çalışmalar bulunmaktadır. 2013 yılında yapılan bir çalışmada kuşak bileşenleri yönteminin Leslie matrisini oluşturma aşaması kolaylaştırılmıştır. “Wood Method” adı verilen yöntem ile yaş, cinsiyet ve diğer değişkenlerin bulunduğu matrisi oluşturmak için ihtiyaç duyulan veri büyüklüğü azaltılmıştır. Geliştirilen yöntem ile 3120 farklı veri seti üzerinde çalışma yapıldıktan sonra sonuçlar değerlendirilmiştir (Sprague, 2013).

Diğer bir çalışma ise gelişmekte olan ülkelerin veri yetersizliğini göz önüne alarak toplam doğum oranını tahmin etmeye yöneliktir. Bu çalışmada anketler, mülakatlar ve farklı kaynaklardan toplanan veriler birleştirilerek Batı Afrika'nın toplam doğum oranı lokal regresyon (local regression) ile tahmin edilmiştir (Alkema ve diğ, 2012). İstatistiksel modelleri kullanarak toplam doğum oranını ve beklenen yaşam süresini tahmin etmeye yönelik farklı bir çalışma daha bulunmaktadır. Bu çalışmada Bayes hiyerarşi modeli (Bayes hierarchical model) ve Markov Monte Carlo algoritması kullanılarak bir model geliştirilmiş ve bütün ülkeler için toplam doğum oranları ve beklenen yaşam süresi tahmin edilmiştir (Alkema ve diğ, 2011; Raftery ve diğ, 2013).

Farklı bir çalışmada ise denetimli yapay sinir ağları yöntemini kullanarak Nijerya'nın ve Hindistan'ın nüfusunu tahmin etmiştir (Folorunso ve diğ, 2010; Bandyopadhyay ve Chattopadhyay, 2006). Diğer bir çalışmada ise yapay sinir ağları yöntemi kullanılarak ülke dışına verilen göçlerin Fiji'nin nüfusu üzerindeki etkisi incelenmiştir (Qiokata ve Khan, 2015). Yapay sinir ağları yöntemi ile geliştirilen bu modeller ile alınan sonuçlar Kuşak Bileşenleri yöntemi kullanılarak üretilen nüfus projeksiyonu sonuçları ile karşılaştırılmıştır.

Yang ve diğ, (2018) Çin'deki doğum senaryoları üzerine çalışma yapmıştır. Çin tarafından yürütülen doğum sınırlaması yasasının değişmesinin ülkenin gelecekteki nüfusu üzerindeki etkisi incelenmiştir. Senaryolara göre ülkenin gelecekteki nüfusu, kuşak bileşenleri yöntemi ve en küçük kareler destek vektör makineleri algoritması ile tahmin edilerek sonuçlar gerçek değerler ile karşılaştırılmıştır. (Li ve diğ, 2018)

Literatürde, kuşak bileşenleri yönteminin uygulama alanlarını ve makine öğrenmesi algoritmalarının veriler üzerindeki tahmin yeteneğini konu alan farklı çalışmalar bulunmaktadır. Ancak, makine öğrenmesi algoritmaları ile nüfus tahmini üzerine yeterli çalışma bulunmamaktadır. Bu çalışmaların bazılarında kuşak bileşenleri yöntemine alternatif olarak yapay sinir ağları ile nüfus tahmin edilmiştir. Yang ve diğ, (2018) yaptığı çalışmada ise nüfus tahmini için sadece destek vektör makineleri algoritması kullanılmıştır. Bu sebeple, nüfusun tahmininde kullanılabilecek farklı makine öğrenmesi algoritmaları değerlendirilecek ve geliştirilen alternatif model ile nüfus projeksiyonları üretilerek literatürdeki bu boşluğun doldurulacağı düşünülmektedir.

2. MAKİNE ÖĞRENMESİ

Makine öğrenmesi terimi ilk olarak 1950 yılında Turing'in makineler öğrenebilir mi sorusuna cevap aramasıyla ortaya çıkmıştır (Turing, 1950). Bu sorunun sorulması ile beraber birçok araştırmacı makinelere farklı öğrenimler kazandırarak uygulama alanları geliştirmişlerdir. Bu yıllarda Marvin Minsky ve Dean Edmonds tarafından geliştirilen yapay sinir ağlarına dayalı ilk bilgisayar olan SNARC (Stokastik Sinirsel Analog Güçlendirme Hesaplayıcısı) (Crevier, 1993), IBM'de Arthur Samuel tarafından geliştirilen satranç oyunu (McCarthy ve Feigenbaum, 1990) makine öğrenmesi uygulamalarına örnek olarak verilebilir. 18. yüzyılda akademiye kazandırılan ve makine öğrenmesinin gelişimine olanak sağlayan Bayes teoremi (Bayes ve Price, 1763; Laplace, 1812), En Küçük Kareler yöntemi (Least Square) (Legendre, 1805) ve Markov zinciri (Markov chains) (Markov, 1906) gibi matematiksel yöntemlerin incelenmesi makine öğrenmesinde kullanılan algoritmaların temelini anlamakta daha faydalı olacaktır.

1950 yılından bu zamana yapay zeka ve derin öğrenme (Deep learning) terimlerinin ortaya çıkmasıyla birlikte makine öğrenmesi uygulamaları birçok farklı alanda gelişim göstermiştir. Özellikle 1990'lı yıllarda insanların teknolojiye olan erişiminin artmasıyla birlikte işlenmeyi bekleyen birçok veri toplanabilir hale gelmiştir. Büyük verilerin toplanması ve rahatça saklanabilmesi, gün geçtikçe bu verilerden faydalı bilgi elde etme ihtiyacını doğurmuştur. Veri madenciliği ile bu veriler işlenerek anlamlı hale getirilmiş, makine öğrenmesi algoritmalarıyla da öğrenilen bilgiler istenilen amaca yönelik hizmet vermesi için makineler tarafından programlanmıştır. 1990'lı yıllarda E-mail spam filtreleme özelliği gibi basit uygulamalar için kullanılırken günümüzde çok çeşitli alanlarda makine öğrenmesi kullanılmaktadır. Bunlara örnek olarak, medikal alanda resimlerin işlenerek hastalıkların daha kolay tespit edilebilmesi, satın alınan ürünlere yönelik reklamların yapılması, daha önce izlenilen ve sevilen filmlere göre sevilebilecek filmlerin tavsiye edilmesi verilebilir. Makine öğrenmesinin en iyi uygulama alanlarından biri de sürücüsüz araç kullanımıdır. Araç üzerindeki yapay

zeka sistemi anlık olarak topladığı verileri makine öğrenmesi algoritmaları ile analiz ederek sürücüsüz araç kullanımına imkan vermektedir (Tekin ve diğ, 2018).

Makine öğrenmesi algoritmaları, elle işlenemeyecek çok büyük verileri istatistiksel modeller kullanarak programlayıp örneklemden anlamlı bir çıkarım yapmaya çalışırlar. Makine öğrenmesi, örnek verileri kullanarak performans kriterini optimize eden bilgisayar programı olarak da tanımlanabilir. Ayrıca, gelecek için tutarlı tahminler yapabileceği gibi geçmişe yönelik anlamlı sonuçlar da çıkartabilir.

Makine öğrenmesi iki adımdan oluşur. Birinci adımda modeli öğrenmek için optimizasyon probleminin çözülmesi gerekir. Kullanılan veri büyük olduğundan çözümlerin saklanması için yeterli alanın bulunması ve çözüm süresinin kısaltılması gibi konular göz önünde bulundurularak problemi çözmek için kullanılan algoritmalar etkili ve verimli olmalıdır. İkinci adımda model öğrenildikten sonra yapılan çıkarımların anlamlı olması gerekir. Bazı uygulamalarda, elde edilen sonucun tutarlılığı önemli olduğu gibi, o sonuca ulaşmak için gereken süre ve işlenecek verinin büyüklüğü de önemlidir (Alpaydin ve Bach, 2014).

2.1 Makine Öğrenmesinin Kısımları

Literatürde pek çok makine öğrenmesi algoritması bulunmaktadır. Bunlardan bazıları; basit (naive) Bayes sınıflandırıcı, karar ağaçları, yapay sinir ağları, destek vektör makineleri, k-en yakın komşu algoritması, lojistik regresyon analizi, k-ortalamalar algoritmasıdır. Bu yaklaşımlardan bir kısmı kümeleme ve sınıflandırma yaparken bir kısmı ileriyi tahmin etme için kullanılmaktadır.

Makine öğrenmesi algoritmaları öğrenim özellikleri bakımından iki ana başlıkta incelenir; denetimli öğrenim (supervised learning) ve denetimsiz öğrenim (unsupervised learning) (Bell, 2014). Bu iki öğrenim sıkça kullanılmakla birlikte literatürde kendine yer bulan birçok öğrenme yöntemleri de vardır. Örneğin; Yarı Denetimli Öğrenme (Semi-Supervised Learning), Transdüktif Akıl Yürütme (Transductive Inference), Çevrimiçi Öğrenme (On-line Learning), Pekiştirmeli Öğrenme (Reinforcement Learning), Aktif Öğrenme (Active Learning) (Mohri ve diğ, 2012).

2.1.1 Denetimli öğrenme (supervised learning)

Denetimli öğrenmenin temel amacı eğitim seti adı verilen giriş verilerini ve çıkış verilerini bir denetmen gözetiminde makineye vererek bu bilgilerden anlamlı bir sonuç çıkartmasını sağlamaktır. Denetimli öğrenme yapılması için öncelikle girdilerin ve çıktıların bulunduğu bir eğitim seti bulunmalıdır. Bu eğitim setindeki her bir veri bir denetmen aracılığı ile makineye öğretilir. Makineye veriler arasındaki ilişki öğretildikten sonra makine, öğrendiği ilişkiyi kullanarak daha önce hiç tanıtılmamış bir örneklem için varsayımlarda bulunabilmektedir.

Denetimli öğrenme sisteminde denetçinin görevi sisteme her bir veri için girdileri ve amaçlanan çıktılar göstermektir. Örnek vermek gerekirse, Twitter sosyal platformunda paylaşılan tweetleri sınıflandırmak isteyen bir kişi öncelikle sisteme atılan her bir tweeti girdi olarak göstermelidir. Ardından her bir girdinin amaçlanan çıktısını yani hangi konuda sınıflandırılabileceğini sisteme gösterir. Sistem bu büyük eğitim setindeki girdilerin ve çıktıların arasındaki bağlantıyı öğrendikten sonra yeni atılan her bir tweeti öğrendiği eğitim setinden faydalanarak konusuna göre sınıflandırma yapabilir (Bell, 2014).

Denetimli öğrenme yöntemi, eğitim setinden faydalanarak öğretilen ile gerçek arasındaki hata farkını en aza indirgemeye çalışır. Hata farkı her bir girdinin algoritma tarafından üretilen çıktısı ile gerçek değeri arasındaki fark olarak tanımlanabilir. Sisteme eğitim seti öğretilmesi aşamasında her bir veri için hata farkı tespit edilerek daha önceden belirlenmiş hata eşik değeri ile karşılaştırılır. Eğer eğitim sürecindeki hata değeri bu eşik değerinden küçük ise sistem eğitilmeye devam ettirilir. Sistem bu amaçlanan eşik değerine ulaştığında eğitim tamamlanır. Model eğitildikten sonra sisteme daha önce görmediği ve eğitim setinde bulunmayan bir girdi verilerek gerçeğe en yakın şekilde bir çıktı vermesi hedeflenir. Maillerin spam olarak sınıflandırılması denetimli öğrenme uygulamasına örnek olarak verilebilir (Mohri ve diğ, 2012).

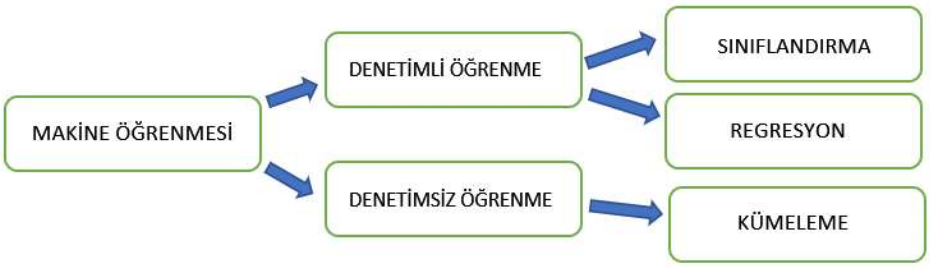
Denetimli öğrenme algoritmaları regresyon ve sınıflandırma olmak üzere iki temel amaç için kullanılır. Zaman serisi ile öngörümleme problemleri denetimli makine öğrenmesi yaklaşımı ile çözülebilmektedir. (Bontempi ve diğ, 2012). Literatürde sıkça kullanılan algoritmalara örnek olarak şunlar verilebilir; Destek Vektör Makineleri, Karar Ağaçları, Yapay Sinir Ağları (YSA), Zaman Serisi Analizi, k-en yakın komşu algoritması, lojistik regresyon analizi, basit (naive) Bayes sınıflandırıcıları.

2.1.2 Denetimsiz öğrenme (unsupervised learning)

Sisteme eğitim seti olarak sadece girdi değişkenleri verilir. Eğitim setinde çıktı ya da etiketleme bilgisi bulunmaz. Ayrıca sisteme öğretim yapan bir denetçi de yoktur. Sistem, girdi değişkenlerinin arasında ilişkiyi inceleyerek sınıflandırma, kümeleme gibi işlemler yapmaya çalışır. Eğitim setinde sınıflandırma ya da çıktı bilgisi bulunmadığı için sistemin öğrenme performansını ölçmek zordur (Mohri ve diğ, 2012).

Denetimsiz öğrenme algoritmaları büyük veri içerisindeki gizli örüntüleri (pattern) bulmaya çalışır. Bu öğrenme yöntemi için doğru ya da yanlış bir çözüm yoktur, ancak verinin sürekli işlenerek örüntülerin ortaya çıkarılması ve anlamlı bir bütünlük sağlanması amacı vardır. Bu açıdan bakıldığında denetimsiz öğrenme modeli daha çok veri madenciliğine benzemektedir (Bell, 2014). Örneğin, bir turizm şirketi düzenlediği organizasyonların ardından müşterileriyle anket düzenleyebilir. Anket sonuçları ile müşteri profilini denetimsiz öğrenme algoritmalarıyla işleyerek müşterileri kümelendirebilir. Bu sonuçları kullanarak müşteri gruplarına özel paketler ya da promosyonlar üretebilir. Diğer bir örnek olarak resimlerin boyutlarının küçültülmesi verilebilir. Birbirine yakın renklerde olan pikseller kümelendirilerek bir görüntü elde edilir. Bu görüntü gerçek görüntünün sıkıştırılmış bir hali olduğu için daha az yer kaplar fakat gerçek görüntünün iyi bir temsilidir (Alpaydin ve Bach, 2014).

Kümeleme (Clustering) ve Boyut Azaltma (Dimensionality Reduction) denetimsiz öğrenme algoritmalarının temel amaçlarındandır. Tekil Değer Ayrıştırması, Korelasyon analizi, Temel Bileşen Analizi, Hiyerarşik Kümeleme, faktör analizi, k-ortalama kümeleme denetimsiz öğrenme yöntemlerinden bazılarıdır. Şekil 2.1'de denetimli öğrenme ile denetimsiz öğrenme yöntemleri sınıflandırılmıştır (Mathworks, n.d.).



Şekil 2.1 : Denetimli öğrenme ve Denetimsiz öğrenme karşılaştırması.

2.1.3 Yarı denetimli öğrenme (semi-supervised learning)

Yarı Denetimli öğrenme yönteminde, kullanılan eğitim setindeki verilerin büyük çoğunluğunu denetimsiz öğrenme yönteminde olduğu gibi etiketleme bilgisi bulunmayan girdi değişkenleri oluşturur. Bununla birlikte, veri setinde az miktarda girdi ve çıktı bilgisi bulunan veriler de vardır. Etiketleme bilgisi bulunan çıktı değerlerini elde etmenin zahmetli ve maliyetli olduğu alanlarda bu öğrenme yöntemi kullanılır. Örnek olarak bir ses kaydının transkripte edilmesi verilebilir. Sistem sadece girdi değişkenlerine bakmakla kalmayıp aynı zamanda etiketleme bilgisi bulunan verilerden de faydalanarak modeli öğrenir. Bu öğrenim ile yeni bir girdi için hedeflenen çıktı bilgisini en iyi şekilde tahmin etmeye çalışır. Yarı denetimli öğrenme yöntemi regresyon, sınıflandırma, sıralama gibi birçok farklı problemi çözmek için kullanılabilir (Mohri ve diğ, 2012).

2.1.4 Transdüktif akıl yürütme (transductive inference)

Yarı denetimli öğrenme yönteminde olduğu gibi bu yöntemde de sisteme etiketlenmiş ve etiketlenmemiş eğitim seti verilir. Fakat Transdüktif Akıl Yürütme yönteminin asıl amacı etiketlenmiş çıktı değerlerini kullanarak etiketlenmemiş verilerin hedeflenen çıktı değerlerini tespit etmektir. Modele sonradan yeni bir veri verilerek varsayım yapması istenmez (Cortes ve diğ, 2009). Bu terim ilk olarak Vapnik tarafından literatüre kazandırılmıştır (Vapnik, 1982).

2.1.5 Pekiştirmeli öğrenme (reinforcement learning)

Pekiştirmeli öğrenme yöntemi ile sistem, karşılaştığı eylemler karşısında kendisini adapte ederek çevreden gelen iyi değerlendirme sonuçlarını maksimize etmeye çalışır (Sutton ve Barto, 1998). Eğitim ve test aşamalarının birlikte yapıldığı bu yöntemde sistem denetçi ile sürekli ilişki içerisindedir. Eğitim setinde her bir girdi için istenilen çıktı hazır bulunmaz. Denetçi sistemin ürettiği çıktıları her defasında girdilere göre iyi veya kötü şeklinde değerlendirerek (ödül) sisteme modeli öğretmeye çalışır. Bu modelde tek bir sonuçtan ziyade süreç önemlidir. Sistem, aldığı çıktıların değerlendirme sonuçlarına göre kendini geliştirerek en çok ödülü almayı hedefler.

Pekiştirmeli öğrenme yöntemine örnek olarak satranç oyunu verilebilir. Satranç oyunu basit kurallar içermesine rağmen birçok farklı hamle kombinasyonları içerdiği için rakibin stratejisinin önceden tahmin edilmesi zordur. Pekiştirmeli öğrenme modeli ile

eğitilen sistem karşıdaki oyuncunun her bir hamlesini değerlendirerek rakibine karşı en iyi hamleyi yapmaya çalışır. Rakibinin stratejisini çözdüğünde gelebilecek hamleleri önceden tahmin ederek ona göre pozisyon alabilir (Alpaydin ve Bach, 2014).

2.2 Makine Öğrenmesi Uygulama Süreci

Bir problem makine öğrenmesi algoritmaları ile çözülmesi istediğinde takip edilmesi gereken başlıca adımlar vardır. Problemin başarıyla çözülebilmesi için bu adımların sırasıyla uygulanması gerekir. Literatürde bu adımlar farklı şekillerde açıklansa da benzer manaları ifade etmektedir. Bell (2014) bu adımları dört ana başlıkta toplamıştır.

1. Verilerin toplanması
2. Verilerin hazırlanması
3. Modelin çalıştırılması
4. Sonuçların raporlanması

Lantz (2013) ise makine öğrenmesi algoritmalarının uygulanması sürecini beş adımda anlatmıştır. Bu adımlar tamamlandıktan sonra istenen amaca yönelik model kullanılabilir. Bu adımlar;

1. Verilerin toplanması
2. Verilerin hazırlanması
3. Modelin eğitilmesi
4. Modelin değerlendirilmesi
5. Modelin geliştirilmesi

Ayrıca, veri madenciliği uygulama sürecinde "Veri Madenciliği için Çapraz Endüstri Standart Süreç Modeli" (CRISP-DM, "Cross-Industry Standard Process for Data Mining") adı verilen yaklaşım kullanılmaktadır. Bu yaklaşım Daimler, Integral Solutions Ltd. (ISL) gibi dünyanın önde gelen 200'ü aşkın firmanın bulunduğu konsorsiyum tarafından geliştirilmiştir (Shearer, 2000). Veri madenciliğinin makine öğrenmesi uygulama sürecine benzerliğinden dolayı CRISP-DM yaklaşımı makine öğrenmesi algortimalarının uygulanmasında da kullanılabilmektedir. CRISP-DM yaklaşımı beş ana başlıktan oluşmaktadır.

1. Problemin tanımlanması
2. Verinin toplanması ve analizi
3. Verinin hazırlanması

4. Modelin seçilmesi ve kurulması
5. Modelin değerlendirilmesi
6. Modelin çalıştırılması ve sonuçların raporlanması

2.2.1 Problemin Tanımlanması

Bir problemin makine öğrenmesi algoritması ile çözülebilmesi için öncelikle problemin iyi tanımlanması gerekir. Öyle ki, bir problemin iyi tanımlanması çözülmesinden daha çok önem arz eder. Problemi tanımlarken şunlar göz önünde bulundurulmalıdır: çözümün neden önemli olduğu, çözümün getireceği faydalar ve problemin ortaya çıkardığı zararlar. Bu maddelerin sağlıklı olarak tamamlanabilmesi için mevcut durumda problem hakkındaki bütün bilgiler toplanmalı ve problemin çözümü için hangi bilgilere ihtiyaç duyulabileceği listelenmelidir. İhtiyaç duyulan bilgiler için araştırma yapılarak, problem açıkça tanımlanmalıdır. Problem tanımını yaparken anahtar kelimeler belirlenerek tanım geliştirilir. Problemin çözüm amacı belirlenir. Problemin çözüm amacının iyi tespit edilmesi makine öğrenmesi modellerinin kurulmasını kolaylaştırmaktadır. Neyi sorguladığını bilmeyen araştırmacı çözüme nereden başlayacağını bilemez (Bell 2014). Problemin çözüm amacı ile beraber, başarı kriterinin de tanımlanması gerekir. Elde edilen sonuçlar başarı kriterine göre değerlendirilerek raporlanır. Ayrıca problemin olası çözüm yollarının makine öğrenmesi algoritmalarına uygunluğu tartışılmalıdır.

2.2.2 Verinin toplanması ve analizi

Probleme uygun olan verilerin sağlıklı toplanması kurulacak olan modelin tahmin yeteneğini etkiler. Araştırmacı başkaları tarafından toplanmış olan verileri alarak probleme olan uygunluğunu gözden geçirip kullanabilir. Verileri internet ortamında ham ya da işlenmemiş şekilde bulunabilir. Türkiye İstatistik Kurumu (TÜİK) tarafından açıklanan işsizlik verileri işlenmiş hazır verilere örnek olarak gösterilebilir. Kağıt üzerinde yapılan anket sonuçları ise analize hazır hale getirilmemiş ham verilere örnektir. Araştırmacı projesine uygun olan verileri bulmak için öncelikle internet üzerinden veritabanlarını araştırır. Gerektiği takdirde ilgili şirketlere hususi başvurular yaparak veri talebinde bulunabilir. Ancak bu noktada bazı şirketler çekimser davranabilmektir. Bunun başlıca sebeplerinden biri şirket verisinin üçüncü şahıslar ile paylaşıldığında güvenlik açığının oluşması endişesidir. Şirket bu verilerin rakip firmalar tarafından rekabet için kullanılabileceği kaygısı duyar. Diğer bir endişe ise

verilerden elde edilen raporların şirketin aleyhinde kullanılması ihtimalidir. Ayrıca şirketlerin ilgili veriyi vermek için yeterli teknik altyapıya sahip olmaması ya da veriyi derlemenin zahmetli ve maliyetli olması gibi unsurlar dezavantaj olarak nitelendirilebilir.

Yeterli veri olmadığı durumlarda araştırmacı kendi imkanlarıyla veri toplayabilir. Veri toplamak için anket çalışması, mülakat, gözlem, deney, şirket raporlarının incelenmesi gibi veri toplama araçlarını kullanabilir. Veriler kağıt üzerinde, yazı formatında, SQL veritabanı gibi farklı platformlarda olabilir. Verinin analize hazır olması için uygun elektronik ortama taşınması gerekir.

Başarılı bir makine öğrenmesi uygulaması yapmanın yolu sağlıklı veri toplamaktan geçer (Bell, 2014). Toplanan verinin formatı, niceliği, her bir sütun ve satırdaki kayıtlar, verilerin konu başlıkları araştırmacı tarafından incelenir. Mod, medyan, ortalama, standart sapma gibi istatistiksel analizler yapılarak ve histogram, pasta grafiği ve dal yaprak grafiği gibi görsel analizler yapılarak veri incelenir. Burada önemli olan toplanan verinin analiz için yeterli olup olmadığıdır. Örneğin, eğer analizde her yıl için ayrı veri gerekiyorsa ancak toplanan veride beş yıllık ortalama veriler bulunuyorsa farklı bir veri setinin araştırılması ya da toplanması yoluna gidilebilir (Shearer, 2000).

2.2.3 Verinin hazırlanması

Veri hazırlanması aşamasında modele öğrenmesi için verilecek olan veri seti hazırlanır. Veri setinin hazırlanması için verilerin seçilmesi, temizlenmesi, formatının uygunluğu, dönüştürülmesi ve entegrasyonu gözden geçirilir. Veri hazırlama süreci her bir veri seti için projenin amacına göre değişiklik göstermektedir.

Veri setinin seçilmesinde verilerin projenin amacına uygunluğu, kalitesi, teknik kısıtlamaları gözden geçirilir. Örneğin, kişilerin ikametgah bilgilerinin bulunduğu bir veri setinde eğer ilçe bilgisi veri analizi için yeterli ise, veri setindeki mahalle ve cadde ayrıntıları göz ardı edilebilir. Bu aşamada verilerin neden veri setinde dahil edildiği ya da edilmediği açıklanmalıdır.

Temiz veri kullanılmadan yapılan makine öğrenmesi uygulamasının sonuçları tartışmaya açıktır (Shearer, 2000). Kullanılan veri seti temiz değilse veri temizleme teknikleri uygulanmalıdır. Buradaki temizlikten maksat gürültü (noisy) adı verilen veri

seti içerisindeki tutarsız, hatalı verilerdir. Bunların temizlenmesi ile ancak veri seti makine öğrenmesi algoritmalarının kullanabileceği şekilde hazır hale gelebilir.

Veri setini gürültüden arındırdıktan sonra kayıp değer (missing value) adı verilen veri setindeki boşluklar gözden geçirilmelidir. Herhangi bir sebepten dolayı veri setinde boşluklar oluşabilir. İlgili bilgiye ulaşım kısıtlanabilir ya da ölçülmesi maliyetli olduğu için araştırmacı bu zamana kadar ölçülmüş olan veri ile yetinmek zorunda kalabilir. Bu tür durumlar makine öğrenmesi algoritmalarının performansını etkileyebilir (Suthar ve diğ, 2012).

Kayıp değerlerin tamamlanması ya da ilgili verinin veri setinden atılması için literatürde birçok yöntem vardır (Rahman ve Davis, 2013; Batista ve Monard, 2003; Suthar ve diğ, 2012):

- Eğer kayıp değerlerin veri setinden atılması çalışmayı önemli ölçüde etkilemeyecek ise ilgili veriler tamamen atılabilir.
- Aynı manayı ifade eden birden çok veri varsa ya da veride büyük boşluklar bulunuyorsa ilgili sütun veri setinden tamamen kaldırılabilir.
- Kayıp değerler yerine ilgili gözlemdeki verilerin ortalaması ya da modu adı verilen en sık tekrar edilen değer atanabilir.
- İnterpolasyon yöntemi ile veri seti üzerindeki kayıp değerler tamamlanabilir (Brubacher ve Wilson, 1976; Moritz ve diğ, 2015).
- Çok katmanlı algılayıcı (Multilayer perception MLP), karar ağacı (decision tree DT), özdüzenleyici haritalar (self-organising maps SOM) gibi bazı makine öğrenmesi algoritmaları kayıp değerlerin tamamlanması için kullanılabilir.

Bazı durumlarda veri seti içerisindeki bazı gözlemlerin tek bir tipe dönüştürülmesi gerekebilir. Verilerin aynı dili konuşuyor olması ve aynı birimde ifade edilmeleri makinenin öğrenme aşamasında hata yapmasını engeller. Basit bir örnek vermek gerekirse; bir değişken için saat biriminde, diğer bir değişken için dakika biriminde veriler bulunuyorsa, bütün verilerin tek bir birime dönüştürülmesi gerekir.

Büyük verilerin kullanılması durumunda makinenin veriyi öğrenmesi ve işlenmesi çok uzun sürebilmektedir. Bu süre bilgisayarın özelliklerine bağlı olarak saatler hatta günler sürebilmektedir. Bu süreyi kısaltmak için veri sayısının azaltılması işlemine veri indirgeme adı verilir.

Veri seti bu ön hazırlık aşamalarından geçtikten sonra makineye öğrenmesi için verilmeden önce öznitelik mühendisliği adı verilen işleme tabi tutulur. Öznitelik mühendisliğinin başlıca uygulamaları arasında normalizasyon, aykırı değerler (outlier), veri ölçekleme, gauss ve log transformasyon vardır.

2.2.1.1 Normalizasyon

Veri setinde yer alan değerler farklı aralıklarda bulunabilir. Örneğin; yıllar 1950-2020 arasında değişirken, yaş seviyeleri 0-80 arasında değişkenlik gösterir. Bu durumda aralığı büyük olan değerler diğerleri üzerinde baskın rol oynayarak sonucu etkileyebilir. Bunun önüne geçilmesi, verideki tekrarları ortadan kaldırmak ve verinin tutarlılığını arttırmak için minimum-maksimum, Z-dönüşümü ve ondalık ölçekleme gibi farklı normalizasyon yöntemleri kullanılabilir.

2.2.1.2 Uç değerler (outliers)

Veri setindeki bazı değerler bulunduğu gözlem verilerinden çok uzak sonuçlar gösterdiği için uç noktalar olarak adlandırılır (Grubbs, 1969). Veri toplanırken olağandışı bir durum yaşanması, veri girişi sırasında hata yapılması ya da açıklanamayan farklı bir sebepten dolayı bu tür uç noktalar ile karşılaşılabilir .

2.2.4 Modelin seçilmesi ve kurulması

Belirlenen problemin çözümü için kullanılacak olan makine öğrenmesi algoritmasının seçilme işlemidir. Modelin seçilmesi kullanılacak olan veri setinin özelliklerine göre değişim gösterir. Bir veri setine uygulanabilecek birçok farklı makine öğrenmesi algoritması bulunmaktadır. Bu noktada bu modellerden birkaçını öğrenip uzmanlaşarak bütün verilere uygulamaya çalışmak çekici gelse de bu tavsiye edilmemektedir. Bütün veri setlerine uygulanabilen ve bütün problemlere çözüm bulabilen tek bir tane ya da bir grup makine öğrenmesi algoritması yoktur. Her bir algoritmanın her bir veri seti ve problem karşısında birbirlerine karşı üstün ve zayıf yönleri vardır. (Wolpert ve Macready, 1997).

Toplanan verinin karakteristik özelliklerine ve problemin çözüm amacına bakılarak uygun olan makine öğrenmesi algoritması seçilir ve model kurulur. Bir tane algoritma seçmektense probleme çözüm bulabilecek birden fazla algoritma seçilerek uygulanması proje sonunda alınacak olan verimi arttıracaktır. Makine öğrenmesi

algoritmaları veri setine bağlı olarak sayısal tahmin, sınıflandırma, kümeleme, örüntü bulma gibi amaçlar için kullanılabilir (Lantz, 2013).

2.2.5 Modelin değerlendirilmesi

İlk aşamada en doğru algoritmayı seçmek zor olacağı için verilerin temizlenmesinde bu yana seçilecek olan model düşünülmeli, model kurulup test edildikten sonra istenen hedef değerlere ulaşılamadığında model tekrar gözden geçirilmelidir. Eğer model hedeflenen çözüme ulaştırmak için yetersiz kalıyor ve beklentileri karşılamıyorsa geliştirilmeli ya da değiştirilmelidir. Modelin performansını değerlendirmek için literatürde birçok yöntem mevcuttur. Bunlara örnek olarak holdout ve çapraz-doğrulama(cross validation) yöntemleri verilebilir (Lantz, 2013).

Holdout yönteminde veri seti eğitim seti ve test seti olmak üzere iki kısma ayrılır. Makine eğitim setini kullanarak modeli öğrenir. Öğrenilen model test seti üzerinde çalıştırılarak sonuçlar değerlendirilir. Genelde verilerin üçte ikisi eğitim setine ayrılırken üçte biri test setine ayrılır. Fakat bu ayrımı yaparken örneklemin rassal dağıtılması sağlanmalıdır. Bu yöntemde veri seti bir defa ayrıldığı için bu ayırma işleminde yapılan herhangi bir hata modelin çalıştırılması sonucu elde edilen çözümü etkileyebilir.

Çapraz-doğrulama yöntemi ise problem çözümü için kullanılan bir veya birden çok modelin performansını ölçmek için kullanılır. Çarpaz-doğrulama yöntemi k-kat çapraz doğrulama(k-fold) ve birini dışarıda bırak çapraz doğrulama (leave one out, LOOCV) olmak üzere iki kısımdır. K-kat çapraz doğrulama yönteminde veri seti rassal olarak k eşit parçaya ayrılır. Her bir parça test seti olarak belirlenir ve kalan parçalar ile model eğitilir. Bu döngü bütün parçalar test seti olana kadar devam ettirilir. Sonuç olarak elde edilen bütün tahmin hatalarının ortalaması alınarak modelin hata ortalaması elde edilmiş olur. Bu ortalamaya göre model değerlendirilir. K değeri herhangi bir sayı olabilir ancak genelde 10 değeri kullanılır. Birini dışarıda bırak yönteminde ise k değeri örneklem sayısına eşit alınır. Veri setinin küçük olduğu zamanlarda kullanılan bu yöntemde sadece 1 gözlem verisi test verisi olarak ayrılır. Geriye kalan veriler ile model eğitilerek test verisi üzerinden değerlendirme yapılır (Refaeilzadeh ve diğ, 2009; Lantz, 2013; Susanto ve diğ, 2016; Modaresi ve diğ, 2018).

Zaman serisi içeren veri seti analizlerinde çapraz doğrulama yöntemi uygulanmaz. Çapraz doğrulama yöntemi veriyi eğitim ve test setini rassal bir şekilde ayırır. Zaman

serisi verilerinde zamanın veriler üzerindeki etkisinde dolayı rastgele bir ayrım yapılması veri setinin yapısına uygun düşmez. Zaman serisi analizlerinde geçmişe bakılarak gelecek öngörülmeye çalışıldığı için çapraz geçerlilik uygulandığında model gelecekteki veri seti ile eğitilerek geçmişi tahmin etmeye zorlanabilir. Bu gibi hataların önüne geçmek için Back Test,Nested Cross Validation, Blocked Cross Validation gibi çözümler kullanılmaktadır. Bu yöntemler ile veri seti zaman serisi problemine uygun olarak eğitim ve test seti şeklinde kısımlandırılır. Test seti olarak ayrılan gözlem değerlerinin zaman açısından eğitim seti olarak ayrılan gözlem değerlerinden sonra gelmesi sağlanır (Bergmeir ve Benítez, 2012).

Modellerin değerlendirilmesi aşamasında tahmin edilen değerler ile test değerleri karşılaştırılarak Ortalama Kare Hata (Mean Squared Error, MSE), R Kare (R Squared, R^2), Ortalama Mutlak Yüzde Hata (Mean Absolute Percentage Error, MAPE) gibi değerlendirme kriterlerine göre raporlanır. Ortalama kare hatası, her bir veri için tahmin edilen değer ile beklenen değer arasındaki farkın karesi hesaplanarak toplamın ortalaması alınır. R kare ise regresyon analizinde varyansı açıklama gücü olarak tanımlanır. Ortalama mutlak yüzde hatası ise regresyon ve zaman serileri analizinde sıkça kullanılmakta olup tahmin edilen değer ile gerçek değer arasındaki farkın gerçek değere nispeti alınarak hesaplanır.

Eğitim seti modele uygulandığında elde edilen MSE değeri küçük ancak, test seti modele uygulandığında elde edilen MSE değeri büyük çıkıyorsa model veri ile aşırı uyum (overfitting) halinde olduğu söylenir. Model eğitim setini öğrenmeye çalışırken veri setinin aslında bulunmayan rassal olarak ortaya çıkmış örüntüleri öğrenebilir. Bu da aşırı uyumun ortaya çıkmasına sebep olabilir. Model ve veri arasında aşırı uyum oluştuğunda eğitim setinde gözlenen örüntü test setinde gözlemlenemeyeceği için MSE değeri çok yüksek çıkacaktır (James ve diğ, 2013).

2.2.6 Modelin çalıştırılması ve sonuçların raporlanması

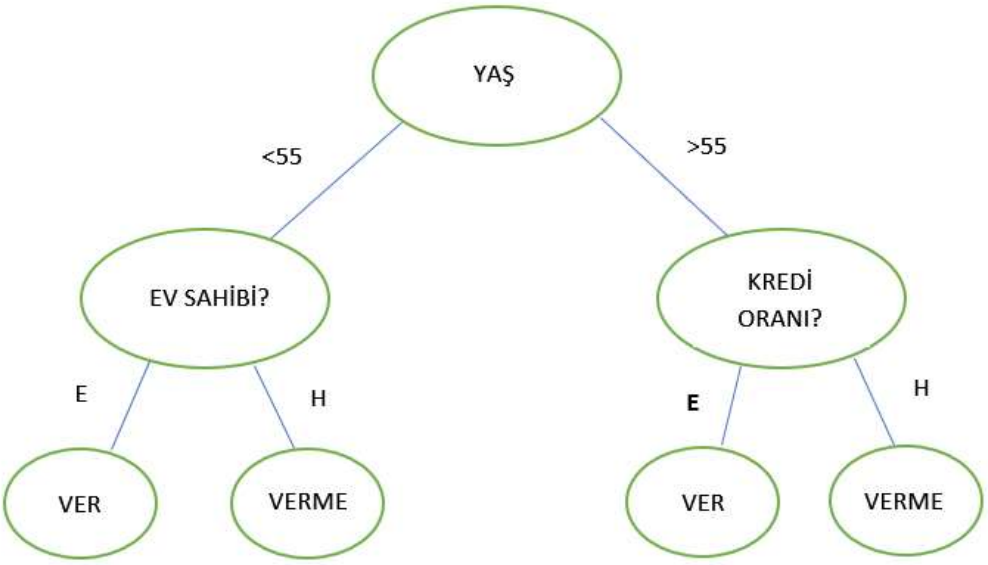
Problem iyi tanımlandıktan sonra çözüm aşaması için veri toplama işlemi yapılır. Toplanan veri temizleme, dönüştürme gibi işlemlerden geçirilerek kullanıma hazır hale getirilir. Problemin çözümü için kullanılacak makine öğrenmesi algoritmaları belirlendikten sonra hazırlanan veri kullanılarak modeller eğitilir. Kullanılan makine öğrenmesi algoritmaları arasından yapılan değerlendirmeler sonucu en iyi performansı gösteren modeller problemin çözümü için kullanılır.

Modellerin çalıştırılması aşamasında veri seti ya da model açısından herhangi bir sorun tespit edilirse geriye dönük adımlar tekrar değerlendirilerek düzeltmeler yapılabilir. Problemin çözümüne en uygun modeller kullanılarak problem için olası çözümler raporlanır. Eğer veri setinde veri dönüştürme işlemleri uygulanmış ise, elde edilen sonuçların birimleri istenen manayı ifade edecek şekilde tekrar dönüştürülür.

3. MAKİNE ÖĞRENMESİ ALGORİTMALARI

3.1 Karar Ağaçları (Decision Trees)

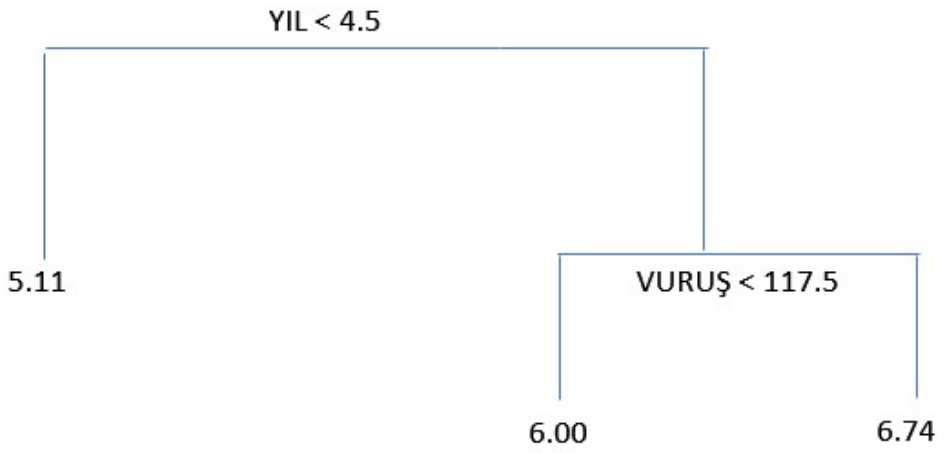
Karar ağaçlarının amacı veri setinde değişkenleri kullanarak hedeflenen değeri tahmin edebilecek bir ağaç modeli oluşturmaktır. Karar ağaçları kök ile başlar. Bütün veriler düğüm adı verilen karar verme noktalarından geçerek dallara ayrılır ve yapraklara ulaşır.Düğüm sonucunda verilen karar çeşitliliğine göre dallar oluşturulur. Son olarak dallar yapraklara yani sınıflandırılan ya da sayısal değer atanan bir çıktı değerlerine ulaşır. Dallar birbirine karışmazlar, düğüm aşamalarından geçerek yukarıdan aşağıya doğru ağaç biçimine benzer bir yapı oluştururlar. Şekil 3.1'de bir karar ağacı örneği verilmiştir (Bell, 2014).



Şekil 3.1 : Karar ağacı sınıflandırma örneği (Bell, 2014).

Şekilde görüldüğü gibi karar ağacı kredi verilmesi kararını sınıflandırmaktadır. Kök düğüm yaşa göre başvuruları yani girdileri dallara ayırır. İkinci seviyedeki düğümlerde ise model 55 yaş altındaki kişilerin ev sahibi olup olmadığını sorgular. Ev sahibi olan kişilere kredi vermeyi onaylarken ev sahibi olmayanlar için kredi verilmeyeceğini

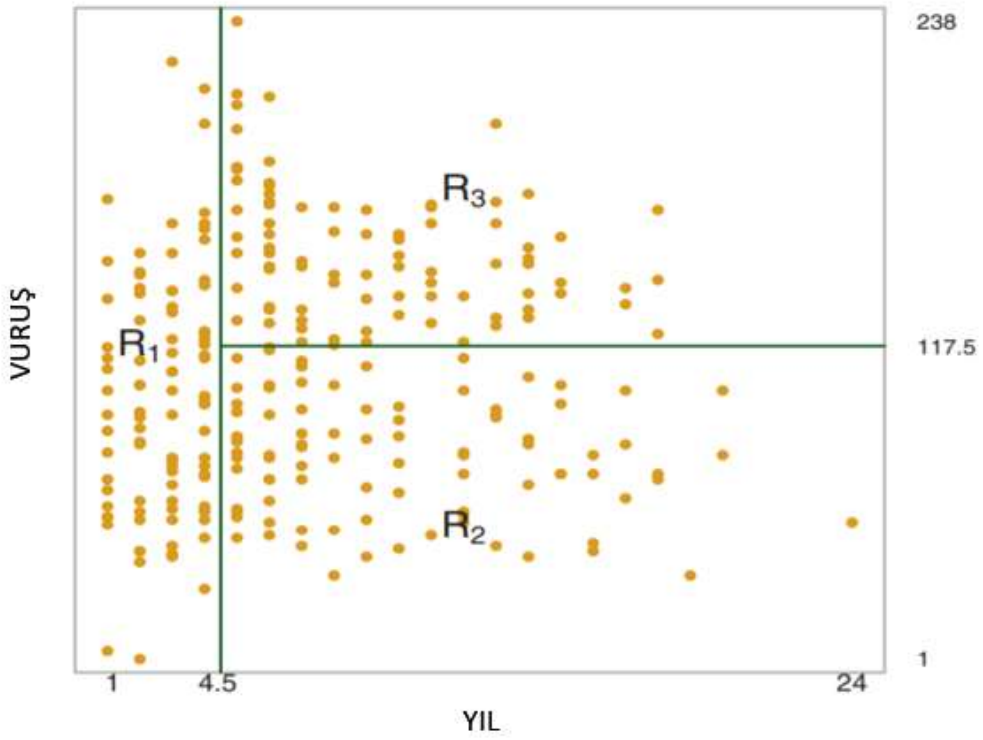
gösterir. Bu karar ağacı denetimli öğrenme ile sınıflandırma yöntemine örnektir. Karar ağaçları regresyon problemleri için de kullanılabilir. Şekil 3.2'de bir karar ağacı regresyon örneği verilmiştir (James ve diğ, 2013).



Şekil 3.2 : Karar ağacı regresyon örneği (James ve diğ, 2013).

Bu örnekte bir beyzbol oyuncusunun oynadığı yıl sayısı ve vurduğu top sayısına göre maaşına karar verme aşaması gösterilmiştir. Kök düğümde oyuncu oynadığı yıl sayısına göre iki dala ayrılır. 4.5 yıldan az oyunculuk yapan girdilerin yani kişilerin ortalama maaş oranı 5.11'dir. Daha uzun yıllar oyunculuk yapanlar için ise tekrar bir karar düğümü oluşturulmuştur. Vurduğu top sayısı 117.5'tan fazla olanların maaş ortalaması 6.74 oranında iken, az olanların maaş ortalaması 6 oranındadır. Bu ağaçta iki düğüm ve üç yaprak bulunmaktadır. Her bir yaprak ilgili hedef değer için log transformasyon yapılmış ortalama maaş oranını göstermektedir.

Yukarıdaki örnekte beyzbol oyuncularının tecrübesine ve vurduğu top sayısına göre ortalama maaşlarını tahmin için karar ağacı regresyonu kullanılmıştır. Öncelikle girdi değerlerinde maaş bilgisi içermeyen gözlemler veri setinden atılmıştır. Daha sonra maaşlar yüzbinler basamağı kullanılarak ifade edildiğinden, dağılımın normale yaklaşması ve çan-eğrisi oluşturması için her bir değerin log transformasyonu alınmıştır. Gözlem verileri oyuncuların tecrübesi, vurduğu top sayısı ve maaşına göre analiz edildiğinde aşağıdaki şekile ulaşılır.



Şekil 3.3 : Beyzbol oyuncularının üç bölgeye ayrılması (James ve diğ, 2013).

Karar ağacı regresyonu, beyzbol oyuncuları R1, R2 ve R3 olarak adlandırılan üç ayrı bölgeye ayırmıştır. R1 tecrübesi 4.5 yıldan az olanları gösterirken R2 değeri vurduğu top sayısı 117.5'tan az olup tecrübesi 4.5 yıldan fazla olanları gösterir. R3 değeri ise tecrübesi 4.5 yıldan fazla olup vurduğu top sayısı 117.5'tan fazla olanları göstermektedir. Her bölgenin maaş ortalaması hesaplandığında R1 bölgesi için \$1000xe^5.107 = \$165.174 ortalama maaş değerine ulaşılır. Diğer bölgeler için de aynı şekilde hesaplanabilir. Şekil 3.3'de görülen üç ayrı bölge karar ağacındaki üç yaprağı gösterir. Görüldüğü üzere yapraklar aşağıda olduğu için karar ağaçları tersten çizilir.

Örnek verilen karar ağacı regresyonu yorumlandığında maaşı en çok etkileyen faktörün tecrübe olduğu görülür. Tecrübesi az olan oyuncuların vurduğu top sayısı oyuncunun maaşını çok az etkilediği görülmektedir. Ancak tecrübeli oyuncuların vurduğu top sayısı maaşını önemli ölçüde etkilemektedir. Sonucu etkileyen faktörlerin önem sırası feature importance yöntemi ile elde edilir.

Yukarıdaki örnek basit bir karar ağacı regresyonu örneğidir. Karar ağacı regresyon yönteminin; sonuçların rahat yorumlanabilmesi, eğitim ve uygulama sürecinin hızlı olması ve anlaşılır görseller sunabilmesi gibi avantajları vardır (James ve diğ, 2013; Swetapadma ve Yadav, 2017).

Karar ağacı yöntemindeki en önemli nokta kök düğümlerden itibaren bütün düğümlerin hangi kritere göre ve nasıl dallanacağına karar vermektir. Bu noktada literatürde sıkça kullanılan ID3, C4.5, CART, CHAID, C5.0, gibi algoritmalar bulunmaktadır (Kass, 1980; Quinlan, 1986; Loh, 2011; Quinlan, 2014; Breiman, 2017).

Literatürde karar ağaçlarını temel alarak geliştirilen ve daha güçlü tahmin sonuçları veren üç farklı topluluk (ensemble) algoritmaları bulunmaktadır; Torbalama (Bagging), Rastgele Orman (Random Forest) ve Artırma (Boosting) (James ve diğ, 2013).

3.2 Torbalama (Bagging - BootStrap Aggregation)

Torbalama yöntemi ile veri setinde modelin eğitilmesi için kullanılacak olan eğitim seti bootstrap örnekleme ile daha küçük parçalara bölünür. Ardından bu farklı eğitim setleri tek bir algoritmanın eğitilmesi için kullanılarak farklı modeller oluşturulur. Modellerin tahmin ettiği sonuçlar oylama (voting) ile elenerek ya da ortalama alınarak sonuca ulaşılır. Buradaki ana düşünce birbirinden farklı eğitim seti ile eğitilen modelin tek bir eğitim seti ile eğitilen modele göre daha iyi sonuç çıkaracağı varsayımıdır (ensemble)(Breiman, 1996; James ve diğ, 2013).

Torbalama yöntemi rastgele orman yöntemine benzerlik gösterir. Aralarındaki en büyük fark, rastgele orman yönteminde ağacın her bir düğümü için değişkenler arasından rastgele bir alt küme seçilir. Seçilen bu alt küme ilgili düğümün dallanması için kullanılır. Bu rassal seçim orman içindeki ağaçların birbirine benzeme ihtimalini düşürür. Bu açıdan rastgele orman yöntemi torbalama yönteminin geliştirilmiş halidir (Breiman, 2001).

3.3 Rastgele Orman (Random Forest)

Rastgele orman yönteminde birden fazla karar ağacı oluşturularak bir orman meydana getirilir. Her bir ağacın verdiği sonuçlar oylama (voting) ile elenerek ya da ortalama alınarak problemin çözümüne ulaşılır.

Rastgele orman yöntemi, torbalama (bagging) ve Random Subspace yöntemlerinin harmanlanmasından ortaya çıkmıştır. Torbalama yönteminde olduğu gibi veri seti alt kümelere ayrılır. Random Space yönteminde uygulandığı üzere ağaç üzerindeki düğümlerden en iyi dallara ayrılan değişken, tüm değişkenler arasından rastgele seçilerek oluşturulan alt küme içerisinden seçilir.

Rastgele Orman algoritması uygulama aşamasında denetçiden iki farklı parametre talep edilir. Bu parametreler orman içindeki ağaç sayısı belirten (N) değeri ve her bir düğümde kullanılacak olan değişken (m) değeridir. M değeri için toplam p değişken sayısından sınıflandırma problemlerinde "√p", regresyon problemlerinde "p/3" değerinin alınması tavsiye edilir (Breiman, 2001). Daha sonra veri seti öğrenme ve test olarak iki kısma ayrılır. Genelde öğrenme seti (inBag) test setinin (Out-of-Bag, OOB) iki katı olur. Ormandaki her bir ağaç boostrap tekniği kullanılarak veri seti alt kümelerine ayrılır. Her bir alt küme için öğrenme ve test seti belirlenmelidir. Ardından her bir alt küme için kök düğüm ve dallanmalar başlar. Her karar düğümünde m değişken arasından rassal seçilen değişkenler ile en iyi dallanma belirlenir. En iyi dallanmayı belirlemek için yukarıda bahsedilen CART gibi algoritmalar kullanılabilir. Bu işlem başta belirtilen N tane ağaç oluşturulana kadar devam ettirilir. Elde edilen çözümler problemin yapısına göre oylama ya da değerlerin ortalaması alınarak nihai çözüme ulaşılır. Test verileri (OOB) kullanılarak elde edilen sonuçların hata oranları tespit edilir.

Her bir ağaç için OOB hata oranı tespit edildiği gibi ormanın da hata oranı tespit edilir. Böylece seçilen alt kümelerin veri setini ne derece yansıttığı ölçülür. Ayrıca her bir ağacın genel orman için bulunan OOB hata verisine olan katkısı hesaplanarak ağaçlar arasında bir değerlendirme yapılır. En az hatayı veren ağaç, yani en çok değişkeni açıklayan ağaç ormana en çok katkıyı veren ağaçtır. Rastgele orman yöntemi, karar ağaçlarına göre daha iyi sonuçlar verse de beraberinden bir takım dezavantajlar getirir. Karar ağaçlarında sonuçları görsel anlamda kolay ve güzel ifade edebilmek mümkün

iken, rastgele orman yönteminde bu işlem zorlaşır. Ancak, her bir değişkenlerin önem sıralaması Gini indeksi ile hesaplanarak görsel oluşturulabilir. (James ve diğ, 2013).

3.4 Artırma (Boosting)

Karar ağaçlarından elde edilen sonuçları iyileştirmenin diğer bir yolu da Artırma (Boosting) yöntemidir. Torbalama yöntemine çok benzemekle beraber ağaçların dallanması açısından farklılık gösterir. Torbalama yönteminde veri seti kopyalanarak her bir kopya için farklı karar ağaçları oluşturulur. Daha sonra bu ağaçlar birleştirildiğinde bir tahmin modeli ortaya çıkar. Bu ağaçların oluşum sırasında birbirlerinden etkilenmezler. Ancak, Artırma (boosting) yönteminde ağaçlar daha önce büyüyen ağaçlardan elde ettikleri bilgileri kullanarak birbirleriyle etkileşim (sequential) halinde büyürler. Diğer bir deyişle, orman hatalarından ders alarak büyür. Artırma yönteminde bootstrap örnekleme kullanılmaz, bunun yerine her bir ağaç sonradan değiştirilmiş veri seti üzerine köklerini salar (James ve diğ, 2013).

3.5 Gradyan Artırma (Gradient Boosting)

Gradyan artırma modeli karar ağaçlarının tahminini iyileştirmek için geliştirilmiş bir modeldir. Rastgele Orman ağaçları modelinden daha önce geliştirilmiştir (Friedman, 2001). Karar ağaçlarında olduğu gibi her bir veri seti için bir ağaç oluştururlar. Bir sonraki ağacı oluşturmadan önce daha önce oluşturulan ağaçların hata oranları göz önünde bulundurulur. Artırma modelinde olduğu gibi diğer ağaçlar ile etkileşim halinde ilerlenerek model kurulur. En büyük amacı, oluşan hata oranını en aza indirmeye çalışarak ağaç oluşturmaya devam etmesidir. Gradyan Artırma Regresyon modeli için uygulama adımları aşağıda sıralanmıştır;

1. Regresyon ağacı oluşturulur.
2. Her bir ağacın yaprağı için hata oranı hesaplanır.
3. 2. adımda hesaplanan hata oranları yeni gözlem verileri olarak kullanılır.
4. İkinci ağaç 2. adımdaki hata oranlarını en aza indirmeyi hedefler
5. 2. ve 4. adımlar tekrar edilir.

3.6 Ekstrem Gradyan Artırma (Extreme Gradient Boost, XGBoost)

Ekstrem Gradyan Artırma (XGBoost) algoritması sonuçların performansı ve çalışma hızı açısından gradyan artırma algoritmasının geliştirilmiş halidir. Makine öğrenmesi problemlerini çözmek için kullanılan bir algoritma paketidir. Milyonlarca örneklemi az miktarda kaynak kullanarak çözebilen verimli bir algoritmadır. Tianqi Chen tarafından 2014 yılında geliştirilip açık kaynak kodlu makine öğrenmesi algoritmalarının bulunduğu DMLC (Distributed Machine Learning Community) adlı kütüphaneye eklenmiştir (Chen ve Guestrin, 2016).

XGBoost algoritması milyonlarca veri setini kısa zamanda ve diğer algoritmalara göre daha az işlemci gücü ve geçici belleğe (ram) ihtiyaç duyarak analiz edebildiği için bir çok makine öğrenmesi ve veri analizi problemlerinin çözümünde kullanılmıştır. Örneğin, Avrupa Nükleer Araştırma Merkezi (CERN), Büyük Hadron Çarpıştırıcısı'nın ürettiği yıllık 3 petabayt (PB, 1024 terabayt) veriyi analiz etmek için XGBoost algoritmasını kullanmıştır (Sundaram, 2018).

3.7 Light Gradyan Artırma (Light Gradient Boosting, LightGBM)

Light Gradyan Artırma (LightGBM) algoritması karar ağaçlarını kullanan farklı bir gradyan artırma modelidir. Diğer karar ağaçları analiz sürecinde dikey bir şekilde dallanma yaparken, LightGBM yapraklar üzerinden yatay dallanma yoluna gider. En büyük delta değerine sahip yaprak üzerinden karar ağacı dallanmaya devam eder. LightGBM diğer gradyan artırma modelleri ile karşılaştırıldığında hızlı işlem yapması, daha büyük verileri işleyebilmesi, daha az bilgisayar hafızasına (ram) duyarak analizleri gerçekleştirebilmesi gibi avantajları vardır. LightGBM modelinin diğer bir avantajı ise kategorik değişkenler bulunan verilerin sayısal olarak analiz edilebilmesi için one-hot encoding gibi işlemlere gerek duymamasıdır. LightGBM, veri seti üzerinde diğer modellere nispetle daha kolay aşırı uyum (overfitting) sağladığından büyük veri setlerinde daha iyi sonuçlar vermektedir. Farklı veri setleri üzerinde yapılan çalışmalarda LightGBM modelinin veriyi öğrenme sürecinin diğer modellere göre 20 kat daha hızlı olduğu sonucuna ulaşılmıştır (Ke ve diğ, 2017).

3.8 Ekstra Rastgele Orman (Extra Tree Regressor)

Rastgele Orman metodunun farklı bir versiyonudur. Rastgele orman metodunda olduğu gibi veri setinin kopyaları kullanılarak model eğitilir, ancak düğümlerin dallara ayrılma aşamasında karar kriteri kullanarak optimum ayrılmayı yapmak yerine rastgele dallanma yolun gider. Bu fikir, bazı veri analizi problemlerinin çözümünde karmaşıklığı ve işlem yükünü azaltmasına rağmen yüksek gürültü barındıran büyük veri setlerinin analizinde performansı düşüktür. İstatistiksel açıdan değerlendirildiğinde bu metod genellikle bias artışına sebep olurken varyansı düşürür (Geurts ve diğ, 2006).

3.9 Adaboost

Topluluk (ensemble) algoritmalarından biri olan bu yöntemde ise torbalama yönteminde olduğu gibi veri seti rastgele parçalara bölünür. Her bir parça farklı bir algoritma ile öğrenilir ve test edilir. Test sonucunda yanlış tahmin edilen veri parçaları farklı bir algoritma ile tekrar öğrenilerek tekrar test edilir. Sistem her defasında doğru tahmin ettiği sonuçları alarak yoluna devam eder (Dietterich, 2000).

3.10 Düzenlileştirme (Regularization)

Makine öğrenmesi algoritmaları ile problem çözümünde karşılaşılan sorunlardan biri makinenin veri setini öğrenmeyi bırakıp ezberleyerek aşırı uyum (overfitting) sağlamasıdır. Bunun önüne geçmek için regresyon modeli bir takım teknikler ile geliştirilir. Bu teknikler arasında literatürde en çok kullanılanlara örnek olarak Ridge, Lasso ve Elastic Net verilebilir (Alpaydin ve Bach, 2014).

3.10.1 Çıkıntı (ridge regression)

Bu teknik ile doğrusal regresyondaki maliyet fonksiyonuna beta katsayılarını sıfıra doğru düşürerek model ağırlıklarının düşmesini sağlayan bir düzenlileştirme terimi eklenir. Bu modeli kullanmadan önce verinin ölçeklendirilmesi gerekebilir.

3.10.2 Lasso (reast absolute shrinkage and selection operator regression)

Bu teknik ile Çıkıntı tekniğinde olduğu gibi maliyet fonksiyonuna farklı bir düzenlileştirme terimi eklenir. Lasso tekniğinin farkı makine öğrenmesi modelindeki

bazı özniteliklerin ağırlıklarını düşürmek hatta sıfır yapmaktır. Çıktı olarak öznetileklerden sadece birkaç tanesinin ağırlığının sıfırdan farklı olduğu ayrık (sparse) bir model verir.

3.10.3 Elastic net

Bu teknikte düzenlileştirme terimi olarak Çıkıntı ve Lasso tekniklerinin bir karışımı olan terim maliyet fonksiyonuna eklenir. Düzenlileştirme teknikleri arasında Çıkıntı tekniğini tüm modeller için kullanılırken, az sayıda özniteliklerin model için daha verimli olacağı düşünülüyorsa Lasso ve Elastic net kullanır.

3.11 Basit Bayes (Naive Bayes) Sınıflandırıcı

Basit Bayes (Naive Bayes) sınıflandırıcısı değişkenlerin birbirlerinden tamamen bağımsız olduğu varsayılan basit ama etkili bir sınıflandırma tekniğidir. Sınıflandırma yapılması istenen verilerin birbirlerinden bağımsız olmaları mümkün değil gibi gözükse de bu varsayımın kullanılarak yapılan araştırmaların iyi sonuçlar verdiği ortaya çıkmıştır (Rish, 2001). Uygulaması basit ve hızlı bir teknik olduğu için sınıflandırma problemlerinde ilk tercih edilen yöntemlerden biridir.

$$P(Cj/X) = P(Cj)\ P(X/Cj) / P(X) \qquad (3.1)$$

P(X) = Herhangi bir örneğin X setinden olma olasılığı
P(Cj) = Örneğin j sınıfından olma olasılığı
P(X/Cj) = J sınıfından olduğu bilinen bir örneği X setinden olma olasılığı
P(Cj/X) = X setinden olduğu bilinen bir örneğin J sınıfından olma olasılığı

P(X) değeri bütün değerler için sabit olduğuna göre bu algoritmanın amacı P(Cj) P(X/Cj) değerini maksimize etmektir.

Örneğin; C=2 için eğer P(C=1 / X) > 0.5 ise veri birinci sınıfa dahil edilir. Eğer ihtimal 0.5 değerinden küçük ise diğer sınıfa dahil edilir.

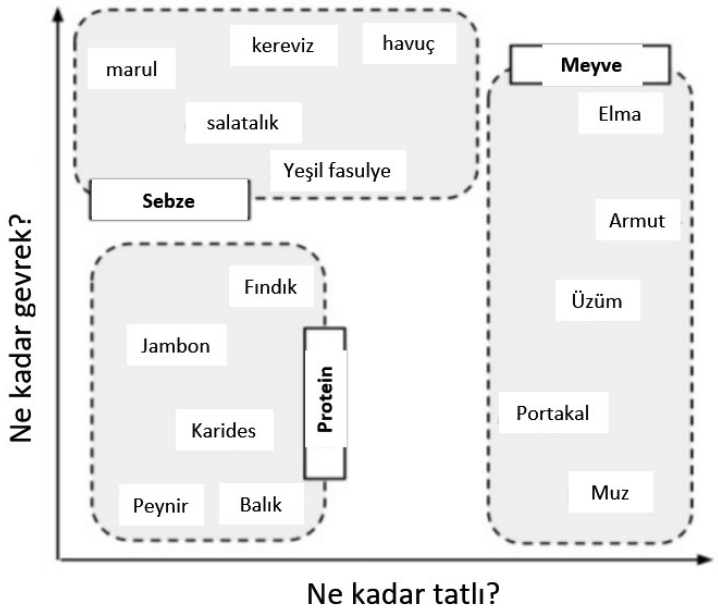
3.12 K-En Yakın Komşu (K-Nearest Neighbor(KNN))

K-En Yakın Komşu (KNN) algoritması etiketleme bilgisi bulunmayan gözlemlerin birbirlerine yakınlık derecesi göz önünde bulundurularak yapılan sınıflandırma işlemidir. Basit bir işlem gibi gözükse de makine öğrenmesi sınıflandırma problemleri

için oldukça kullanışlı bir yöntemdir. İlk olarak Fix ve Hodges tarafından literatüre kazandırılmıştır (Fix ve Hodges, 1951). Yüz tanıma teknolojisi, kullanıcının seyrettiği ve sevdiği filmlere göre yeni film tavsiyesi sunabilen sistemler, genetik üzerinde spesifik protein yapılarının tespit edilmesi gibi çalışmalar bu algoritmanın kullanım alanlarına örnek olarak verilebilir. Bu algoritmanın güçlü yanlarından bazıları; uygulamanın basit ve verimli olması, algoritmanın verinin dağılımı hakkında varsayım yapmaması ve eğitim sürecinin hızlı olmasıdır. Zayıf yönleri ise sınıflandırmanın yavaş yapılması, yüksek bilgisayar hafızası gerektirmesi ve veri setindeki kayıp değerlerin (missing value) işlem yükünü artırmasıdır.

Bu algoritmanın uygulama aşamasında öncelikle eğitim verisi kullanılarak sistem eğitilir. Eğitim setinde sınıflandırma bilgisi bulunan veriler mevcuttur. Eğitim seti verildikten sonra kullanıcı tarafından bir k değeri tespit edilir. Sistem öğrendiği eğitim setindeki sınıfları göz önüne alarak yeni verilen gözlem verilerinden birbirine en yakın bulunan k tane komşu veriyi sınıflandırır. Verilerin birbirine olan yakınlığı ölçmek için Öklid uzaklığı, Jaccard Distance, Simple Matching Distance, Manhattan Distance gibi yöntemler kullanılır.

Örneğin; yiyeceklerin bulunduğu 15 gözlem verisi tatlı ve gevrek oluşlarına göre 1-10 arasında değer verip iki boyutlu düzleme yerleştirildiğinde Şekil 3.4 elde edilir. Bu yiyecekler k=5 değeri için tatlarının yakınlık derecesine göre sınıflandırıldığında 3 farklı sınıf oluşur. Tadı gevrek olup tatlı olmayanlara sebze grubu, gevrek olmayıp tatlı da olmayanlara protein grubu, tatlı olup gevrek olan ya da gevrek olmayanlara da meyve grubu adı verilebilir (Lantz, 2013).



Şekil 3.4 : Yiyeceklerin tatlarına göre sınıflandırılması örneği (Lantz, 2013).

KNN algoritması için kaç komşu (k) kullanılacağının belirlenmesi, modelin verileri nasıl sınıflandıracağını etkiler. Makine öğrenmesi algoritmaları ile çözülen bir problemin sonuçlarını etkileyen en büyük faktörler gürültü(noisy), varyans ve dizgesel hata(bias)dır. Modelin eğitim veri seti ile aşırı uyum (overfitting) ya da yetersiz uyum (underfitting) halinde olması bias ve varyans arasındaki dengeyi (bias-variance tradeoff) kurmak için önemlidir. Büyük bir k değeri seçmek gürültü verilerinin varyans etkisini düşürürken, araştırmacıyı yanıltarak (bias) küçük ama önemli detayları gözden kaçırmasına sebep olabilir (James ve diğ, 2013). Literatürde k değeri genellikle 3 ile 10 arasında ya da eğitim veri seti sayısının karekökü olarak alınmaktadır (Lantz, 2013).

3.13 Zaman Serisi Analizi

Makine öğrenmesi algoritmaları bir çok farklı alanda tahminleme yapmak üzere kullanılmaktadır. Yapılan bu çalışmalarda kullanılan veri setleri periyodik ya da zamana bağlı bir şekilde toplanmış verilerden oluştuğunda bu analize zaman serisi analizi adı verilir (Grim ve Gradvohl, 2018). Zaman serisi ile öngörümleme problemleri denetimli makine öğrenmesi yaklaşımı ile çözülebilir (Bontempi ve diğ, 2012).

3.13.1 Üstel düzeltme (exponential smoothing, ES)

Basit Üstel Düzeltme ile bir sonraki zamanın tahmini yapılırken veri üzerindeki son değişim ve önceki tahmine ait hatanın bir kısmı kullanılır. Bu yöntem ile geçmiş gözlem değerlerine daha düşük ağırlık verilirken son döneme yakın gözlem verilerine daha büyük ağırlık verilerek tahmin yapmaktadır. Gözlemlere verilen ağırlık üstel biçimde geçmiş gözlemlere doğru azalmaktadır (Hyndman ve Athanasopoulos, 2018). Üstel Düzeltme yönteminin basit, ikili ve üçlü üstel düzeltme olmak üzere üç farklı çeşidi vardır.

Basit Üstel Düzeltme yönteminde kullanılan veri seti tek değişkenden oluşur ve trend ya da mevsimsellik barındırmaz. Denkleminde tek bir parametre bulunur; ortalama düzey düzleştirme katsayısı (alpha). Bu parametre geçmiş verilerin yapılacak tahmin üzerindeki ağırlığını belirler. 1 değerine yakın olan parametre hızlı öğrenmeyi (son döneme yakın veriler tahmini daha çok etkiler); 0 değerine yakın olan parametre ise yavaş öğrenmeyi (geçmiş verilerin tahmin üzerindeki etkisi daha büyüktür) ifade eder (Shmueli ve Lichtendahl, 2016).

İkili üstel düzeltme yöntemi basit üstel düzeltme yönteminin bir uzantısı olup trend içeren tek değişkenli veri seti üzerinde kullanılmaktadır. Ortalama düzey düzleştirme katsayısı ile beraber trenddeki değişkenliğin tahmin üzerine etkisini kontrol eden eğim düzleştirme katsayısı (beta) bulunur. Bu metoddaki trendin değişimi iki farklı şekilde olur; doğrusal ve üstel. Doğrusal değişime sahip olan trende Additive Trend, üstel değişime sahip olan trende Multiplicative Trend adı verilir. Additive Trend bulunan veri seti ile yapılan İkili üstel düzeltme yöntemi Holt's doğrusal trend modeli olarak da bilinir.

Üçlü üstel düzeltme yöntemi ise trend ve mevsimsellik içeren veri setinin analizi için kullanılır. Holt-Winters Üstel Düzeltme yöntemi olarak da adlandırılır. Alfa ve beta katsayıları ile beraber mevsimselliğin etkisini kontrol eden gama katsayısı da kullanılır (Shmueli ve Lichtendahl, 2016).

3.13.2 Bütünleşik otoregresif hareketli ortalama (autoregressive integrated moving average, ARIMA)

ARIMA modeli zaman serisi veri setini analiz etmek için kullanılan yöntemlerden biridir. Bu yöntem ile durağan olmayan zaman serileri fark alma işlemiyle durağan hale (stationary series) dönüştürülür. İstatistiksel özellikleri zaman içerisinde

değişmeyen zaman serisi verilerine durağan adı verilmektedir. ARIMA notasyonu AR, I ve MA modellerinin birleşimidir. AR modeli Otokorelasyon (Autoregressive) mertebesidir ve geçmiş verilerin ağırlıklı hareketli ortalamasını göstermektedir. I modeli bütünleşmeyi ifade etmekte olup doğrusal veya polinomiyal trendi göstermektedir. MA modeli ise hareketli ortalama (Moving Average) modeli olup geçmiş hataların ağırlıklı hareketli ortalamasını göstermektedir. Bu modelin genel gösterimi ARIMA (p, d, q) şeklindedir. Buradaki p otokorelasyon (AR), d fark alma derece ve q ise hareketli ortalama (MA) modelinin derecesini göstermektedir. Zaman serileri mevsimsellik içeriyor ise Mevsimsel Bütünleşik Otoregresif Hareketli Ortalama (Seasonal Autoregressive Integrated Moving Average, SARIMA) adı verilen model kullanılmaktadır (Chou ve Tran, 2018).

3.14 Prophet Tahmin Modeli

Prophet, Python ve R programlama dillerinde zaman serisi analizi ile tahmin yapmayı sağlayan açık kaynak kodlu bir yazılımdır. Prophet modeli Facebook Veri Bilimi merkezi tarafından yayımlanmıştır. Prophet modelinin dökümantasyonu ve nasıl uygulanacağına dair bilgiler Taylor ve Letham tarafından paylaşılmıştır (2018). Prophet modeli ile veri seti üzerindeki haftalık veya yıllık trend ve mevsimsellik analiz edilir. Tatil ve özel günlerin tarihleri modele önceden öğretilebilir. Prophet algoritması basit kullanımı ile veri analizi alanında uzman olmayan kimseler tarafından tahminleme işleminde kullanılabileceği iddia edilmektedir. Model üzerinde kullanılacak olan veri seti iki en az iki sütuna ihtiyaç duymaktadır. Birinci sütunda "ds" yani zaman serileri, yıllar gösterilir. İkinci sütun ise "y" yıllara karşılık gelen gözlem değerlerini ifade eder. Bu sebeple, mevsimsel zaman serileri kolaylıkla analiz edilebilmektedir. Bu kolaylıklardan biri, model kurulurken veri setindeki mevsimselliğin yıllık, haftalık ya da günlük olduğu bilgisi modele gösterilebilir (Taylor ve Letham, 2018).

4. NÜFUS PROJEKSİYONU

4.1 Nüfus Projeksiyonu Kavramı

Ülkelerin veya toplulukların geleceğe yönelik alacağı kararlarda oluşturdukları nüfus projeksiyonlarının rolü büyüktür. Nüfusun zaman içerisindeki değişimi sosyal, ekonomik, çevresel ve politik alanlarda alınacak kararları önemli ölçüde etkilemektedir. Bu sebeple, nüfus projeksiyonları ile herhangi bir konuda alternatif senaryolar üretilerek yaşanabilecek problemler önceden tahmin edilmeye çalışır. Böylece, yaşanabilecek problemleri önleyici proaktif tedbirlerin alınması için tavsiyeler verilebilir veya aksiyonlar alınabilir. Nüfus projeksiyonları sayesinde herhangi bir değişkenin zaman içerisindeki etkisi ölçülebilir. Örneğin; doğum oranlarındaki %25 düşüş ülkenin nüfusunu 20 yıl içerisinde nasıl değiştirir? Sigara sebebiyle yaşanan ölümlerin %50 azaltılması ülke nüfusunu nasıl etkiler ? Bir bölgeye 2000 işçi kapasiteli yeni bir fabrika açılması o bölge üzerinde nasıl değişikliklere sebep olur ?

Üretilen nüfus projeksiyonları, diğer alanlarda da projeksiyonlar üretilmesine olanak sağlar (aile planlaması, gelir düzeyi, işsizlik seviyesi vb.). Ulusal ve yerel nüfus projeksiyonlarının birçok kullanım alanı vardır. Örneğin; gelecekteki sağlık harcamalarının tahmin edilmesi (Lee ve Miller, 2002), ortalama yaşam süresinin değişimi (Olshansky ve diğ, 2005), bölgedeki özel okul ihtiyacının öngörülmesi (Swanson ve diğ, 1998), itfaiye istasyonu ihtiyacının tahmin edilmesi (Tayman ve diğ, 1997), işletmelerde ürün talep tahmini yapılması (Thomas, 1997), populasyonun işgücü piyasası üzerindeki etkisi (Börsch-Supan, 2003), diyabet hastalığının nüfus üzerindeki etkisi (Boyle ve diğ, 2010), nüfus değişiminin iklim, su kaynakları ve doğa üzerindeki etkisi (Arnell, 2004; Goudie, 2018), göçlerin Türkiye nüfusu üzerindeki etkisi (Ediev ve Yüceşahin, 2016; Sirkeci, 2017).

Nüfus projeksiyonlarının doğruluğu ve güvenilirliği için girdi olarak kullanılan verilerin sağlıklı toplanması gerekir. Nüfusun gelecekteki değişimi geçmişe bağlıdır.

Geçmişte yaşanan değişimler ve trendler bize gelecek hakkında ipuçları verir. Nüfus, doğurganlık, ölüm, göç oranları geçmişten bu zamana hassas, doğru ve güvenilir bir biçimde toplanıp arşivlenirse üretilecek olan nüfus projeksiyonları daha başarılı olacaktır (Swanson ve diğ, 2004).

Nüfus projeksiyonlarının tahmin edilmesi için birçok yöntem bulunmaktadır. Bunlardan biri geçmiş verilerin trendine bağlı olarak geleceği tahmin eden trend ekstrapolasyon (trend extrapolation) yöntemidir. Ayrıca, doğrusal büyüme (linear growth), üstel büyüme (exponential growth), korelasyon büyümesi (correlation growth) gibi matematiksel modeller de kullanılmaktadır. Literatürde en çok kullanılan yöntem ise nüfustaki her bir değişkeni kendi içerisinde ele alan kuşak bileşenleri yöntemidir (Cohort Component Model)(Smith ve diğ, 2006). Bu çalışmada kuşak bileşenleri yöntemi ile nüfus projeksiyonu üretilmiştir.

4.2 Kuşak Bileşenleri Yöntemi (Cohort Component Model, CCM)

Kuşak Bileşenleri yöntemi ülke bazında nüfus projeksiyonları üretmek için en çok kullanılan modellerden biridir. İlk olarak Cannan tarafında 1895 yılında kullanılmış, daha sonra Bowley ve Whelpton tarafından geliştirilerek günümüzdeki halini almıştır (Cannan, 1895; Bowley, 1924; Whelpton, 1928). Bir çok farklı uygulaması bulunmasıyla birlikte ana çalışma prensibi değişmemiştir (Swanson ve diğ, 2004).

Kuşak Bileşenleri yöntemi ile nüfus, yaş grupları ve cinsiyete göre kuşaklara ayrılır. Her bir kuşak için doğurganlık hızı, ölüm oranı ve göç oranı belirlenir. Kuşak bileşenleri yöntemi nüfusun, doğurganlık, ölüm ve göç oranlarının geçmiş zaman içindeki davranışlarını hesaba katarak gelecekteki yıllar için nüfus projeksiyonu üretir. Geriye dönük nüfus verisi bulunmayan yılların projeksiyonu için de kullanılabilir. Girdi olarak kullanılan verilerin güvenilir olması üretilen nüfus projeksiyonlarının tahmin doğruluğunu artırır.

4.3 Kuşak Bileşenleri Yönteminde İhtiyaç Duyulan Temel Veriler

4.3.1 Nüfus

Kuşak bileşenleri yöntemi ile nüfus projeksiyonu üretmek için öncelikle bir başlangıç yılı belirlenir. Başlangıç yılında her yaş için cinsiyete göre ayrılmış nüfus sayımlarına ihtiyaç duyulur. 0-79 yaşları arası yaşlar birer yaş, 80 ve üstü yaşlar ise bir grup olarak

kullanılır. Ayrıca, 0-4 ile 75-79 arasında 5'erli yaş gruplarına göre ayrılmış nüfus istatistikleri de kullanılmaktadır. Yaş grupları ırk, etnik köken gibi nüfus projeksiyonunun amacına göre farklı gruplara da ayrılabilir. Nüfus istatistikleri, TUİK gibi devletlerin resmi istatistik kurumlarından ya da dünya genelinde istatistikler üreten Population Bureau, World Bank gibi kurumlardan temin edilebilir.

4.3.2 Doğurganlık oranları

Doğurganlık oranları toplam doğurganlık oranı (TDO) (Total Fertility Rate, TFR) ve yaşa özel doğurganlık oranı (YÖDO) olmak üzere iki kısımdır. Toplam doğurganlık oranı bir kadının yaşamı boyunca vereceği çocuk sayısının ortalamasıdır. Yaşa özel doğurganlık oranı ise doğum yapma potansiyeli olan her 1000 kadın için canlı olarak dünyaya gelen çocuk sayısını ifade eder. Yaşa özel doğurganlık oranlarını her bir yaş için toplamak zor olacağından yaş grupları halinde kullanılır. Kuşak bileşenleri yöntemi doğurganlık oranlarını kullanarak her yıl için nüfusa eklenecek olan çocuk sayısını elde eder. Doğurganlık oranları nüfus projeksiyonu için en önemli etkendir (Swanson ve diğ, 2004).

4.3.3 Doğumdaki cinsiyet oranı

Doğumdaki cinsiyet oranı dünyaya gelen her 100 kız çocuğu için erkek çocuğu oranını verir. Nüfus projeksiyonunda bir sonraki yıl için nüfusa eklenecek olan sayının cinsiyeti önemlidir. Nüfus projeksiyonları 50 sene sonrasını tahmin edebileceği için dünyaya gelen kız çocuğu ve bunların arasından doğum yapma yaşına gelenlerin istatistiği önemlidir. Doğumdaki cinsiyet oranı dünyada genelinde toplanan istatistiklere göre 103-107 arasındadır.

4.3.4 Ölüm oranları

Nüfusu önemli ölçüde etkileyen diğer bir faktör ölüm oranlarıdır. Çocuk ölümleri ve yetişkin ölümleri birbirlerinden farklı bir istatistiksel yapı oluşturmaktadır. Ülkelerin gelişmişlik yapısına göre yetişkinlerin yaşam süresi uzamaktadır. Model Hayat Tablosu adı verilen istatistikler ile çocuk ölümleri, her yaş grubu için hayatta kalma oranı, doğumda beklenen yaşam süresi verilerine ulaşılabilir. Örneğin; Türkiye'de yeni doğan bir bireyin ortalama yaşam süresi 1980 yılında 58 iken, 2010 yılında bu rakam 74'e yükselmiştir.

4.3.5 Göçler

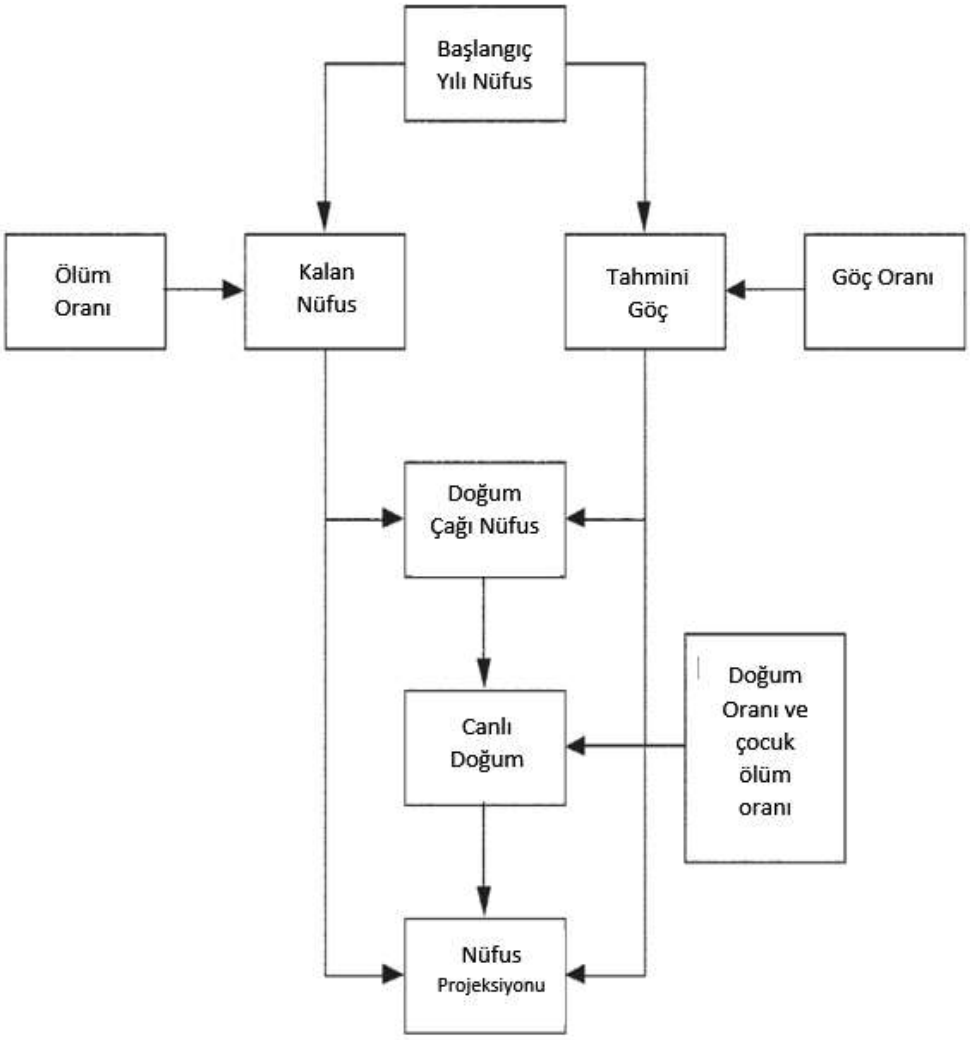
Nüfusu etkileyen diğer bir faktör alınan ve verilen göç oranlarıdır. Nüfus projeksiyonu üretilecek olan bölge veya ülke için dışarıdan alınan göç oranı ve dışarıya verilen göç oranı istatistiklerine ihtiyaç duyulur. Bu istatistikler her ülke için yeterli seviyede toplanmamış olabilir. Özellikle gelişmişlik seviyesi düşük ve dışarıya göç veren ülkelerin istatistiklerine ulaşılamadığı durumlarda, göç edenlerin verilerine ulaşmak için göç ettikleri ülkelerin kayıtlarından yararlanılablir.

4.4 Nüfus Projeksiyonun Üretilmesi

Nüfus projeksiyonunu üretmek için öncelikle bir başlangıç yılı belirlenir. Başlangıç yılına kadar hayatta kalan kişi sayısı kadın-erkek cinsiyet gruplarına göre ayrılarak her bir yaş için nüfus verisi toplanır. İkinci adımda projeksiyon süreci boyunca her yaş grubu için bölgeye alınan ve bölgeden verilen göçler hesaba katılır. Böylece başlangıç yılı için yaşayan insan sayısı belirlenmiş olur.

Üçüncü adımda bir sonraki yıla kadar dünyaya gelecek çocuk sayısı belirlenir. Bunun için yaşa özel doğurganlık oranları kullanılarak her yaş grubu için dünyaya gelecek çocuk sayısı belirlenir ve cinsiyetlerine göre ayrılarak nüfusa eklenir. Son adımda ise bir sonraki yıla kadar meydana gelecek ölümler, cinsiyetlerine göre ayrılarak hesaba katılır. Hastalık, kaza vb. sebepler ile her yaş grubu için meydana gelecek ölüm oranları tespit edilerek toplam nüfustan azaltılır. Böylece başlangıç yılından bir sonraki yıl için hayatta kalan nüfusa ulaşılır.

Bu işlemler hedeflenen projeksiyon yılına ulaşana kadar her yıl için tekrar edilir. Her yıla özel önceden tahmin edilmiş doğum, ölüm ve göç oranları kullanılır. Şekil 4.1'de nüfus projeksiyonu üretilmesi için takip edilen adımlar özetlenmiştir (Smith ve diğ, 2006).



Şekil 4.1 : Kuşak bileşenleri yöntemi ile nüfus projeksiyonu üretim adımları (Smith ve diğ, 2006).

5. UYGULAMA

5.1 Aşamalar ve Metodoloji

Bu çalışmanın amacı farklı makine öğrenmesi algoritmaları ile toplam nüfus tahmini yapmaktır. Bunun için 6 farklı makine öğrenmesi algoritmaları seçilerek 2017 yılı toplam nüfus değeri tahmin edilecektir. 2017 yılının tahmin edilmesinin sebebi, yapılan tahminlerin gerçek nüfus değeri ile karşılaştırarak modellerin tahmin başarısının ortaya konulmasıdır. Böylece hangi makine öğrenmesi algoritmasının nüfus tahmini üzerinde daha iyi sonuç verdiği gözlemlenecektir. Bu doğrultuda, çalışmanın amacına yönelik farklı makine öğrenmesi algoritmaları seçilmesi gerekmektedir. Bütün problemlere uygulanabilen tek bir model yoktur. Her bir algoritmanın birbirine göre üstün ya da zayıf yönleri bulunmaktadır. Bu sebeple farklı çalışma prensiplerine sahip algoritmalar üzerinden nüfus tahmini yapılacaktır.

Karar Ağaçları (Classification and Regression Tree, CART) algoritma ailesinden LightGBM algoritması seçilmiştir. LightGBM algoritmasının seçilmesinin sebebi diğer karar ağaçları algoritmalarına göre daha az işlem gücüne ihtiyaç duyarak daha hızlı işlem yapabilmesidir. Bununla birlikte, LightGBM algoritması kategorik değişkenler bulunan verilerin sayısal olarak analiz edilebilmesi için one-hot encoding gibi işlemlere gerek duymamaktadır. Bu özellik analiz süresini kısaltmaktadır. LightGBM algoritması iyi sonuç vermediği takdirde, benzer çalışma sistemine sahip olan diğer karar ağaçları algoritmalarının farklı bir sonuç doğurması beklenmemektedir.

Literatürde zaman serileri problemi için tavsiye edilen diğer modeller ise Doğrusal Regresyon ve Ridge Regresyon algoritmalarıdır. Nüfus tahmininde kullanılan verilerin zamana dayalı olması, belirli sebepler ile artış ya da azalma göstermesi, zaman serisi verilerinin trend içermesi ve çalışmanın amacının bir sonraki yılı tahmin etmek olması sebebiyle regresyon algoritmalarının bu çalışmada daha iyi sonuç verebileceği

öngörülmektedir. Düzenlileştirme modelleri arasından Ridge modeli Lasso modeline göre daha iyi sonuç verdiği için tercih edilmiştir.

Literatürde trend içeren ve zamana bağlı toplanan veri seti için tavsiye edilen diğer bir algoritma grubu ise zaman serisi analizidir (Grim ve Gradvohl, 2018). Bu analiz grubu içerisinden Holt-Winters, ARIMA ve Prophet adı verilen makine öğrenmesi algoritmaları kullanılarak nüfus tahmini yapılacaktır. Bu algoritmalar çalışma prensibi gereği tek bir değişken üzerinden tahmin yapabildiği için, 1960-2017 yılları arasındaki toplam nüfus değerleri kullanılarak 2017 yılı toplam nüfusu tahmin edilecektir. Holt-Winters modeli trend ve mevsimsellik içeren veri setinin analizi için kullanılmaktadır (Shmueli ve Lichtendahl, 2016). ARIMA ise durağan olmayan zaman serilerini fark alma işlemiyle durağan hale dönüştürerek analiz yapmaktadır. ARIMA modeli için p, d, ve q değerleri sırasıyla 1,1 ve 0 olarak seçilmiştir. Bunun sebebi veri setindeki değerlerin her bir yıl için bir defa toplanması, veri setini durağan hale getirmek için 1 yıllık fark derecesinin alınması ve hareketsel ortalama uygulanmamasıdır.

Bu çalışmanın uygulama kısmı için ihtiyaç duyulan veriler Dünya Bankası tarafından temin edilmiştir. 262 farklı ülke için 1960-2017 yılları arasındaki demografik göstergeler ile 6 farklı makine öğrenmesi algoritması uygulanacaktır. Temin edilen veri seti üzerinden hedef değişken olarak 2017 yılı toplam nüfusu olarak belirlenmiştir. Veri üzerinde makine öğrenmesi algoritmaları eğitilmeden önce öznitelik mühendisliği (feature engineering) kapsamında veri ön hazırlığı yapılmıştır. Veri seti hazırlandıktan ve model kurulduktan sonra, veri seti makine öğrenmesi algoritmaları ile eğitilerek algoritmaların performansları ölçülüp, sonuçlar karşılaştırılmıştır. Makine öğrenmesi algoritmaları bütün veri seti ile eğitildikten sonra, eğitilen modeller kullanılarak Türkiye'nin verileri üzerinden 2017 yılı toplam Türkiye nüfusu tahmin edilecektir. 6 farklı makine öğrenmesi algoritması ile tahmin edilen 2017 yılı Türkiye nüfusu, farklı bir nüfus tahmin modeli olan ve dünya genelinde ülkelerin nüfus birimleri tarafından sıkça kullanılan kuşak bileşenleri yöntemi ile de tahmin edilerek sonuçlar karşılaştırılacaktır.

5.2 Problem Tanımı

Dünya üzerinde bir çok farklı ülke bulunmaktadır. Bu ülkeler varlığını devam ettirebilmek için içerisinde yaşayan insanları yöneten bir idareye ihtiyaç duyarlar. Bu idare ya da devlet ülkeyi yönetebilmek ve ülke hakkında kararlar alabilmek için

öncelikle yönetmekle sorumlu olduğu insan sayısı yani ülkenin nüfusu hakkında fikir sahibi olmalıdır. Bu sebeple ülke genelinde belirli aralıklar ile nüfus sayımı yapılması ihtiyacı doğmuştur. Böylece ülke hakkında eğitim, sağlık, yatırım, barınma gibi ulusal ihtiyaçlar planlanabilecektir. Ülkenin geleceğine dair sosyal, ekonomik, politik ve çevresel kararlar daha sağlıklı alınabilecektir.

Nüfus sayımı, ülkede yaşayan bütün insanlar hakkında belirli zaman aralıkları içerisinde demografik, sosyal ve ekonomik verilerin toplanması, derlenmesi, analiz edilmesi ve yayınlanması olarak tanımlanmıştır (Bamgbose 2009). Yapılan nüfus sayımları neticesinde elde edilen veriler ülkenin gelecekteki nüfusunu tahmin etmek için kullanılır. Böylece ülkenin geleceğine dair kararlar alınabilir. Bu sebeple, toplanan verilerin gerçeği yansıtması önemlidir.

Nüfus sayımının maliyeti ülke için çok yüksektir. Bu sebeple genellikle on yılda bir ülke genelinde nüfus sayımları yapılmaktadır. Bununla beraber, yapılan nüfus sayımları neticesinde ülkenin toplam nüfusu hakkında net verilere ulaşmak oldukça meşakkatlidir. Bu da ülkenin nüfusunu tahmin etmeyi daha çok zorlaştırmaktadır. Bu sebeple, kimi zaman tutarsız sonuçlar doğuran ve milyonlarca lira harcanarak ülke için yüksek giderlere mal olan nüfus tahmini ülke yönetimi için önemli bir problemdir. Bu problemin çözülmesi ya da minimize edilmesi önem arz etmektedir.

Ülkenin nüfusunu tahmin etmeyi zorlaştıran bir takım sebepler vardır. Bu sebepler şöyle sıralanabilir;

- Doğum oranları düşmesine rağmen doğum yapma çağındaki kadın sayısının artış göstermesi,
- Doğum ve ölüm oranları açısından ülke içinde bölgeler bazında çeşitlilik gözlenmesi,
- Ortalama yaşam ömrünün yükselmesi,
- Gelişmekte olan ülkelerde HIV virüsü ve benzer hastalıklar nedeniyle ölen insan sayısında artış yaşanması,
- Son dönemdeki iç ve dış göç oranlarında büyük değişim gözlenmesi.

Makine öğrenmesi algoritmaları, demografik veriler üzerindeki belirsizlikleri algılayıp, geçmiş verileri analiz ederek ülkenin nüfusunu daha iyi tahmin edebilir. Önceki verilere göre geleceği tahmin edebilecek en iyi uygulama olan makine

öğrenmesi günümüzde sıkça kullanılmaktadır. Makine öğrenmesi algoritmalarının nüfus tahmini alanında da önemli bir katkı sağlayacağı düşünülmektedir.

Makine öğrenmesi ile ülkenin nüfus tahmini için regresyon alanında tahmin başarısı yüksek 6 farklı makine öğrenmesi algoritması belirlenmiştir. 2017 yılı nüfus tahmini için her bir algoritmanın başarısı ayrı ayrı incelenerek karşılaştırılacaktır. Gerçek nüfus değeri ile karşılaştırma yaparak algoritmaların tahmin performanslarının ölçülmesi için 2017 yılı seçilmiştir. Algoritmaların performanslarının karşılaştırılması neticesinde hangi algoritmanın nüfus tahmini üzerinde daha iyi sonuç vereceği hakkında fikir sahibi olunacaktır. Böylece ülkenin nüfus tahmininde en iyi performansa sahip algoritmanın regresyon tahmininden faydalanılacaktır.

5.3 Verilerin Toplanması

Nüfus projeksiyonu üretmek için kullanılacak olan veriler, ilgili bölgenin istatistik kurumlarından temin edilebileceği gibi, dünya genelinde araştırmalar yürüten kurumlardan da elde edilebilir. Bunlara örnek olarak; Türkiye için nüfus ve hayati istatistikleri üreten Türkiye İstatistik Kurumu (TÜİK) bulunmaktadır. Dünya genelinde ise Dünya Bankası (World Bank), gelişmekte olan ülkeler için hayati istatistikler raporlayan önemli bir veri merkezidir.

5.3.1 Dünya Bankası (World Bank)

Birleşmiş milletler çatısı altında global ekonomik kalkınma ve refahın artırılmasını hedefleyen iki farklı kuruluş bulunmaktadır; IMF ve Dünya Bankası. Dünya Bankası ilk olarak 1946 yılında faaliyete geçmiştir. İlk amacı 2. Dünya Savaşında yıkıma uğrayan ülkelerin kalkınması için finansal yardım sağlamaktır. Daha sonra üçüncü dünya ülkelerinin kalkınması için faaliyetler başlatılmıştır. Dünya Bankası çatısı altında 189 üye ülke ve dünya çapında 130 farklı noktada çalışma ofisi bulunur. Gelişmekte olan ve fakir ülkelerde refahın artırılması, yoksulluğun azaltılması ve yaşam kalitesinin iyileştirilmesi için destek ve kaynak çalışmaları yürütülür. Ayrıca iklim değişikliği, göç ve salgın hastalıklar gibi küresel sorunların çözümü hedeflenmektedir. Tüm bu amaçlara ulaşmak için ülkelerden ciddi anlamda veri toplaması gerekir. Veriler Dünya Bankası'nın ofisleri aracılığı ile her ülkenin kendi istatistik kurumundan temin edilir. Ayrıca, Dünya Bankası gelişmemiş ülkeler için ülke içerisinde veri toplanması aşamasına destek sağlar. 262 farklı ülkeden toplanan

çeşitli veriler akademik çalışmalarda kullanılması için Dünya Bankası'nın internet sayfasında yayınlanır. Demografik göstergeler adı altında 1960-2017 arasındaki 1595 farklı indikatör her ülke için yayınlanmaktadır. Bu indikatörler arasında ülke ile ilgili nüfus istatistikleri, göç oranları, kalkınma göstergeleri, eğitim düzeyleri, sağlık, beslenme ve yoksulluk istatistikleri gibi çok çeşitli ve kapsamlı veriler bulunmaktadır.

Bu çalışmada ihtiyaç duyulan veriler Dünya Bankası tarafından temin edilmiştir. Dünya Bankası veri tabanında 262 ülke için 1960-2017 yılları arasındaki demografik göstergeler kullanılmıştır. Veri setinin boyutu 116 Mb olup Microsoft Excel üzerinden düzenlenebilmektedir. Veri boyutu açısından küçük veri sınıfına girmektedir. Veri setinden bir kesit Çizelge 5.1 'de gösterilmiştir.

CountryName	CountryCode	Year	SP,POP,TOTL	SP,POP,BRTH,MF	SP,DYN,CBRT,IN	SP,DYN,CDRT,IN	SP,DYN,TFRT,IN	.....
San Marino	SMR	2015	32960		43504	43472		.....
Sao Tome and Principe	STP	2015	195553	43525	34326	6806	4518	.....
Saudi Arabia	SAU	2015	31557144	43525	19921	19784	25789	.....
Senegal	SEN	2015	14976994	1036	36208	60710	30773	.....
Serbia	SRB	2015	7095383	1052	43533	43630	16803	.....
Seychelles	SYC	2015	93419	43617	17	43592	15008	.....
Sierra Leone	SLE	2015	7237025	1018	35611	130279	4561	.....
Singapore	SGP	2015	5535002	1073	43655	43681	45292	.....
Sint Maarten (Dutch part)	SXM	2015	38824			43620		.....
Slovak Republic	SVK	2015	5423801	43586	43534	43717	43556	.....
Slovenia	SVN	2015	2063531	1054	10	43625	20821	.....
Solomon Islands	SLB	2015	587482	1067	29269	4852	3907	.....
Somalia	SOM	2015	13908129	43525	43661	116290	6365	.....
South Africa	ZAF	2015	55291225	43525	21290	10102	2485	.....
South Sudan	SSD	2015	11882136	43556	36315	11244	4938	.....
Spain	ESP	2015	46444832	1064	9	43474	12055	.....
Sri Lanka	LKA	2015	20966000	1042	43631	6814	2063	.....
St. Kitts and Nevis	KNA	2015	54288					.....
St. Lucia	LCA	2015	177206	43525	12239	7472	14709	.....
St. Martin (French part)	MAF	2015	31754		43570	43500	29587	.....
St. Vincent and the Grenadines	VCT	2015	109455	43525	16	7115	1953	.....
Sudan	SDN	2015	38647803	43556	33315	7522	4595	.....
Suriname	SUR	2015	553208	1072	18448	7254	2396	.....
.....	....	....	.....	.....	.....	.....	.....	.....

Çizelge 5.1 : Veri setinden bir kesit.

Bu veri setinde ülke hakkında sosyal, ekonomik, demografik bilgiler içeren 1595 farklı gösterge bulunmaktadır. Bu göstergeler ile bir ülkenin karakteristik özellikleri, gelişmişlik düzeyi, nüfus dağılımı, ekonomik durumu ve daha farklı birçok konu hakkında fikir sahibi olunabilir. Özellikle gelişmişlik düzeyi düşük olan ülkeler için bazı gösterge değerleri hakkında yeterli veri bulunmamaktadır. Veri setinde bulunan ve nüfus tahmini üzerinde etkisi olan göstergeler kısaca aşağıda özetlenmiştir. Bu değişkenler kuşak bileşenleri yöntemi ile nüfus tahmini yapmak için de kullanılmaktadır.

Nüfus: 1960-2017 arası her yıl için toplam ülke nüfusun ve cinsiyete göre dağılım bilgisi bulunmaktadır. Ayrıca 5'erli yaş grupları için cinsiyete göre ayrılmış nüfus istatistikleri bulunmaktadır.

Doğum ve Ölüm oranları: 1960-2017 yılları arasında her ülkede meydana gelen doğum ve ölüm oranını ifade etmektedir. Değerler her bin kişi için meydana gelen doğum ve ölüm oranını ifade etmektedir. Toplam doğum oranı ve yaşa özel doğurganlık oranları toplam ülke nüfusunu etkileyen faktörlerdendir. Toplam doğum oranı bir kadının ömrü boyunca verdiği ortalama canlı doğum sayısını ifade etmektedir.

Doğumdaki cinsiyet oranı: 2001 yılında bu zamana sağlanan verilere bakıldığında 105 oranı sabit kalmaktadır. Dünya genelinde bu oranın biyolojik sebeplerden dolayı 103-107 arasında değiştiği gözlemlenmektedir. Bu sebeple, bu çalışmanın kuşak bileşenleri ile nüfus tahmini aşamasında, farklı çalışmalardaki nüfus projeksiyonlarında olduğu gibi, doğumdaki cinsiyet oranı 105 olarak sabit alınmıştır.

Beklenen yaşam süresi: Bir bireyin doğumda beklenen yaşam süresini ifade eder. Örneğin; Türkiye'de 1960 yılında doğan bir birey için beklenen yaşam süresi 45 iken, 2016 yılında bu rakam 75'e yükselmiştir.

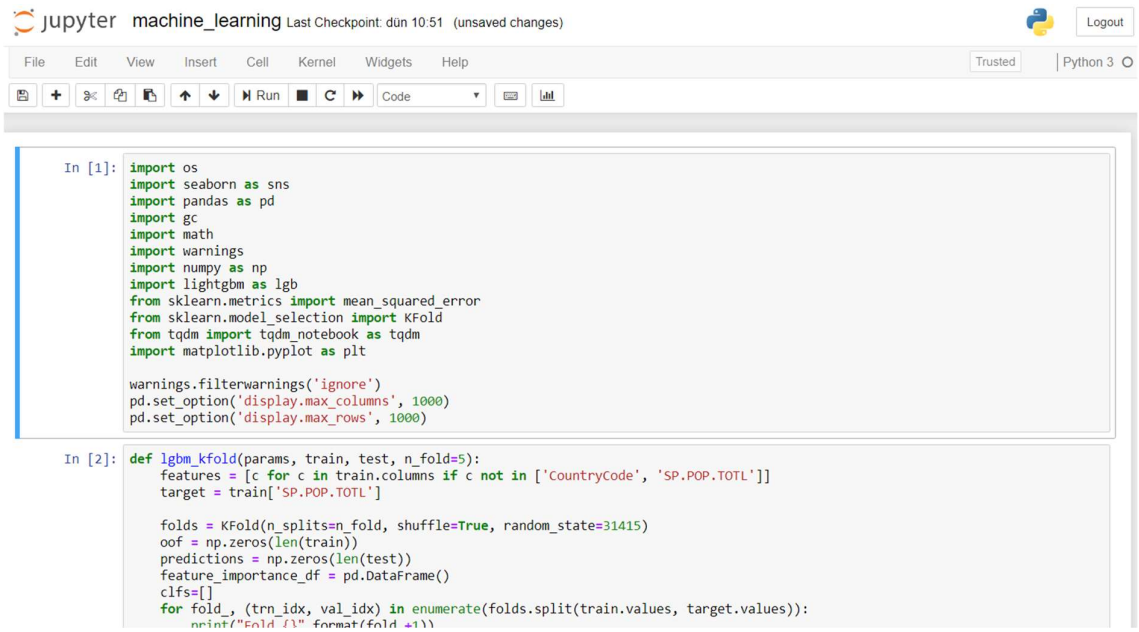
Göçler: Ülkenin dışarıya verdiği göç ile aldığı göç arasındaki fark net göç olarak ifade edilmiştir. Toplam ülke nüfusunu etkileyen faktörlerden biridir.

5.4 Çalışmada Kullanılan Yazılımlar

Makine öğrenmesi algoritmaları ile analiz imkanı veren birçok açık ve kapalı kaynak kodlu program bulunmaktadır. Bunlara örnek olarak SPSS, KXEN, RapidMiner, WEKA, Oracle, Jupyter, Pyzo, Spyder, WinPhyton, BigML, Datarobot, Google Cloud AutoML, IBM Watson Studio, Microsoft Azure Machine Learning Studio, Paxata, Trifacta verilebilir. Bu çalışmada makine öğrenmesi algoritmalarının çalıştırılması ve modelin kurulması için Python dili kullanılmıştır. Python geliştirme dili dünya genelinde yapay zeka ve veri madenciliği alanlarında sıkça kullanılmaktadır.

Python birçok açık kütüphane ile beraber kullanılarak makine öğrenmesi analizine imkan vermektedir. En çok kullanılan kütüphanelere örnek olarak Scikit-learn, Scrapy, NumPy, SciPy, SymPy, Dask, Numba, HPAT, Cython verilebilir. Bu çalışmada Python dili çalışma arayüzü olarak Jupyter kullanılmıştır. Jupyter farklı program dillerini bir arada kullanılmasını sağlayan yazılımdır. Etkileşimli shell sunması,

otomatik tamamlama özelliği gibi kullanışlı bir arayüze sahip olması, tarayıcı tabanlı defter (notebook) özelliği gibi avantajları vardır. Açık kaynak kodlu bir yazılım olup "https://jupyter.org/" adlı internet sitesinden ücretsiz olarak kurulabilir ve akademik araştırmalar için kullanılabilir. Kurulum için sistemde Python 3.3 veya yukarısı kurulu olmalıdır. Anaconda aracılığı ile kurulabileceği gibi Python Package Index aracı olan "pip" ile de kurulabilir. Jupyter'e ait arayüz aşağıdaki Şekil 5.1'de görülmektedir.



Şekil 5.1 : Jupyter program arayüzünden bir kesit.

Nüfus projeksiyonu üretmek için geliştirilen birçok yazılım bulunmaktadır. İlk yazılım 1995 yılında DOS işletim sistemine uygun olarak geliştirilmiş ve FIVFIV adı verilmiştir. Şu anda kullanımda olan PDE Population Projection, Demographic Analysis & Population Projection System (DAPPS), Spectrum gibi yazılımlar bulunmaktadır. Türkiye İstatistik Kurumu (TÜİK) nüfus projeksiyonları üretmek için Microsoft Excel ortamında çalışan Visual Basic for Applications (VBA) programlama dilinde yazılmış makro kodlar kullanmaktadır. Bu çalışmada kuşak bileşenleri ile nüfus tahmini için Birleşmiş Milletler ve Avenir Health tarafından geliştirilen, kullanışlı bir arayüz tasarımına sahip olan Spectrum adlı program kullanılmıştır. Analizlerin yapılması için kullanılan bilgisayarın özellikleri şu şekildedir: Intel Core i7-2530QM 2.00 GHz, 14 GB ram, Windows 10 64Bit işletim sistemi.

5.5 Verilerin Hazırlanması ve Modelin Kurulması

Dünya Bankası'ndan elde edilen veri setinde yıl bilgisi içeren değerler farklı sütunlarda bulunmaktadır. Algoritmaların analizi daha hızlı yapabilmesi için yıl değişkeni sütundan satıra aktarılmıştır. Her bir sütunda bulunan göstergelerin ülkelere ve yıllara karşılık değerleri satırlar halinde verilmektedir. Veri setinde 262 ülke için 1595 farklı gösterge bulunmaktadır. Ülkeler arasından, Kuveyt, Bati Filistin, Sırbistan, Sint Maarten ve Eritre ülkelerinin demografik göstergeleri arasından nüfus istatistiklerinin çoğu boş olduğu için bu ülkeler veri setinden çıkarılmıştır.

Veri seti içerisinde bazı değişkenlerde boşluklar olduğu tespit edilmiştir. Her ne kadar nüfus ile ilgili göstergelerin doluluk oranı yüksek olsa da, diğer değişkenlerin bazılarında kayıp değerler bulunmaktadır. Değişkenler kendi içerisinde değerlendirildiğinde kayıp değer sayısı %33 ve üzeri olan değişkenler veri setinden çıkarılmıştır. Geriye kalan değişkenlerdeki kayıp değerler interpolasyon yöntemi ile doldurulmuştur. Zaman serisi analizlerinde kayıp değerlerin tamamlanması için tavsiye edilen yöntemlerden biri de interpolasyondur (Brubacher ve Wilson, 1976; Moritz ve diğ, 2015). 257 ülkenin 1595 farklı demografik göstergesi üzerinden model eğitilecektir. Tüm veri seti 3406x2836 boyuta sahiptir.

Ayrıca veri seti üzerinde box-cox transformasyon işlemi uygulanmıştır. Ancak bu işlem model performansını düşürdüğü için uygulama sürecinde kullanılmamıştır.

Veri setine bütün ülkeleri dahil etmekteki amaç modelin daha iyi öğrenmesini sağlayarak tahmin sonuçlarını iyileştirmektir. Veri setine 257'den daha az ülke dahil edildiğinde model performansı düşmektedir. Her ne kadar model bazı ülkeler için kötü tahmin sonuçları üretse de, o ülkelerin parametresinden farklı bilgiler öğrendiği için bütün ülkelerin dahil edilmesi modelin genel performansını arttırmaktadır. Bu sebeple modelin öğrenme aşamasında bütün ülkelerin verileri kullanılmıştır.

Elimizdeki veri setini zenginleştirmek ve değişken (feature) sayısını arttırarak makinenin daha iyi öğrenmesini sağlamak için LightGBM, Lineer regresyon ve Ridge regresyon modellerinde iyileştirmeler yapılmıştır. Bu iyileştirmeler yapılırken yanlılık/varyans ikilemi (Bias/Variance Tradeoff) göz önüne alınmıştır. Yapılan denemeler neticesinde en iyi sonuçları veren parametreler seçilmiştir. Buna göre, her bir değişkenin son 5 yıl değerinin yeni değişken olarak atanması ve 2000-2017 yılları

arasındaki değerleri kullanarak modelin eğitilmesi LightGBM, Doğrusal regresyon ve Ridge regresyon modellerinin en iyi sonucu vermesini sağlamıştır. 1960'a kadar olan veriler kullanılarak LightGBM modeli ile analiz yapıldığında daha yüksek RMSE ve MAPE değerleri elde edildiği görülmüştür. Buna sebep olarak geçmişe doğru gidildikçe değişkenlerin varyansının artması, dolayısıyla modelin öğrenmesinin zorlaşması söylenebilir. Model öğrenme sürecinde 2017 yılını tahmin etmeyi öğrendiği gibi, geçmişe yönelik örneğin 1970 yılını da analiz ederek verileri uydurmaya (fitting) devam etmektedir. Veri setindeki nüfus değerleri, yapısı gereği yıllar geçtikçe belirli bir trendi takip ettiğinden, geçmişe doğru bütün değerlerin alınarak modelin eğitilmesi model performansını düşürmektedir. Yanlılık/Varyans ikilemi (Bias/Variance Tradeoff) göz önüne alınarak denemeler yapıldığında model en iyi performansı 2000-2017 yılları için göstermiştir.

Bu sebeple, LightGBM, Doğrusal Regresyon ve Ridge Regresyon analizleri için 2000-2017 yılları arasındaki veriler kullanılacaktır. Bu algoritmalar için uygulanacak olan parametre (lag) değeri 5 olarak seçilmiş ve değişkenler içerisinde kayıp değer sayısı %33 ve üzeri olan değişkenler veri setinden çıkarılmıştır. Geriye kalan 2838 değişken üzerinden model eğitilecektir. Bütün değişkenler üzerinden model eğitildikten sonra sadece kuşak bileşenleri değişkenleri ile model eğitilerek sonuçları karşılaştırılacaktır. Yukarıda bahsedildiği gibi, veri setinin yapısı gereği sadece kuşak bileşenleri değişkenleri kullanılarak model eğitildiğinde LightGBM algoritmasının performansında bir değişim beklenmezken, Doğrusal ve Ridge regresyonu analizlerinin performansında artış beklenmektedir. Literatür kısmında ayrıntılı şekilde bahsedildiği üzere kuşak bileşenleri değişkenleri demografik göstergelerden nüfus, doğum oranları, beklenen yaşam süresi, ölüm oranları ve göç oranları ile ilgili olanlardır.

Holt-Winters, ARIMA ve Prophet analizleri için ise 1960-2017 yılları arasındaki toplam nüfus değişkeni kullanılarak 2017 yılının toplam nüfusu tahmin edilecektir.

Hedef değişken olarak çalışmanın amacına yönelik tahmin edilmesi planlanan 2017 yılı toplam nüfus değeri seçilmiştir. Böylece, gerçek değer ile tahmin edilen değer karşılaştırılabilecektir. Bu süreçte 2017 yılı için tahmin edilen toplam nüfus, 2016 ve 2017 yılındaki gerçek toplam nüfus değeri ile karşılaştırılarak çalışmanın başarısı ortaya konulacaktır.

5.6 Algoritmalardan Elde Edilen Bulgular

Nüfus tahmini için daha önceden belirlenen LightGBM, Doğrusal Regresyon, Ridge Regresyon, Holt-Winters, ARIMA ve Prophet algoritmaları ile bir önceki bölümde bahsedilen hazırlık aşamalarından geçen veri seti Jupyter arayüzü ile Python yazılım dili üzerinden analiz edilmiştir. Veri seti zaman serisi içerdiği için model 2016 yılına kadar olan veriler ile eğitilmiş, 2017 yılındaki toplam nüfus ile test edilmiştir. Bu süreçte 6 tane farklı makine öğrenmesi algoritması kullanılarak performansları karşılaştırılmıştır. RMSE ve MAPE kriterleri üzerinden performans sonuçlarına ilişkin tablo Çizelge 5.2'dedir.

Modeller	Versiyon	RMSE	MAPE	CPU Süresi (sn)
Doğrusal Regresyon	Tüm Veri Seti	2278905,76	2,580%	13
Çıkıntı Regresyon	Tüm Veri Seti	2479951,04	2,760%	15
LightGBM	Tüm Veri Seti	25098273,89	7,490%	650
Doğrusal Regresyon	Filtreli Veri Seti	363726,94	2,280%	8
Çıkıntı Regresyon	Filtreli Veri Seti	334646,14	1,850%	9
LightGBM	Filtreli Veri Seti	30830492,06	7,480%	450
Holt Winters	Sadece Nüfus	157461,34	0,080%	42
ARIMA	Sadece Nüfus	136570,57	0,100%	43
Prophet	Sadece Nüfus	2427166,07	1,430%	86
Son Sene Nüfusu	Sadece Nüfus	11555505,55	1,530%	0

Çizelge 5.2 : Makine öğrenmesi algoritmaları performans karşılaştırması

Bu çizelgede bütün veri seti işleme alınarak yapılan analizlerin sonucu "Tüm Veri Seti Versiyonu" olarak; sadece kuşak bileşenleri değişkenleri ile yapılan analizlerin sonucu "Filtreli Veri Seti Versiyonu" olarak gösterilmiştir. Son sene nüfusu ise, herhangi bir analiz ve tahmin yapılmadan 2017 yılı için bir önceki yılın nüfus değeri yazıldığında ortaya çıkan hata oranıdır.

Performans sonuçları değerlendirildiğinde tüm veri seti ve filtreli veri seti üzerinden eğitilen Doğrusal Regresyon ve Ridge Regresyon modellerinin, LightGBM modeline göre daha iyi sonuç verdiği görülmektedir. Buna sebep olarak Doğrusal Regresyon ve Ridge Regresyon modellerinin çalışma prensibi açısından zaman serisi veri setine daha uygun olması söylenebilir. Daha çok sınıflandırma problemleri için kullanılan karar ağaçları algoritmaları, her bir ağacı oluştururken dallanmalar suretiyle karar verdiği

için zaman serisi içeren veri setleri üzerinde trendi ve mevsimselliği tespit etmesi zorlaşmaktadır. Ayrıca, LightGBM algoritması diğer algoritmalara nispetle daha kolay aşırı öğrenme (overfitting) göstermektedir. Bu sebepler ile regresyon algoritmaları daha iyi sonuç vermektedir.

Doğrusal Regresyon ve Ridge Regresyonu filtrelenmiş veri seti üzerinden eğitildiğinde tüm veri seti ile eğitilmesine göre daha iyi sonuç vermesinin sebebi, çalışma prensibi gereği bütün parametreleri regresyon denklemine eklenmesidir. LightGBM algoritması veri seti üzerinden rassal bir şekilde değişken seçimi (feature selection) uygulayarak çalışmaktadır. Yani, LightGBM çalışma prensibi gereği anlamsız verileri tespit ederek göz ardı etmektedir. Bunun için, modelin performansını artıran değişkenlere daha çok değer verirken, sonuca anlamlı bir katkı sağlamayan parametrelere daha az önem vererek tahmin yapmaktadır. Bu sebeple, LightGBM ile analiz edilen veri setinde anlamsız bilgilerin bulunması sonuç üzerinde büyük bir etkiye sahip değildir. Bu sebeple veri seti değiştirildiğinde LightGBM algoritmasının performansı önemli ölçüde değişmemiştir. Ancak, Doğrusal Regresyon ve Ridge Regresyonu değişken seçimi (feature selection) yapma özelliği bulunmadığı için filtrelenmiş veri üzerinden model eğitildiğinden daha iyi performans göstermektedir.

Regresyon algoritmaları arasında değerlendirme yapıldığında özellikle filtrelenmiş veri seti üzerinden yapılan analizlerde Ridge Regresyon modelinin daha başarılı sonuç verdiği görülmektedir. Buna sebep olarak Ridge Regresyonunun, Lineer Regresyon modelinin farklılaşmış bir versiyonu olması söylenebilir. Ridge Regresyonu analiz yaparken düzenlileştirme terimi kullanarak yüksek olan model ağırlıklarının düşmesini sağlar. Böylece modeli düzenleyerek daha iyi öğrenmesini sağlamaktadır. Böylelikle Ridge Regresyonu Doğrusal Regresyona göre aşırı öğrenme (overfitting) problemini minimize ederek kısmen daha iyi sonuç vermiştir.

Ayrıca, LightGBM algoritması filtrelenmiş veri seti üzerinden eğitildiğinde performansı düşmektedir. Bu sonuca dayanarak LightGBM algoritmasının tüm veri seti içerisinde uygulandığında farklı öğrenmeler gerçekleştirdiği söylenebilir. RMSE değeri yükselirken MAPE değeri sabit kalmaktadır. Bunun sebebi olarak model bütün veri seti ile eğitildiğinde rassal olarak ortaya çıkmış örüntüleri öğrenmiş olabilir. Bu da aşırı öğrenmeye (overfitting) sebep olabilir. Çalışmanın ilerleyen kısımlarında anlatılan Türkiye örneği ile bu yorum desteklenecektir.

Bütün modellerin performansları karşılaştırıldığında Holt-Winters, ARIMA ve Prophet analizlerinin diğer modellere nispetle daha başarılı sonuçlar verdiği görülmektedir. Kendi içlerinde ise Holt-Winters ve ARIMA birbirlerine yakın ve Prophet modelinden daha iyi sonuçlar vermektedir. Veri seti zaman serisi problemi olduğu için bu algoritmalar daha iyi sonuç vermektedir. Holt-Winters, ARIMA ve Prophet zaman serisi problemleri için sıkça kullanılan algoritmalardır (Grim ve Gradvohl, 2018; Chiru ve Posea, 2018).

257 farklı ülke için model eğitilerek model performansları değerlendirildiğinde ARIMA ve Holt-Winters modellerinin Doğrusal ve Ridge Regresyonu modellerine göre daha iyi sonuçlar verdiği görülmektedir. Ancak, sonuçların birbirine yakın olması sebebiyle ülke bazında analizler yapıldığında modellerin birbirlerine karşı üstünlükleri değişebilir.

Ayrıca, 2017 yılı nüfusunu tahmin etmek yerine 2016 yılı nüfusunu kullanmak, LightGBM modeli ile tahmin yapmaya göre daha iyi bir sonuca ulaştırmaktadır. Ancak, LightGBM modeli uzun yıllar sonrasını tahmin etmek için kullanıldığında bu yorum değişebilir.

Tüm veri seti üzerinden LightGBM ile analiz yapıldıktan sonra hangi değişkenlerin modelin performansına daha büyük katkısının olduğu görmek için değişkenler (feature importance) önem sırasına göre listelenmiştir. Nüfusu etkileyen ölüm, doğum ve nüfus istatistikleri gibi değerlerin ön plana çıkması beklenirken, nüfus ile ilgisi bulunmayan değişkenler daha üst sıralarda yer almıştır. Örneğin; Fransa ve Japonya ülkelerinden net bağış akışı, kömür fiyatları, azot oksit emisyon değerleri gibi değişkenlerin önem (feature importance) sıralamasında nüfus değerlerinden üst sıralarda gösterilmektedir. Buna sebep olarak, LightGBM algoritması veri seti içerisinde öğrenim uygularken, bir değişken üzerinden öğrendiği bilginin bir benzerini farklı bir değişken üzerinde gördüğünde, ilk değişken üzerinden o bilgiyi öğrendiği için ilk değişkene daha yüksek önem sırası verecektir. Aynı örüntüye sahip diğer değişkenden farklı bir bilgi öğrenmediği için ikinci değişkenin önem sırasını düşürecektir. Değişkenler arasında korelasyon haritası çıkarıldığında kömür fiyatları ile başka bir değişken arasında korelasyon beklenebilir.

Değişken önem (feature importance) sıralamasında nüfus değerlerinin alt sıralarda olmasının diğer bir sebebi, ülke bazında nüfus istatistikleri üzerindeki trend

değerlerinin ülkeler arasında farklılık göstermesi olabilir. Model, bütün ülkeler üzerinden veri setini öğrendiği için, nüfus oranları arasındaki trend farklılığı nüfus değişkeninin önem sırasını düşürmeye sebep olabilir.

5.6.1 Türkiye'nin 2017 yılı Nüfus Tahmini

Modeller	RMSE	MAPE %	Tahmin edilen 2017 Nüfusu
Doğrusal Regresyon	39836,04	0,0493%	80784860
Çıkıntı Regresyon	5966,25	0,0074%	80750990
Light GBM	6718648,60	8,3208%	87463669
Holt Winters	8360,00	0,0104%	80753380
ARIMA	6434,02	0,0080%	80751454
Prophet	813871,04	1,0100%	79931150
Kuşak Bileşenleri	427708,00	0,5297%	80317312
TÜİK Tahmini	979008,00	1,2125%	79766012
2017 yılı Türkiye Nüfusu			80745020

Çizelge 5.3 : 2017 yılı Türkiye nüfusu tahmin sonuçları

Çizelge 5.3'te 7 farklı modelin 2017 yılı Türkiye nüfusu tahmin performansları gösterilmiştir. Ayrıca TÜİK tarafından 2013 yılında açıklanan nüfus projeksiyonunun 2017 yılı için nüfus tahmini çizelgeye eklenmiştir.

Makine öğrenmesi algoritmaları bütün veri seti ile eğitildikten sonra, Türkiye'nin veri setini kullanarak 2017 yılı toplam nüfusu tahmin edilmiştir. Tahmin sonuçlarına göre gerçek değere en yakın tahmini Ridge Regresyon modeli vermektedir. ARIMA algoritmasının performansı da oldukça yüksektir. Ridge Regresyonu, Holt-Winters ve ARIMA modelleri 2017 yılı toplam nüfusunu 10 bin kişinin altında bir sapma ile tahmin etmiştir.

Türkiye'nin 2016 yılı için toplam nüfus, doğum, ölüm ve göç oranları, doğumda beklenen yaşam süresi ve doğumdaki cinsiyet oranı verileri veri ile kuşak bileşenleri yöntemi kullanılarak 2017 yılı nüfus tahmini üretilmiştir. Kuşak bileşenleri yöntemi ile nüfus tahmini için literatür kısmında açıklanan işlemler uygulanmıştır. Nüfus tahmini için sıkça kullanılan kuşak bileşenleri yöntemi ile yapılan nüfus tahmini, gerçek nüfusun yaklaşık 400 bin kişi uzağındadır. TÜİK tarafından 2013 yılında yapılan nüfus projeksiyonuna göre 2017 yılı nüfus tahmini gerçek değerin yaklaşık 1 milyon kişi uzağındadır. Makine öğrenmesi algoritmaları kuşak bileşenleri yöntemine göre çok daha tutarlı ve gerçeğe yakın sonuçlar vermiştir.

Ayrıca, Türkiye'nin kendi veri seti ile model eğitilerek nüfus tahmini yapıldığında gerçek değerin yaklaşık 25 milyon uzağında bir rakama ulaşılmaktadır. Bu da modelin eğitilmesinde birden fazla ülkenin veri setinin kullanılmasının sağladığı katkıyı göstermektedir.

Doğrusal ve Ridge Regresyon modelleri tüm veri seti ile eğitildikten sonra yapılan nüfus tahmini sonuçları Holt-Winters ve ARIMA modellerine göre daha uzak sonuçlar vermiştir. Holt-Winters ve ARIMA modelleri ile nüfus tahmini yapmak için veri setindeki değişkenler arasından sadece toplam nüfus oranı kullanılarak bu modeller eğitilmiştir. Bu dört modelin birbiri ile aynı düzlemde karşılaştırılması için, sadece toplam nüfus değerleri ile modeller eğitilerek elde edilen sonuçlar Çizelge 5.4'te gösterilmiştir.

	TÜM ÜLKELER		**TÜRKİYE**		
Modeller	**RMSE**	**MAPE %**	**RMSE**	**MAPE %**	**Tahmin edilen 2017 Nüfusu**
Doğrusal Regresyon	106982,05	0,653%	20813,08	0,0258%	80765830
Çıkıntı Regresyon	106982,13	0,653%	20813,41	0,0258%	80765830
Holt Winters	157461,34	0,080%	8360,00	0,0104%	80753380
ARIMA	136570,57	0,100%	6434,02	0,0080%	80751454
2017 yılı Türkiye Nüfusu					80745020

Çizelge 5.4 : Toplam nüfus değişkeni ile eğitilen makine öğrenmesi algoritmalarının bütün ülkeler ve Türkiye üzerindeki performans sonuçları.

Bütün ülkelerin toplam nüfus değerleri kullanılarak modeller eğitildiğinde, Doğrusal ve Ridge regresyon modellerinin Holt-Winters ve ARIMA modellerine göre daha iyi sonuç verdiği gözlemlenmektedir. Ancak, Türkiye verileri üzerinden tahmin yapıldığından Holt-Winters ve ARIMA modellerinin daha iyi sonuç verdiğini görülmektedir. Bu modellerin performansı LightGBM ve Prophet modellerine göre oldukça başarılıdır. Genel anlamda bütün ülkeler üzerinden modellerin performansları karşılaştırıldığında Doğrusal ve Ridge Regresyon modellerinin nüfus tahmini için daha anlamlı sonuçlar vereceği düşünülmektedir.

Modeller	Versiyon	RMSE	MAPE
Doğrusal Regresyon	Tüm Veri Seti	2278905,76	2,580%
Çıkıntı Regresyon	Tüm Veri Seti	2479951,04	2,760%
Doğrusal Regresyon	Filtreli Veri Seti	363726,94	2,280%
Çıkıntı Regresyon	Filtreli Veri Seti	334646,14	1,850%
Doğrusal Regresyon	Sadece Nüfus	106982,05	0,653%
Çıkıntı Regresyon	Sadece Nüfus	106982,13	0,653%

Çizelge 5.5 : Doğrusal ve Ridge Regresyon modellerinin farklı eğitim setlerine göre performans sonuçları.

Ayrıca, Doğrusal ve Ridge Regresyon modelleri bütün veri seti ile eğitildikten sonra elde edilen sonuçlar ile sadece toplam nüfus verileri ile eğitildikten sonra alınan sonuçlar değerlendirildiğinde büyük oranda performans artışı gözlemlenmektedir. Çizelge 5.5'te bu fark açıkça görülebilmektedir. Buna sebep olarak, Doğrusal ve Ridge regresyon modelleri bütün veri seti ile eğitildiklerinde çalışma prensibi gereği değişken seçimi (feature selection) özelliği bulunmadığı için, bütün değişkenleri regresyon denkleminde hesaba katmaktadır. Bununla birlikte, model bütün veri seti içerisinde aslında bulunmayan ve rassal olarak ortaya çıkmış örüntüleri öğrenmiş olabilir. Ancak, sadece nüfus değerleri ile eğitildiğinde model performansı gözle görülür oranda artmıştır.

6. SONUÇ VE ÖNERİLER

Bu çalışmada, farklı makine öğrenmesi algoritmaları ile nüfus tahmini yapılmıştır. Farklı modeller geliştirilerek en iyi performans gösteren modeller uygulanmıştır. Elde edilen sonuçlara göre, makine öğrenmesi algoritmaları nüfusu binde bir yanılma payı ile tahmin edebilmektedir. Türkiye örneğinde ulaşılan sonuçlara göre Türkiye'nin nüfusu 6 bin kişilik fark ile kuşak bileşenleri yönteminden daha iyi tahmin edilmiştir. Makine öğrenmesi algoritmaları ülkenin nüfusunu tahmin etmeyi zorlaştıran sebepleri minimize edip, demografik veriler üzerindeki belirsizlikleri analiz ederek ülkenin gelecekteki nüfusunu daha iyi tahmin edebilmektedir. Bu sebeple, makine öğrenmesi algoritmalarının nüfus tahmini üzerinde kullanılması ülke için önemli bir katkı sağlayacaktır. Böylece ülke hakkında ulusal ihtiyaçların planlanması kolaylaşacak ve sosyal, ekonomik, çevresel kararların daha tutarlı alınması sağlanacaktır.

Bu çalışmada makine öğrenmesi algoritmaları kuşak bileşenleri yönteminden daha tutarlı nüfus tahmini yapmıştır. Dünya genelinde nüfus projeksiyonu üretmek için en sık kullanılan yöntemlerden biri kuşak bileşenleri yöntemidir. Kuşak bileşenleri yöntemi ile ülkenin toplam nüfusu, doğum, ölüm ve göç oranları, doğumda beklenen yaşam süresi ve doğumdaki cinsiyet oranı verileri kullanılarak nüfus tahmini yapılır. Ancak, makine öğrenmesi algoritmaları ile demografik göstergeler arasından öngörülemeyen, farklı değişkenler de hesaba katılır. Böylelikle farklı değişkenlerden öğrendiği bilgiler ile daha tutarlı tahmin sonuçları üretebilmektedir. Bu sebeple, nüfus tahmini için makine öğrenmesi algoritmalarının kullanımı üretilen nüfus tahminlerini geliştirecektir.

LightGBM analizi kuşak bileşenleri değişkenleri ile filtrelenmiş veri seti üzerinden yapıldıktan sonra değişkenlerin önem değerleri (feature importance) sıralandığında, doğumdaki cinsiyet oranının en yüksek önemi taşıdığı görülmektedir. Doğumdaki cinsiyet oranı kuşak bileşenleri yönteminde kullanılan önemli değişkenlerden bir tanesidir. Ancak, TÜİK tarafından yapılan nüfus projeksiyonlarında bu değişken sabit

alınmaktadır. Nüfus tahminini etkileyen önemli bir değişken olduğu için sabit alınmaması üretilecek olan nüfus projeksiyonlarının tutarlılığını arttıracaktır.

Araştırmanın sınırlarından bir tanesi 257 farklı ülkenin bir havuzda toplanarak modelin eğitilmesi olarak söylenebilir. Yapılan çalışmaya göre bütün ülkelerin demografik göstergeleri ile modelin eğitilmesi, daha az ülkenin veri seti ile eğitilmesine göre daha iyi sonuç vermektedir. Ancak, ülkelerin gelişmiş düzeylerine, kültürel özelliklerine ya da coğrafi konumlarına göre gruplandırarak modellerin eğitilmesi farklı sonuçlar doğurabilir.

Çalışmanın sınırlarından bir diğeri ise veri büyüklüğüdür. Makine öğrenmesi algoritmaları büyük veri setleri (big data) ile daha anlamlı sonuçlar vermektedir. Ancak, problemin doğası gereği ülkelerin nüfus istatistikleri ve demografik göstergeleri kısıtlı veri setine sahiptir. Bununla birlikte, özellikle gelişmemiş ülkeler için veri seti yetersiz olmakla beraber mevcut verilerin tutarlılığı da şüphelidir. Toplanan verilerin tutarlı olması ve veri sayısının artırılması nüfus tahminini iyileştirecektir.

İleriye yönelik çalışmalarda kuşak bileşenleri yöntemi ile makine öğrenmesi algoritmaları sentezlenebilir. Kuşak bileşenleri yönteminde tahmin edilmek istenen seneye kadar olan bütün girdi verilerine ihtiyaç duyulur. Örneğin, 2050 yılı için nüfus tahmini yapılmak istendiğinde, aradaki 31 yıl boyunca her bir yıl için doğum oranları gibi bütün değişkenlerin tahmin edilmesi gerekir. Her bir yıl için doğum, ölüm, göç oranları gibi bu girdi verileri makine öğrenmesi ile tahmin edildikten sonra elde edilen veriler kuşak bileşenleri yönteminde kullanılabilir. Bu çalışma, tahmin edilecek olan yıl sayısı artırılarak genişletilebilir. Ülkeler nüfus projeksiyonu üretirken 50 yıl sonrasını da tahmin etme ihtiyacı duyabilirler. Uzun yıllar sonrasını tahmin etmek için makine öğrenmesi algoritmalarının kullanımı önemli bir katkı sağlayabilir.

Bu çalışmada tüm veriler analiz edildikten sonra, daha iyi karşılaştırma yapabilmek için, kullanılan değişkenler kuşak bileşenleri yönteminde kullanılan değişkenler ile filtrelenmiştir. Bu boyut düşürme işlemi gelecekteki araştırmalarda farklı makine öğrenmesi yöntemleri kullanılarak yapılabilir. Örneğin, Temel Bileşen Analizi (Principal Component Analysis, PCA) ile bütün veri seti üzerinden çalışma yapılarak veri seti farklılaştırılabilir.

KAYNAKLAR

Ahmad, A. S., Hassan, M. Y., Abdullah, M. P., Rahman, H. A., Hussin, F., Abdullah, H., & Saidur, R. (2014). A review on applications of ANN and SVM for building electrical energy consumption forecasting. Renewable and Sustainable Energy Reviews, 33(1), 102-109.

Alkema, L., Raftery, A. E., Gerland, P., Clark, S. J., & Pelletier, F. (2012). "Estimating trends in the total fertility rate with uncertainty using imperfect data: Examples from West Africa." Demographic research 26(15).

Alpaydin, E. ve F. Bach (2014). Introduction to Machine Learning, MIT Press.

Arnell, N. W. (2004). "Climate change and global water resources: SRES emissions and socio-economic scenarios." Global environmental change 14(1): 31-52.

Bamgbose, J. A. (2009). "Falsification of population census data in a heterogeneous Nigerian state: The fourth republic example." African Journal of Political science and International relations 3(8): 311.

Bandyopadhyay, G. and S. Chattopadhyay (2006). "An Artificial Neural Net approach to forecast the population of India." arXiv preprint nlin/0607058.

Batista, G. E. A. P. A. ve M. C. Monard (2003). "AN ANALYSIS OF FOUR MISSING DATA TREATMENT METHODS FOR SUPERVISED LEARNING." Applied Artificial Intelligence 17(5/6): 519.

Bayes, T., ve M. Price, (1763), An essay towards solving a problem in the doctrine of chances: Philosophical Transactions, 53, 370–418, https://doi.org/10.1098/rstl.1763.0053.

Bélisle, E., Huang, Z., Le Digabel, S., & Gheribi, A. E. (2015). "Evaluation of machine learning interpolation techniques for prediction of physical properties." Computational Materials Science 98: 170-177.

Bell, J. (2014). Machine Learning : Hands-On for Developers and Technical Professionals. Somerset, UNITED STATES, John Wiley & Sons, Incorporated.

Bergmeir, C. ve J. M. Benítez (2012). "On the use of cross-validation for time series predictor evaluation." Information Sciences 191: 192-213.

Bontempi, G., Taieb, S. B., & Le Borgne, Y. A. (2012, July). Machine learning strategies for time series forecasting. In European business intelligence summer school (pp. 62-77). Springer, Berlin, Heidelberg.

Boon, M. E. ve L. Kok (1995). Classification of cells in cervical smears. Applications of Neural Networks, Springer: 113-131.

Bowley, A. (1924). "Births and population in Great Britain." The Economic Journal 34(134): 188-192.

Box, G. E. and N. R. Draper (1987). Empirical model-building and response surfaces, John Wiley & Sons.

Boyle, J. P., Thompson, T. J., Gregg, E. W., Barker, L. E., & Williamson, D. F. (2010). "Projection of the year 2050 burden of diabetes in the US adult population: dynamic modeling of incidence, mortality, and prediabetes prevalence." Population health metrics 8(1): 29.

Börsch-Supan, A. (2003). "Labor market efftects of population aging." Labour 17: 5-44.

Breiman, L. (1996). "Bagging predictors." Machine learning 24(2): 123-140.

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

Breiman, L. (2017). Classification and regression trees, Routledge.

Brubacher, S. R., & Wilson, G. T. (1976). Interpolating time series with application to the estimation of holiday effects on electricity demand. Journal of the Royal Statistical Society: Series C (Applied Statistics), 25(2), 107-116.

Cannan, E. (1895). "The probability of a cessation of the growth of population in England and Wales during the next century." The Economic Journal 5(20): 505-515.

Chen, K.-Y. and C.-H. Wang (2007). "A hybrid SARIMA and support vector machines in forecasting the production values of the machinery industry in Taiwan." Expert Systems with Applications 32(1): 254-264.

Chen, T. ve C. Guestrin (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, ACM.

Chiru, C.-G. and V.-V. Posea (2018). Time Series Analysis for Sales Prediction. International Conference on Artificial Intelligence: Methodology, Systems, and Applications, Springer.

Chou, J.-S. ve D.-S. Tran (2018). "Forecasting energy consumption time series using machine learning techniques based on usage patterns of residential householders." Energy 165: 709-726.

Cortes, C., Mohri, M., Pechyony, D., & Rastogi, A. (2009). Stability Analysis and Learning Bounds for Transductive Regression Algorithms.

Crevier, D. (1993). AI: The tumultuous search for artificial intelligence: Basic Books.

Cruz, J. A. ve D. S. Wishart (2006). "Applications of machine learning in cancer prediction and prognosis." Cancer informatics 2: 117693510600200030.

Cukier, K. (2010). Data, data everywhere: A special report on managing information, Economist Newspaper.

Cakmak, T., Tekin, A. T., Senel, C., Coban, T., Uran, Z. E., & Sakar, C. O. (2019). Accurate Prediction of Advertisement Clicks based on Impression and Click-Through Rate using Extreme Gradient Boosting.

Dietterich, T. G. (2000). Ensemble methods in machine learning. International workshop on multiple classifier systems, Springer.

Dorn, H. F. (1950). "Pitfalls in population forecasts and projections." Journal of the American Statistical Association 45(251): 311-334.

Ediev, D. ve M. M. Yüceşahin (2016). "Contribution of migration to replacement of population in Turkey." Migration Letters 13(3): 377-392.

Fix, E. ve J. Hodges (1951). "An important contribution to nonparametric discriminant analysis and density estimation." International Statistical Review 3(57): 233-238.

Folorunso, O., Akinwale, A. T., Asiribo, O. E., & Adeyemo, T. A. (2010). "Population prediction using artificial neural network." African Journal of Mathematics and Computer Science Research 3(8): 155-162.

Friedman, J. H. (2001). "Greedy function approximation: a gradient boosting machine." Annals of statistics: 1189-1232.

Garrido, A. P. (2016). "What Is the Difference between Bagging and Boosting?" Quantdare, 20 Apr. 2016, quantdare.com/what-is-the-difference-between-bagging-and-boosting/. [Ziyaret tarihi: 28.02.2018]

Geurts, P., Ernst, D., & Wehenkel, L. (2006). "Extremely randomized trees." Machine learning 63(1): 3-42.

Goudie, A. S. (2018). Human impact on the natural environment, John Wiley & Sons.

Grim, L. F. L. ve A. L. S. Gradvohl (2018). High-Performance Ensembles of Online Sequential Extreme Learning Machine for Regression and Time Series Forecasting. 2018 30th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD), IEEE.

Grubbs, F. E. (1969). "Procedures for detecting outlying observations in samples." Technometrics 11(1): 1-21.

Heuveline, P. (2003). "HIV and population dynamics: A general model and maximum-likelihood standards for East Africa." Demography 40(2): 217-245.

Hu, Y., Peng, L., Li, X., Yao, X., Lin, H., & Chi, T. (2018). "A novel evolution tree for analyzing the global energy consumption structure." Energy 147: 1177-1187.

Huang, W., Nakamori, Y., & Wang, S. Y. (2005). "Forecasting stock market movement direction with support vector machine." Computers & Operations Research 32(10): 2513-2522.

Hyndman, R. J. ve G. Athanasopoulos (2018). Forecasting: principles and practice, OTexts.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). Statistical learning. An Introduction to Statistical Learning, Springer: 15-57.

Kalles, D. ve C. Pierrakeas (2006). "Analyzing student performance in distance learning with genetic algorithms and decision trees." Applied Artificial Intelligence 20(8): 655-674.

Kandananond, K. (2011). "Forecasting electricity demand in Thailand with an artificial neural network approach." Energies 4(8): 1246-1257.

Kara, Y., Boyacioglu, M. A., & Baykan, Ö. K. (2011). "Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange." Expert Systems with Applications 38(5): 5311-5319.

Kass, G. V. (1980). "An exploratory technique for investigating large quantities of categorical data." Journal of the Royal Statistical Society: Series C (Applied Statistics) 29(2): 119-127.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems.

Khandelwal, I., Adhikari, R., & Verma, G. (2015). "Time series forecasting using hybrid ARIMA and ANN models based on DWT decomposition." Procedia Computer Science 48: 173-179.

Khashei, M. ve M. Bijari (2011). "A novel hybridization of artificial neural networks and ARIMA models for time series forecasting." Applied Soft Computing 11(2): 2664-2675.

Kim, K.-j. (2003). "Financial time series forecasting using support vector machines." Neurocomputing 55(1-2): 307-319.

Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2004). "PREDICTING STUDENTS'PERFORMANCE IN DISTANCE LEARNING USING MACHINE LEARNING TECHNIQUES." Applied Artificial Intelligence 18(5): 411-426.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). "Machine learning applications in cancer prognosis and prediction." Computational and structural biotechnology journal 13: 8-17.

Lantz, B. (2013). Machine Learning with R. Olton, UNITED KINGDOM, Packt Publishing Ltd.

Laplace, P., 1812, Théorie analytique des probabilités: Courcier.

Lee, R. ve T. Miller (2002). "An approach to forecasting health expenditures, with application to the US Medicare system." Health Services Research 37(5): 1365-1386.

Legendre, A. M., (1805). Nouvelles méthodes pour la détermination des orbites des comètes: Firmin Didot.

Leslie, P. H. (1945). "On the use of matrices in certain population mathematics." Biometrika: 183-212.

Li, S., Yang, Z., Li, H., & Shu, G. (2018). "Projection of population structure in China using least squares support vector machine in conjunction with a Leslie matrix model." Journal of Forecasting 37(2): 225-234.

Lin, W. Y., Hu, Y. H., & Tsai, C. F. (2012). "Machine learning in financial crisis prediction: a survey." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42(4): 421-436.

Loh, W. Y. (2011). "Classification and regression trees." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1(1): 14-23.

Mair, C., Kadoda, G., Lefley, M., Phalp, K., Schofield, C., Shepperd, M., & Webster, S. (2000). "An investigation of machine learning based prediction systems." Journal of Systems and Software 53(1): 23-29.

Markov, A. A., (1906). Extension of the law of large numbers to dependent quantities [in Russian]: Izvestiya Fiziko Matematicheskikh Obschestva Kazan University, 15,135-156.

Mathworks (n.d.). machinelearning_supervisedunsupervised.png. <https://de.mathworks.com/help/stats/machinelearning_supervisedunsupervised.

McCarthy, J., ve E. Feigenbaum, (1990), Arthur Samuel: Pioneer in machine learning: AI Magazine, 11, no. 3, 10–11.

Miller, T. H., Gallidabino, M. D., MacRae, J. I., Owen, S. F., Bury, N. R., & Barron, L. P. (2019). "Prediction of bioconcentration factors in fish and invertebrates using machine learning." Science of The Total Environment 648: 80-89.

Minaei-Bidgoli, B., Kashy, D. A., Kortemeyer, G., & Punch, W. F. (2003). Predicting student performance: an application of data mining methods with an educational web-based system. 33rd Annual Frontiers in Education, 2003. FIE 2003., IEEE.

Modaresi, F., Araghinejad, S., & Ebrahimi, K. (2018). "A Comparative Assessment of Artificial Neural Network, Generalized Regression Neural Network, Least-Square Support Vector Regression, and K-Nearest Neighbor Regression for Monthly Streamflow Forecasting in Linear and Nonlinear Conditions." Water Resources Management 32(1): 243-258.

Mehryar Mohri, A. R. (2012). Foundations of Machine Learning. Cambridge, UNITED STATES, MIT Press.

Moritz, S., Sardá, A., Bartz-Beielstein, T., Zaefferer, M., & Stork, J. (2015). Comparison of different methods for univariate time series imputation in R. arXiv preprint arXiv:1510.03924.

Olshansky, S. J., Passaro, D. J., Hershow, R. C., Layden, J., Carnes, B. A., Brody, J., ... & Ludwig, D. S. (2005). "A potential decline in life expectancy in the United States in the 21st century." New England Journal of Medicine 352(11): 1138-1145.

Pai, P.-F. and C.-S. Lin (2005). "A hybrid ARIMA and support vector machines model in stock price forecasting." Omega 33(6): 497-505. png>. [Erişim tarihi: 10.03.2019].

Pritchett, H. S. (1891). "A formula for predicting the population of the United States." Publications of the American Statistical Association 2(14): 278-286.

Qiokata, V. ve M. Khan (2015). Modeling emigration of Fiji's population using Artificial Neural Network. 2015 2nd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), IEEE.

Quinlan, J. R. (1986). "Induction of decision trees." Machine learning 1(1): 81-106.

Quinlan, J. R. (2014). C4. 5: programs for machine learning, Elsevier.

Raftery, A. E., Chunn, J. L., Gerland, P., & Ševčíková, H. (2013). "Bayesian probabilistic projections of life expectancy for all countries." Demography 50(3): 777-801.

Rahman, M. M. ve D. N. Davis (2013). Machine Learning-Based Missing Value Imputation Method for Clinical Datasets. IAENG Transactions on Engineering Technologies: Special Volume of the World Congress on Engineering 2012. G.-C. Yang, S.-l. Ao and L. Gelman. Dordrecht, Springer Netherlands: 245-257.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). "Cross-validation." Encyclopedia of database systems: 532-538.

Rish, I. (2001). An empirical study of the naive Bayes classifier. IJCAI 2001 workshop on empirical methods in artificial intelligence.

Sean J. Taylor & Benjamin Letham (2018) Forecasting at Scale, The American Statistician, 72:1, 37-45, DOI: 10.1080/00031305.2017.1380080

Shearer, C. (2000). "The CRISP-DM model: the new blueprint for data mining." Journal of data warehousing 5(4): 13-22.

Shin, K. S., Lee, T. S., & Kim, H. J. (2005). "An application of support vector machines in bankruptcy prediction model." Expert Systems with Applications 28(1): 127-135.

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., ... & Ray, T. S. (2002). "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning." Nature medicine 8(1): 68.

Shmueli, G. ve K. C. Lichtendahl (2016). Practical time series forecasting: A hands-on guide, Axelrod Schnall Publishers.

Shorter, F. C., Sendek, R., & Bayoumy, Y. (1995). Computational methods for population projections: with particular reference to development planning, HPN Technologies Inc.

Siegel, J.S. ve Swanson, D.A. (2004) The Methods and Materials of Demography. 2nd Edition, Elsevier Science & Technology, San Francisco, USA.

Sirkeci, I. (2017). "Turkey's refugees, Syrians and refugees from Turkey: a country of insecurity." Migration Letters 14(1): 127-144.

Smith, S. K., Tayman, J., & Swanson, D. A. (2001). State and Local Population Projections : Methodology and Analysis. Hingham, UNITED STATES, Kluwer Academic Publishers.

Smith, S. K., Tayman, J., & Swanson, D. A. (2006). State and local population projections: Methodology and analysis, Springer Science & Business Media.

Sprague, W. W. (2013). Wood's Method--a Method for Fitting Leslie Matrices from Age-Sex Population Data, with some Practical Applications, UC Berkeley.

Sundaram, R. B. (2018, September 06). Understanding the Math behind the XGBoost Algorithm. Retrieved March 02, 2019, from https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/

Susanto, F., Budi, S., De Souza, P., Engelke, U., & He, J. (2016). "Design of environmental sensor networks using evolutionary algorithms." IEEE Geoscience and Remote Sensing Letters 13(4): 575-579.

Suthar, B., Patel, H., & Goswami, A. (2012). "A survey: classification of imputation methods in data mining." International Journal of Emerging Technology and Advanced Engineering 2(1): 309-312.

Sutton, R. S. ve A. G. Barto (1998). Introduction to reinforcement learning, MIT press Cambridge.

Swanson, D. A., Hough, G. C., Rodriguez, J. A., & Clemans, C. (1998). "K-12 enrollment forecasting: merging methods and judgment." ERS spectrum 16(4): 24-31.

Swetapadma, A. ve A. Yadav (2017). "A Novel Decision Tree Regression-Based Fault Distance Estimation Scheme for Transmission Lines." IEEE Transactions on Power Delivery 32(1): 234-245.

Taylor, S. J. and Letham, B.(2018). Prophet: Forecasting at scale, https://facebook.github.io/prophet/, [Ziyaret Tarihi:15.04.2019].

Tayman, J., Parrott, B., & Carnevale, S. (1997). "Locating fire station sites: The response time component." RAND-PUBLICATIONS-MR-ALL SERIES-: 203-217.

Tekin, A. T., Ayhan, G., & Özkale, L. (2018). TURKISH AUTOMOTIVE INDUSTRY AND PREPARATION FOR INDUSTRY 4.0.

Tekin, A. T., Ozkale, N. L., & Oztaysi, B. (2019). Big Data Concept in Small and Medium Enterprises: How Big Data Effects Productivity. In Industrial Engineering in the Big Data Era (pp. 363-376). Springer, Cham.

Tekin, A. T., Özkale, L., & Ayhan, G. (2018). The Importance of R & D Investments in Information and Communication Technologies and Tax Policies Applied Through Information and Communication Technologies. In International R&D, Innovation and Technology Management Congress, In IRDITECH (pp. 155-166).

Thomas, J. R. and S. J. Clark (2011). "More on the cohort-component model of population projection in the context of HIV/AIDS: A Leslie matrix representation and new estimates." Demographic research 25: 39.

Thomas, R. K. (1997). "Using demographic analysis in health services planning: A case study in obstetrical services." RAND-PUBLICATIONS-MR-ALL SERIES-: 159-179.

Trafalis, T. B. ve H. Ince (2000). Support vector machine for regression and applications to financial forecasting. Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium, IEEE.

Turing, A., 1950, Computing machinery and intelligence: Mind, 59, 433–460.

Vandamme, J. P., Meskens, N., & Superby, J. F. (2007). "Predicting academic performance by data mining methods." Education Economics 15(4): 405-419.

Vapnik, V. (1982). Estimation of Dependences Based on Empirical Data. Springer, Berlin,

Whelpton, P. K. (1928). "Population of the United States, 1925 to 1975." American Journal of Sociology 34(2): 253-270.

Wolpert, D. H. ve W. G. Macready (1997). "No free lunch theorems for optimization." IEEE transactions on evolutionary computation 1(1): 67-82.

Ye, Q. H., Qin, L. X., Forgues, M., He, P., Kim, J. W., Peng, A. C., ... & Ma, Z. C. (2003). "Predicting hepatitis B virus–positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning." Nature medicine 9(4): 416.

Yip, S. C., Wong, K., Hew, W. P., Gan, M. T., Phan, R. C. W., & Tan, S. W. (2017). "Detection of energy theft and defective smart meters in smart grids using linear regression." International Journal of Electrical Power & Energy Systems 91: 230-240.

Yu, Q., Miche, Y., Séverin, E., & Lendasse, A. (2014). "Bankruptcy prediction using extreme learning machine and financial expertise." Neurocomputing 128: 296-302.

ÖZGEÇMİŞ

Ad-Soyad **: Fatih Veli ŞAHİNARSLAN**

Doğum Tarihi ve Yeri **: 07.10.1993 / İSTANBUL**

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

ÖĞRENİM DURUMU:

- **Lisans** **:**2015, Bahçeşehir Üniversitesi, Mühendislik Fakültesi, Endüstri Mühendisliği
- **Yükseklisans** **:**2019, İstanbul Teknik Üniversitesi, Sosyal Bilimler Enstitüsü, İşletme

MESLEKİ DENEYİM VE ÖDÜLLER:

- İstanbul Teknik Üniversitesi'nde İşletme Yüksek Lisans kazandı.
- Bahçeşehir Üniversitesi'nde %100 eğitim bursu kazandı.