BÜYÜK VERİ ARAÇLARI VE R KULLANARAK AMERİKAN HAVAYOLU FİRMALARININ SORUNLARININ KEŞFEDİLMESİ

Mustafa Vahit KESKİN

Doğan YILDIZ

Yıldız Teknik Üniversitesi, Fen Edebiyat Fakültesi, İstatistik Bölümü

Öz

Büyük veri sistemleri günümüzün büyük ölçekli veri analitiği ihtiyaçlarını karşılamaktadır. Bununla birlikte R yazılımının istatistiksel hesaplama ve veri görselleştirme gücü ile büyük veri araçlarının büyük ölçekli uygulamalar gerçekleştirebilme yetenekleri birleştiğinde başarılı analiz sistemleri ortaya çıkmaktadır. Çalışma kapsamında büyük veri araçları Apache Hadoop ve Apache Spark'a değinilmiş, disk bazlı çalışan MapReduce programlama modeli ile bellek içi çalışan Apache Spark'ın içyapısı arasındaki farklılıklara dikkat çekilmiştir. Makine öğrenmesi yaklaşımları ele alınmış ve denetimli öğrenme ile denetimsiz öğrenme metotları arasındaki farklılıklar ifade edilmiştir. Teorik olarak denetimsiz öğrenme yöntemlerinden kümeleme yöntemlerine, denetimli öğrenme yöntemlerinden karar ağaçlarına değinilmiştir. Bu analiz yöntemleri Amerika Birleşik Devletleri havayolu firmalarına ait 1987-2008 yılları verilerine uygulanmıştır. Havayolu şirketlerinin uçuş mesafeleri ve uçuş gecikme performanslarına yönelik kümeleme analizi yapılmıştır. Spesifik olarak bir havayolu şirketinin gecikme sürelerine ilişkin çıkarımlar karar ağacı kullanılarak yapılmıştır. Ana akım havayolu firmaları, analiz sonuçlarından hareketle en sorunlu kümeyi oluşturmuştur. Bu sorunların özellikle kalkıştaki gecikme ve mesafeden kaynaklandığı gerçeği, kümeleme analizi ve karar ağaçları sonuçlarından ortaya çıkmıştır.

Anahtar Sözcükler: *Büyük Veri, Makine Öğrenmesi, Apache Hadoop, Apache Spark, Kümeleme Analizi, Karar Ağaçları*

DISCOVERING THE PROBLEMS OF THE AMERICAN AIRLINE COMPANIES BY USING R AND BIG DATA TOOLS

Abstract

Big data systems solve today's large scale data analytics needs. In addition, succesfull analysis systems emerge when we combine statistical power of R programming and data visualization with power of making big scalability applications of big data tools. In the scope of the study, the major big data tools Apache Hadoop and Apache Spark are mentioned, the difference between the disk based MapReduce programming model and the in-memory Apache Spark's internal structure is highlighted. Machine learning approaches are discussed and the differences between supervised learning and unsupervised learning methods have been stated. Theoretically, clustering methods of unsupervised learning approaches, decision trees of supervised learning approaches are mentioned. These analysis methods were applied to the data of the United States airline companies between 1987-2008. Cluster analysis was conducted for airline companies flight distances and flight delays. Specifically, the conclusions regarding the delays of an airline company were made using the decision tree. The mainstream airline companies formed the most problematic cluster with the result of the analysis. The fact that these problems are resulted from take off delay and distance is revealed by results of clustering analysis and decision trees.

Keywords: *Big Data, Machine Learning, Apache Hadoop, Apache Spark, Cluster Analysis, Decision Trees*

GİRİŞ

Havayolu endüstrisinde, havayolu firmalarının en çok sorun yaşadığı bölüm, uçakların zamanında kapılara (gates) ulaşmasıdır. Günümüzde havayolu firmaları en büyük çalışmalarını zamanlama performansının kalitesini arttırmak üzerine kurmuştur. İnsan ve bagaj taşıma kurallarına rağmen, yolcuların ve havayolu firmalarının zamanından tasarruf etme ihtiyacı öncelik olarak belirlenmiştir. Bu çalışmadaki ana hedefimiz Amerika Birleşik Devletleri içinde 1987-2008 yılları arasında çalışma yürütmüş olan havayolları firmaları verileri üzerinden, havayolu sektörü üzerindeki sorunların neler olabileceğine dair genel bir öngörü sağlamak, özel ve değerli ilişkileri bulmak, keşfetmektir. Buna bağlı olarak bu çözümlemelerin ve keşiflerin yapılması için kümeleme (clustering) ve karar ağaçları (decision trees) algoritmalarından yararlanılmıştır.

Büyük veri analiz sistemlerinde yapılan ve bu çalışmaya katkı sağlayan bilgi ve literatür çalışması bu bölümde verilmiştir. Yengi yüksek lisans tez çalışmasında büyük veride duygu analizine dayalı öneri sistemi uygulaması gerçekleştirmiştir (Yengi, 2016). Salur Apache Hadoop kullanarak veri madenciliği uygulaması çalışmıştır (Salur, 2016). Özdeş büyük veri araçları ile duygu analizi çalışmıştır (Özdeş, 2017). Akgün Apache Spark ile akan veride sınıflandırma üzerine bir uygulama gerçekleştirmiştir (Akgün, 2016). Çetinkaya Apache Hadoop kullanarak hızlı tüketim sektörüne yönelik uygulamalar yapmıştır (Çetinkaya, 2016). Hallaç büyük veri araçlarını kullanarak büyük veride makine öğrenmesi algoritmaların ile ilgili uygulamalar yapmıştır (Hallaç, 2014).

Çalışmanın birinci bölümü büyük veri ve ilgili yazılımların genel bilgi ve kavramsal çerçevesinden oluşmaktadır. Çalışmanın sonraki bölümünde veri setinin nasıl oluşturulduğu ve içeriği hakkında bilgiler verilmiştir. Kümeleme ve karar ağaçları algoritması hakkında detaylı bilgilerle beraber analiz sonuçlarına dair açıklamalar bu bölümü oluşturmaktadır. Çalışmanın son bölümünde analiz sonuçları, kavramsal çerçeve dikkate alınarak sorunlar ortaya konulmuş ve analizlerini havayolu sektörüne olan etkileri incelenmiştir.

KAVRAMSAL ÇERÇEVE

Büyük Veri Nedir?

Geleneksel yöntemler ile işlenemeyen verilere büyük veri denir. Ortaya çıkan veri işleme güçlüğü verinin boyutu, çeşitliliği ve hızı ile alakalıdır. Büyük veriyi ifade eden özellikler hacim, çeşitlilik ve hızdır (De Mauro ve Grimaldi, 2015:97). Büyük verinin tanımlanmasında büyük verinin bileşenleri de kullanılabilmektedir. Büyük verinin bileşenleri bazı kaynaklarda 3V olarak bazı kaynaklarda 5V olarak geçer (SAS, What is big data, 2015).

Büyük Verinin Bileşenleri

Hacim: Veri boyutunun geleneksel yaklaşımlar ile işlenemeyecek boyutta olması. Verinin hacminin artması ile veri odaklı uygulamalar istenilen zamanlarda istenilen performansları sağlayamamaya başlamıştır.
Hız: Büyük boyutun hızla artması.

Çeşitlilik: Farklı türlerde veri yapılarından meydana gelmesi.

Büyük Veri Analitiği Türleri

İş dünyasında uygulama alanı günden güne artan büyük veri teknolojileri kullanılarak çeşitli analitik yaklaşımlar sergilenmektedir. Bu yaklaşımlar betimleyici analitik, teşhis analitiği, öngörüsel analitik ve reçeteli analitiktir (SAS, What is big data, 2015).

Betimleyici analitik

Araştırma süreci "Ne oldu?" sorusuna yanıt aramaktadır. Veriyi resmetmek genel durumunu ortaya koymayı ifade etmektedir.

Teşhis analitiği

Araştırma süreci "Nasıl oldu?" sorusuna yanıt arar. Veri içerisinde oraya çıkan yapıların sebepleri araştırılır.

Öngörüsel analitik

Araştırma süreci "Ne olacak?" sorusu ile ilgilenir. Gelecek ile ilgili tahminlerde bulunmak ya da gözlemlenen olayları meydana getiren koşulların anlaşılarak henüz gerçekleşmemiş olaylarla alakalı çıkarımlarda bulunmak ile ilgilenilir. Borsadaki hisse senetlerinin gelecek değerlerinin tahmin edilmesi, belirli özelliklere sahip bir aracın fiyatının ne olabileceği ve belirli bir yaş grubunun hangi ürünleri satın alacağının tahmin edilmesi gibi problemler ile ilgilenilir.

Reçeteli analitik

Veri seti üzerinden gelecek ile ilgili ne olacağının belirlenmesinin ardından gerçekleşebilecek olaylara karşı ne tür eylemler yapılabileceğine yönelik bilgiler sunar. Örneğin bir müşterinin marka kullanımını bırakabileceğinin tahmin edilmesinden sonra marka kullanımının bırakılmaması adına müşteriye ne tür yaklaşılması gerektiği gibi.

Büyük Veri Araçları

Apache Hadoop

Apache Hadoop açık kaynak kodlu, güvenilir, ölçeklenebilir paralel hesaplama yazılımı projesidir (Apache Hadoop, 2018). Büyük veri teknolojilerinin temelini oluşturan bu yazılım geleneksel yöntemler ile efektif olarak işlenmesi mümkün olmayan verilerin işlenebilmesine olanak

sağlamaktadır. Apache Hadoop bir bilgisayar kümesinin belirli bir işi yapmak için tek bir bilgisayar gibi birlikte hareket etmesini sağlamaktadır.

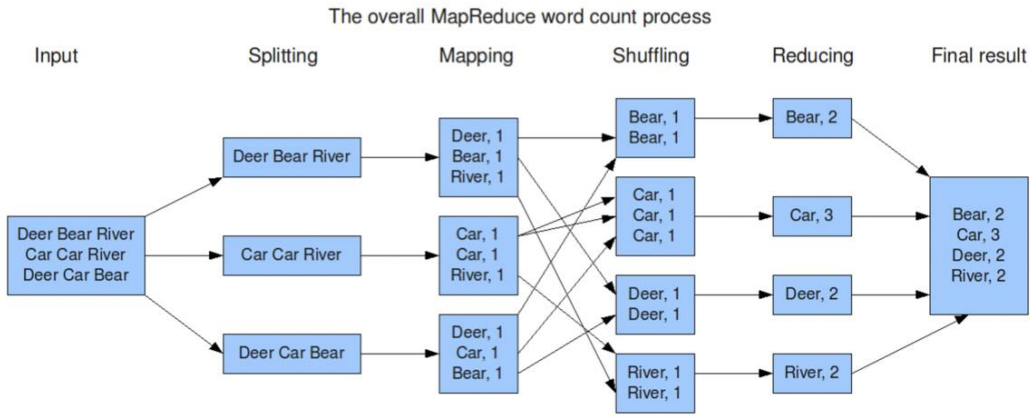
Apache Hadoop'un temel yapıları

Apache Hadoop, Hadoop'un temel gerekliliklerini barındıran Hadoop Common, Hadoop'un dosya sistemini ifade eden Hadoop Dağıtık Dosya Sistemi (HDFS), kaynak yönetimi için kullanılan Hadoop YARN ve dağıtık hesaplama paradigmasını uygulama anlamında hayata geçiren Hadoop MapReduce yapılarından meydana gelmektedir (Apache Hadoop, 2018).

Hadoop MapReduce

Büyük veride dağıtık hesaplama yapabilmek için kullanılan MapReduce bir programlama yöntemidir. Bilgisayarlardan oluşan küme üzerinde belirli bir amaç için yapılan görev paylaşımının gerçekleştirilmesi için MapReduce kullanılır (Gonzolez, 2016). Map ve Reduce aşamalarından oluşan bu süreçte dağıtık olarak tutulan veri Map aşamasında işlenir ve Reduce aşamasında tüm bilgisayarların yaptığı işlemler toparlanarak değerlendirilir.

Şekil 1: MapReduce kelime sayma örnek uygulaması (Stackoverflow, 2018)



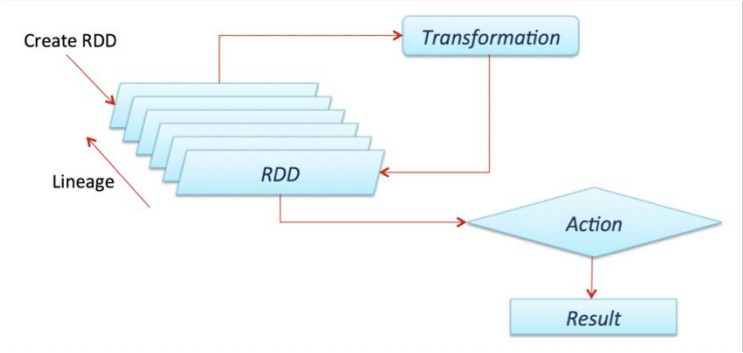
Apache Spark

Apache Spark Apache Hadoop'a benzer şekilde bilgisayarlardan oluşan bir küme üzerinde dağıtık olarak genel amaçlı ve hızlı bilgi işleme yapabilmek için geliştirilmiştir. MapReduce modelinde yer alan disk bazlı çalışma sisteminin yarattığı maliyetlerden dolayı ortaya çıkmıştır. Bellek içi veri işleyebilmesiyle iteratif işlemler gerektiren büyük ölçekli uygulamalarda Apache Hadoop'a göre 100 kat daha hızlı çalışmaktadır (Apache Spark, 2018).

Apache Spark temel bileşenleri ve RDD (Spark Core ve Resilient Distributed Datasets)

Bellek içi veri işleme işlemi RDD (Resilient Distributed Datasets) aracılığı ile gerçekleşir. Bellek içi veri işleme üç aşamada gerçekleşir: RDD oluşturulur, dönüştürülür ve aksiyon alınır. Fiziki diskte yer alan veri RDD basamağında geçici belleğe taşınır. Oluşturulan RDD'ler üzerinden yeni RDD'ler oluşturulabilir. Aksiyon komutları verilene kadar gerçek bir hesaplama yapılmaz ve ancak aksiyon komutları geldiğinde hesaplama yapılır (Zaharia, 2013).

Şekil 2: Spark tembel (lazy) çalışma yapısı



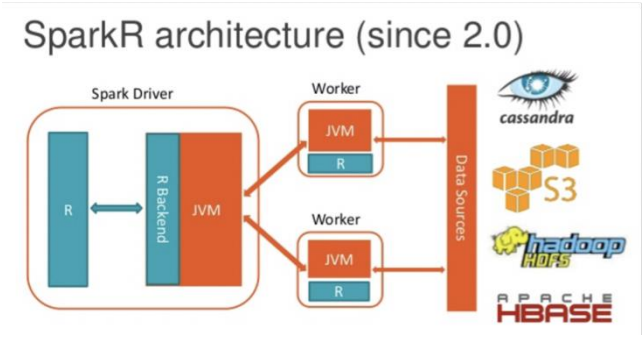
R programalama dili

R açık kaynak kodlu istatistiksel hesaplama ve görselleştirme aracıdır (R-project, 2018). R dili son yıllarda veri bilimi ve büyük veri dünyasında en sık kullanılan araçlardan birisi olmuştur. Zengin kütüphaneleri ve büyük bir topluluk tarafından desteklenmesi R programlama dilini vazgeçilmez hale getirmiştir. R programlama dilinin tek başına büyük veri setleri ile kullanımında performans kayıpları meydana gelebilmektedir. Verinin boyutunun büyümesi ile kişisel bilgisayarlarda R ile veri analizleri zorlaşmıştır. Büyük veri teknolojilerinde veri analizi ve istatistiksel hesaplamaların gerçekleştirilmesi de oldukça zordur. Büyük miktardaki verilerin işlenmesine olanak sağlayan büyük veri yazılımları ile istatistiksel hesaplamalar ve veri analizi uygulamalarında dünyanın en sık kullanılan araçlarından olan R programlama dilinin bir arada kullanılması heyecan verici bir gelişmedir. Büyük veri teknolojileri ve R programlama dilinin bir arada kullanılmasına olanak sağlayan birçok kütüphane duyurulmuştur. Çalışma kapsamında Apache Spark tarafından geliştirilen SparkR ve R-Studio tarafından geliştirilen sparklyr kütüphanelerine yer verilecektir.

SparkR

SparkR, Apache Spark kullanımı için arayüz sağlayan bir R kütüphanesidir. Spark 2.2.1 versiyonundan sonra büyük veri setleri üzerinde dağıtık veri çerçevesi işlemleri yapılmasına olanak sağlamıştır (R veri çerçevesi ve "dplyr" işlemleri gibi). Ayrıca SparkR, MLlib kullanarak dağıtık makine öğrenmesi uygulamalarının yapılabilmesine de olanak sağlamaktadır (Apache Spark, 2018).

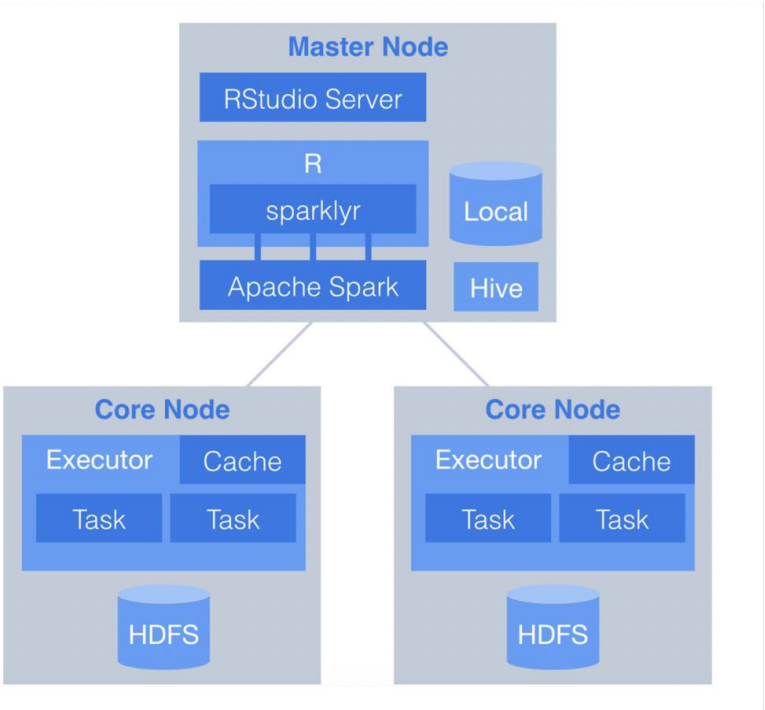
Şekil 3: SparkR Mimarisi



sparklyr

R-Studio'nun geliştirdiği sparklyr, Apache Spark için arayüz sağlayan bir kütüphanedir. SparkR'a benzer olarak Apache Spark üzerinde dağıtık hesaplamalar yapmaya imkan sağlar (R-studio, 2018). Ana düğüm noktasına kurulan R-Studio Server ve sparklyr Apache Spark ile entegre edilerek R üzerinden bilgisayar kümesinde Spark temelli işlemler yapılır.

Şekil 4: sparklyr mimarisi (R-studio, 2018)



Makine Öğrenmesi Nedir?

Yapay zeka altında bir çalışma alanı olan makine öğrenmesi, öğrenme ile ve öğrenme yaklaşımlarının optimizasyonu ile ilgilenir (Alpaydın, 2010). Makine öğrenmesi algoritmalarına kendilerine verilen örnek veri setleri içerisindeki yapıları öğrenerek genelleme yapabilme yeteneği kazandırılmaya çalışılır (Shalev-Shwartz ve Ben-David, 2014).

Öğrenme çeşitleri

Gözetimli öğrenme

Veri seti içerisinde bağımlı ve bağımsız değişkenlerin bir arada yer aldığı dolayısıyla hangi girdi değerlerine karşılık hangi çıktı değerlerinin oluştuğunun belli olduğu durumlarda kullanılan öğrenme yöntemdir. Etiketli veri olarak adlandırılan bu veri setleri üzerinde algoritmalar girdiler ile çıktılar arasındaki ilişkiyi tanımlamaya çalışır (Mohri ve diğerleri, 2012). Regresyon modelleri, karar ağaçları, topluluk öğrenme yöntemleri, yapay sinir ağları, destek vektör makineleri gözetimli öğrenme yöntemlerine örnek olarak verilebilir. Çalışma kapsamında karar ağaçlarına yer verilmiştir.

Gözetimsiz öğrenme

Etiketli verinin olmadığı veri setleri üzerinde gerçekleştirilen öğrenme biçimidir. Sadece girdi değerleri olduğu için bu tür veriler üzerinde değerlendirme yapmak zordur. Girdi değerleri üzerinden çıkarımlar yapılmaya çalışılır, boyut indirgeme, çok boyutlu ölçekleme ve kümeleme yöntemleri gözetimsiz öğrenme yöntemlerindendir (Alpaydın, 2010).

MATERYAL VE METOD

Çalışmanın Amacı

Çalışmada havacılık sektöründe sık karşılaşılan problemlerden birisi olan rötarlar bu rötarların sebeplerine ilişkin çıkarımlarda bulunmaya çalışılmıştır. Buna yönelik olan hava yolu şirketleri için kümeleme analizi uygulanmış ve belirli bir hava yolu şirketi için karar ağacı ile gecikme süresi değişkenine yönelik incelemelerde bulunulmuştur.

Veri Seti

ABD Ulaştırma Bakanlığı (DOT) Ulaşım İstatistikleri Bürosu (BTS) büyük hava yolu şirketlerinin iç hat uçuşlarının zamanlama konusundaki performanslarına yönelik istatistikler tutmuştur. Zamanında, gecikmeli, iptal edilmiş ya da yönlendirilmiş uçuşlara ait istatistikler düzenli olarak yayınlanmaktadır. 2003'ten bu yana uçuşlar ve uçuşların gecikme nedenlerine yönelik veriler düzenli olarak tutulmuş ve bir süre sonra halka açık hale getirilmiştir. Uygulama kapsamında 2008 yılına ait veriler kullanılmıştır.

Tablo 1: Flights veri setinin değişkenleri

	DEĞİŞKENLER	**TANIM**
1	Year	2008
2	Month	1-12
3	DayofMonth	1-31

4	DayOfWeek	1(Pazartesi)- 7(Pazar)
5	DepTime	gerçek kalkış zamanı
6	CRSDepTime	planlanan kalkış zamanı
7	ArrTime	gerçek varış zamanı
8	CRSArrTime	planlanan varış zamanı
9	UniqueCarrier	hava yolu şirket isimlerinin kısaltmaları
10	FlightNum	uçuş numarası
11	TailNum	uçak kuyruk numarası
12	ActualElapsedTime	uçuş süresi (dk)
13	CRSElapsedTime	planlanan uçuş süresi(dk)
14	AirTime	uçuş zamanı
15	ArrDelay	varış geçikmesi(dk)
16	DepDelay	kaçış gecikmesi(dk)
17	Origin	kalkış noktası uluslararası havaalanı kısaltma kodu
18	Dest	varış noktası uluslararası havaalanı kısaltma kodu
19	Distance	mesafe (mil)
20	TaxiIn	taksi bulma süresi (dk, varış lokasyonunda)
21	TaxiOut	taksi bulma süresi (dk, kalkış lokasyonunda)
22	Cancelled	uçuş iptal edildi mi? (1=evet)
23	CancellationCode	uçuş iptal sebebi: (A: havayolu şirketi, B: hava şartları, C: NAS, D: güvenlik)
24	Diverted	yönlendirme oldu mu? (1=evet)
25	CarrierDelay	havayolu şirketi kaynaklı gecikme süresi (dk)
26	WeatherDelay	hava şartları kaynaklı gecikme süresi (dk)
27	NASDelay	NAS kaynaklı gecikme süresi (dk)
28	SecurityDelay	güvenlik kaynaklı geçikme süresi (dk)
29	LateAircraftDelay	geç gelen uçak gecikmesi (dk)

Uygulanacak iki farklı analiz yaklaşımında veri iki farklı şekilde kullanılacaktır. Gerçekleştirilen kümele analizi için her bir hava yolu şirketine yönelik olarak büyük veri seti içerisinden hava yolu şirketlerine ait kalkış gecikme süreleri ortalamaları ve yapılan uçuşların mesafelerinin ortalamaları elde edilmiştir. Oluşturulan iki değişken üzerinden hava yolu şirketlerinin kümelenmeleri incelenmiştir. Karar ağacı yönteminde ise American Airlines şirketine ait veriler kullanılmış ve gecikme sürelerine ilişkin çıkarımlar karar ağacı yapısı üzerinden değerlendirilmiştir. Bu analizde kalkış gecikme süresi (DepDelay), planlanan kalkış zamanı (CRSDepTime), gün (dayofweek) ve uzaklık değişkeni (distance) kullanılmıştır. Bu değişkenler bağımlı değişkeni açıklamada anlamlı olan değişkenlerdir. Uygulanan karar ağacı yönteminde anlamlılıklarını ifade eden p-value değerleri ifade edilmiştir.

K-Means Kümeleme (K-means Clustering)

Hiyerarşik olmayan kümeleme (gruplandırma) yöntemleri içinde en çok kullanılan metod K-ortalamalar (K-means) yöntemidir. Bu yöntemle, küme (grup) içerisindeki kareler-toplamını minimum olacak şekilde k sayıda kümeye (gruba) bölmek amaçlanmaktadır.

Bu yöntemde küme sayısı; en az küme sayısı 2, en fazla küme sayısı ise gözlem sayısına eşit veya daha az olacak şekilde hesaplanır. K-means (ortalamalar) metodunun amacı, gözlemleri, sayısı araştırmacı tarafından belirlenen kümelere sınıflamaktır. Böylece, k-means metod yöntemleri yardımıyla birimler, kümeler arasındaki değişkenlik en büyük, kümeler ya da gruplar içi değişkenlik en küçük olacak şekilde farklı kümelere gönderilir. Bu yöntemle her bir birimin (gözlemin) ya da kişinin en yakın merkezli (ortalama) kümeye (gruba) gönderilmesi amaçlanmıştır. *K*-ortalama metodu, gözlemleri (birimleri) kümelerin önceden belirlediğimiz sayısına (araştırmacı tarafından) göre küme durumuna getirilerek analizi başlatır. Böylece herbiri tek gözlemden (birimden) oluşan *k* adet küme (grup) ile analize başlanır ve herbir yeni birim en yakın ortalamalı kümeye eklenerek olmaları gereken kümeler hesaplanır. Kümeye yeni bir birim eklendikten sonra grup ortalaması yeniden hesaplanır ve süreç yanı şekilde işlemeye devam eder. Bu süreç tüm gözlemler gruplara hesaplanıncaya kadar süreç devam eder. Bütün birimler kümelere girdikten sonra ait oldukları küme ortalamasından daha yakın küme ortalaması varsa, birimlerin kümeleri değiştirilerek daha güvenilir (yakın merkezli) kümeler oluşturulmalıdır.

Tablo 2: K-means kümeleme analizi algoritma mantığı

K-Means Kümeleme Analizi İçin Kullanılan Algoritma Mantığı
Sınıf merkezlerinin belirlenmesi
Örneklerin mesafelere göre sınıflandırılması
Yapılan sınıflandırma sonrasında yeni merkezlerin belirlenmesi
Kümeleme işlemi güvenilir hale gelinceye kadar bir önceki iki adımın tekrarlanması

Genel olarak k-means yönteminin amacı diğer kümeleme yöntemlerinde olduğu gibi, gerçekleştirilen kümeleme analizi işlemi sonucunda elde edilen kümelerin, küme içi benzerliklerinin maksimum, kümeler arası benzerliklerinin ise minimum olmasını sağlamaktır. Küme benzerliği (similarity) tanımı, kümenin ağırlık-merkezi olarak alınan bir gözlemle kümedeki (gruptaki) diğer gözlemler arasındaki mesafelerin ortalama değeri ile hesaplanmaktadır (Zaharia, 2013).

K- ortalamalar uygulaması

K-means (ortalamalar) kümeleme analizi yapılarak elde edilen çıktılar dikkate alındığında, özellikle ABD içi büyük hava yolu şirketlerinin iç hat uçuşlarının zamanlama konusundaki performansları için kalkıştaki gecikme ve ilgili uzaklık ortalamasının en etkili değişkenler olduğu göze çarpmaktadır. Bu değişkenlerden benzerlik (similarity) ya da benzemezlik (dissimilarity) ölçütleri dikkate alınarak değerlendirme yapıldığında ABD Havayollarının üç farklı küme seçiminde (k=3) en iyi sonucu verdiği görülmektedir. Bu sonuçlardan hareketle uzaklık ortalamaları ve kalkıştaki gecikme değişkenlerinin şirketler üzerinde oluşturduğu 3 küme için birçok değerlendirme söz konusudur.

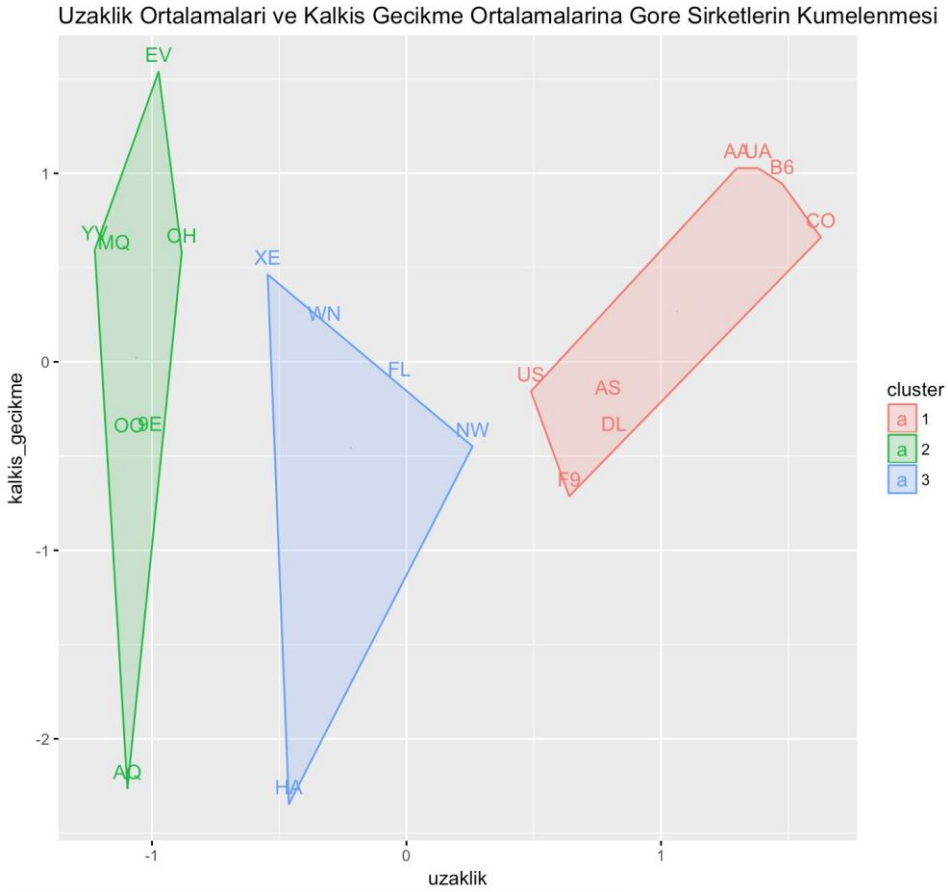
Tablo 3: k-means sonuçları

1.KÜME (CLUSTER)		**2.KÜME (CLUSTER)**		**3.KÜME (CLUSTER)**	
Havayolu Firma Kodu(IATA)	**Firma Adı**	**Havayolu Firma Kodu(IATA)**	**Firma Adı**	**Havayolu Firma Kodu(IATA)**	**Firma Adı**
AA	American Airlines	EV	ExpressJet	XE	JetSuiteX
UA	United Airlines	YV	Mesa Airlines	WN	Southwest Airlines
B6	JetBlue Airways	MQ	Envoy Air	FL	AirTran Airways
CO	Cobaltair Ltd	CH	Bemidji Airlines	NW	Northwest Airlines
AS	Alaska Airlines	OO	SkyWest Airlines	HA	Hawaiian Airlines
DL	Delta Air Lines	9E	Endeavor Air		
US	-	AQ	9 Air Co Ltd		
F9	Frontier Airlines				

ABD havayolları firmalarının oluşturduğu kümeler ve bu küme grafiği üstteki tablolar ve şekilde verilmiştir. Bu sonuçlardan hareketle 1.kümeyi oluşturan havayollarının ABD içinde ana akım

havayolları firmalarından olduğu rahatlıkla görülmektedir. 2. ve 3. Kümeyi oluşturan havayolları ise genellikler ulusal veya kargo firmalarından oluşmaktadır. Kümeleme analizinde grafikte merkeze yaklaştıkça (centroid yöntemi) havayolu firmalarının mesafe ve kalkıştaki gecikme sorunlarına duyarlılığının derecesinin giderek azaldığı söylenebilir. Yani firmalar merkeze yaklaştıkça sorunları daha da küçülmektedir. Küme içi homojenlik ve kümeler arası heterojenliğin sağlandığı kümeleme analizi grafiğinden görülebilir.

Şekil 5: k-means görsel çıktısı



Birinci kümeyi oluşturan firmalar daha uzun mesafe harcayan ve kalkıştaki gecikme süreleri en fazla olan şirketlerdir. Bu kümede şirketler için kalkıştaki gecikme süreleri önemli bir seviyede sorun olarak görünmektedir. Gecikme ve kalkış ölçütünün ABD içinde özellikle ana akım havayolu firmalarının ortak sorunu olduğu rahatlıkla söylenebilir. Küme içinde sorunu en fazla yaşayan firmalar ise American Airlines, United Airlines ve JetBlue Airways firmalarıdır. Uzaklık ölçütü en çok birinci kümeyi etkilemektedir.

İkinci kümeyi oluşturan havayolları firmaları ExpressJet, Mesa Airlines, Envoy Air, Bemidji Airlines, SkyWest Airlines, Endeavor Air ve 9 Air Co Ltd şirketleridir. Bu kümede mesafe ölçütünün değişimi kümeyi oluşturan bütün birimler için aynı seviyededir. Bu kümede önem arz eden değişkenlik ise kalkıştaki gecikme sürelerindedir. Birinci küme ile benzerliği de bu değişimlerdir. ExpressJet firması kalkıştaki gecikme küme içinde en çok etkilenen havayolu firmasıdır.

Üçüncü küme analiz içinde sorunları en az olan havayolu firmalarını temsil etmektedir. Bu firmalar ise JetSuiteX, Southwest Airlines, AirTran Airways, Northwest Airlines ve Hawaiian Airlines şirketleridir. Bu şirketler analiz sonuçlarından da görüldüğü gibi merkeze en yakın kümedir. Diğer kümelerle karşılaştırıldığında bu firmaların genellikle kısa mesafe yolcu ve kargo uçuşları yaptığı, bu nedenlerle kalkıştaki gecikme ve mesafe sorunlarından en az derecede etkilenen küme olduğu görülmektedir.

Karar Ağaçları (Decision Trees)

Karar ağacı, karar ve olası sonuçları ortaya koymak için ağaç (tree) benzeri bir şekil ve ya model kullanan bir karar alma sistemidir. Bir algoritmayı ortaya koymanın görsel bir şeklidir. Karar ağaçları çalışmalarında, ana amaç sonuca ulaşılabilecek olası yolların bir haritasının çıkarılmasıdır(Hallaç, 2014).

Karar ağaçları karar vermede kullanılan bir metottur. Karar ağacı, kararlar ve karar alma işlemini görsel (visual) bir şekilde ortaya koyar. Veri madenciliğinde (data mining) ise bu tip analiz datayı tanımak için kullanılır. Diğer yandan veri madenciliğinde kararları açıklamaz. Karar ağaçları akış diyagramı-benzeri yapılar olduğu söylenebilir. Kararları organize etmek ve açıkça göstermek için kullanılabilir. Karar ağacı analizinde görsel (visual) ve içsel kullanım ve anlaşılma artısı bulunur.

Tablo 4: k-means sonuçları Karar ağacı algoritması birimleri (Greco ve Girmaldi, 2015)

Karar Ağacı Algoritması Birimleri
1. İç düğüm: Bu düğüm bir özelliği değerlendirir ya da test eder.
2. Dal: İç düğümdeki değerlendirmenin sonucunu gösterir.
3. Uç düğüm ya da sonuç düğümü: Ele alınacak kararı gösterir.

Karar verme süreci ağaç (tree) analizinin başlangıcında ilk adımını atar, iç düğümde belirlenen bir nitelik ele alınır, ele alınan değerlendirme sonucu bir sonraki dala geçilir. Seçilen dal başka bir değerlendirme düğümüne götürür ve bu süreç son düğüme ulaşılıncaya kadar devam eder. Bu yolla değerlendirilen farklı özelliklerin değerleri farklı kararlara yol gösterir (Greco ve Girmaldi, 2015).
Karar ağaçları analizlerinde en çok yer alan algoritmalar(ID3, C4.5, C5.0, CART, CHAID ve QUEST) grafikte verilmiştir. Bu algoritmalar ve özellikleri ve kullanım alanlarıyla beraber verilen grafikten incelenebilir.

Tablo 5: Karar ağaçlarında en çok kullanılan algoritmalar

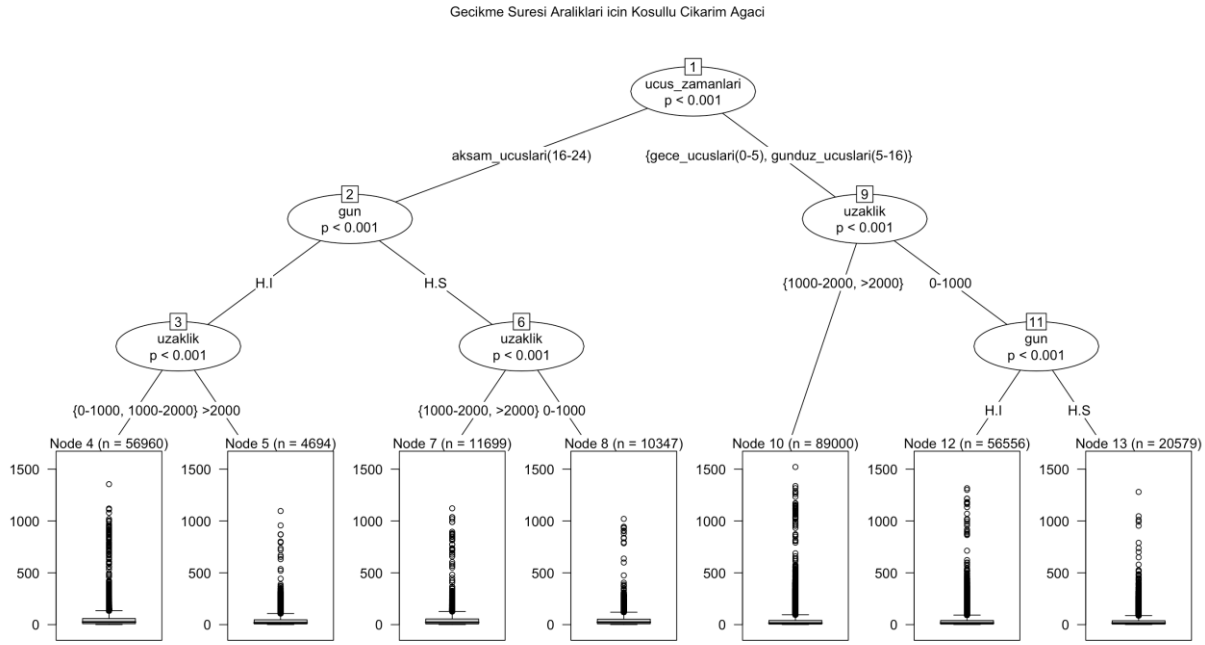
KARAR AĞACI ALGORİTMASI	ÖZELLİKLER
C&RT	Gini'ye dayalı ikili bölme işlemi mevcuttur. Son veya uç olmayan her bir düğümde iki adet dal bulunmaktadır. Budama işlemi ağacın karmaşıklık ölçüsüne dayanır. Sınıflandırma ve regresyonu destekleyici bir yapıdadır. Sürekli hedef değişkenleri ile çalışır. Verinin hazırlanmasına gereksinim duyar.
C4.5 ve **C5.0** (ID3 karar ağacı algoritmasının ileri versiyonları)	Her düğümden çıkan çoklu dallar ile ağaç oluşturur. Dalların sayısı tahmin edicinin kategori sayısına eşittir. Tek bir sınıflayıcı da birden çok karar ağacını birleştirir. Ayırma işlemi için bilgi kazancı kullanır. Budama işlemi her yapraktaki hata oranına dayanır.
CHAID (Chi-Squared Automatic Interaction Detector)	Ki-kare testleri kullanarak bölme işlemini gerçekleştirir. Dalların sayısı iki ile tahmin edicinin kategori sayısı arasında değişir.
SLIQ (Supervised Learning in Quest)	Hızlı ölçeklenebilir bir sınıflayıcıdır. Hızlı ağaç budama algoritması mevcuttur.
SPRINT (Scalable Parallelizable Induction of Decision Tree)	Büyük veri kümeleri için idealdir. Bölme işlemi tek bir niteliğin değerine dayanır. Tüm bellek sınırlamaları üzerinde nitelik listesi veri yapısı kullanarak işlem yapar.

Karar ağacı uygulaması

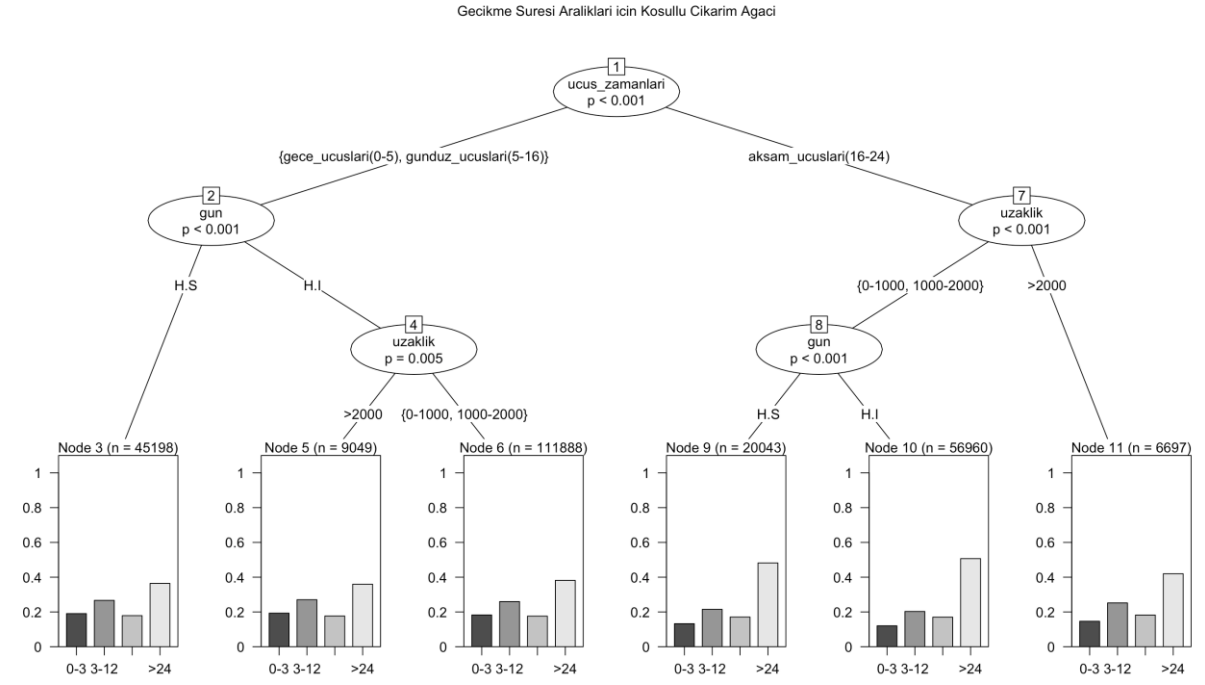
Çalışma kapsamında American Airlines hava yolu şirketinin uçuşlarına ait verilere odaklanılıp gecikme sürelerini oluşturan yapılar incelenmek istenmiştir. Karar ağacı yöntemi ile etkin çalışabilmek adına değişkenlerde bazı dönüşümler yapılmıştır. Hedef değişken (bağımlı değişken) olan gecikme süresi değişkeni "0-3","3-12","12-24" ve ">24" olacak şekilde kategorik değişkene dönüştürülmüştür. Planlanan kalkış zamanı değişkeni (crsdeptime) uçuş zamanları olarak adlandırılmış "gece_ucusları(0-5)", "gunduz_ucuslari(5-16)" ve "aksam_ucuslari(16-24)" şeklinde değişken dönüşümü sağlanmıştır. Uçulacak uzaklığı ifade eden uzaklık değişkeni (distance) "0-1000", "1000-2000" ve ">2000" şeklinde dönüştürülmüştür. Gün değişkeni (dayofweek) hafta içi (H.I) ve hafta sonunu (H.S) ifade edecek şekilde dönüştürülmüştür.

Dönüştürme işlemleri sonrasında American Airlines hava yolu şirketinin gecikme sürelerinin oluşma yapısı uçuş zamanları, uzaklık, gun değişkenlerince incelenmiştir. Analiz için "party" kütüphanesinde bulunan koşullu çıkarım ağacı fonksiyonu "ctree" fonksiyonu kullanılmıştır. Nominal, ordinal, sürekli ya da sensörlü verilere uygulanabilen parametrik olmayan bir ağaç yöntemi olan koşullu çıkarım ağacı yöntemi hedef değişkeni oluşturan yapıların görülmesinde kolaylık sağlamaktadır. Bağımlı değişken olan gecikme süresi değişkeni öncelikle sürekli değişken olarak incelenmiş ve aşağıdaki sonuçlar elde edilmiştir (Şekil 6). Bu sonuçlar bağımlı değişken içerisinde yapının okunması güç olduğundan sürekli değişken kategorik değişkene çevrilerek gecikme süreleri aralıklara ayrılmış ve böylece ikinci görseldeki (Şekil 7) gibi gecikme süreleri dağılımının daha net incelenebilmesi sağlanmıştır.

Şekil 6: Karar ağacı çıktısı(1)



Şekil 7: Karar ağacı çıktısı(2)



Gecikme süresi aralıkları için oluşturulan koşullu çıkarım ağacı sonuçları incelendiğinde bütün değişkenlerin anlamlı olduğu ve en önemli değişkenin uçuş zamanları değişkeni olduğu

görülmektedir. Akşam uçuşları ve diğer uçuşlar olarak iki ayrı dala ayrılmıştır. Akşam uçuşları içerisinde önce uzaklığın sonra uçuş zamanının (hafta içi, hafta sonu) etkili olduğu görülmektedir. Gecikme süreleri dağılımı akşam uçuşları bazında incelendiğinde 2000 milden kısa olan yolculuklarda 2000 milden uzun olan yolculuklara göre 24 saatten fazla gecikme vakası olması oranının daha fazla olduğu gözlenmektedir. Diğer uçuş zamanları (gece_ucuslari ve gunduz ucuslari) incelendiğinde ucuçların hafta içi ya da hafta sonu olmasını ifade eden gün değişkenin uzaklık değişkenine göre daha önce olduğu görülüyor. İki dal karşılaştırıldığında akşam saatlerinde yapılan uçuşlar için gecikme sürelerine en fazla etki eden değişken uzaklık iken diğer uçuş zamanlarına gün değişkeninin daha fazla etki ettiği görülmektedir. 0-3 saat arası gecikmelerin oranının en az hafta içi 2000 milden az gerçekleşen akşam uçuşlarında meydana gelmesi dikkat çekicidir. 2000 milden uzun yolculuklar için gecikme süresine yolculuk gününün etkisi olmadığı görülmektedir. Benzer şekilde hafta sonu gerçekleşen diğer uçuş zamanlarındaki uçuşlarda gecikme süresine uzaklık değişkenin etkisi olmadığı görülmektedir.

SONUÇ

İki bölümden oluşan çalışmada birinci bölümde büyük veri araçları ile R entegrasyonu ikinci bölümde havacılık sektörüne yönelik olarak yapılmış bir uygulama yer almaktadır. Büyük veri teknolojileri incelenmiş ve bunların sık kullanılan veri analizi aracı R programı ile arasındaki entegrasyon ifade edilmiştir. Denetimli ve denetimsiz öğrenme yaklaşımları kabaca incelendikten sonra ABD Ulaştırma Bakanlığı Ulaşım İstatistikleri bürosunca yayınlanan verilerle uygulama yapılmıştır. İlk uygulamada havayolu şirketlerinin uçuş mesafeleri ortalamaları ile rötar süreleri arasındaki ilişkiye dayalı olarak kümeleme analizi yapılmıştır. Bu analiz sonucunda hava yolu şirketleri üç kümeye ayrılmıştır. Bu sonuçlardan hareketle 1. kümeyi oluşturan havayollarının ABD içinde ana akım havayolları firmalarından olduğu 2. ve 3. Kümeyi oluşturan havayollarının ise genellikle ulusal veya kargo firmalarından oluştuğu gözlenmiştir. Birinci kümeyi oluşturan firmalar daha uzun mesafeli uçuşlar gerçekleştiren ve rötar süreleri en fazla olan şirketlerdir. Bu kümede şirketler için rötar süreleri önemli bir seviyede sorun olarak görünmektedir. İkinci kümede mesafe ölçütünün değişimi kümeyi oluşturan bütün birimler için aynı seviyededir. Bu kümede önem arz eden değişkenlik ise rötar sürelerindedir. Birinci küme ile benzerliği de bu değişimlerdir. Üçüncü küme analiz içinde sorunları en az olan havayolu firmalarını temsil etmektedir. Bu şirketler analiz sonuçlarından da görüldüğü gibi merkeze en yakın kümedir. Diğer kümelerle karşılaştırıldığında bu firmaların genellikle kısa mesafe yolcu ve kargo uçuşları yaptığı, bu nedenlerle rötar sorunu en az olan kümedir.

Birinci küme içerisinde bulunan ve en fazla sorun yaşayan şirketlerden olan American Airlines firması ikinci uygulama kapsamında odak noktasına alınmıştır. AA firmasında gerçekleşen gecikmeler koşullu çıkarım ağacı kullanılarak değerlendirilmiştir. Öncelikle değişkenlerde birtakım dönüşümler

uygulanmış ve gecikme süresini etkileyen anlamlı değişkenler ile çalışma gerçekleştirilmiştir. Dönüşüm sonrasında bağımlı değişken hem sürekli hem de kategorik olarak incelendiğinde yorumlanması daha kolay olduğu için kategorik formattaki durumuna odaklanılmıştır. Karar ağacı yapısı en önemli değişken olarak uçuş zamanlarını belirlemiş ve uçuş zamanlarını da akşam uçuşları ve diğer uçuşlar olarak dallandırmıştır. Akşam uçuşları içerisinde önce uzaklığın sonra uçuş zamanının (hafta içi, hafta sonu) etkili olduğu görülmüştür. Gecikme süreleri dağılımı akşam uçuşları bazında incelendiğinde 2000 milden kısa olan yolculuklarda 2000 milden uzun olan yolculuklara göre 24 saatten fazla gecikme vakası olması oranının daha fazla olduğu gözlenmiştir. 0-3 saat arası gecikmelerin oranının en az hafta içi 2000 milden az gerçekleşen akşam uçuşlarında meydana gelmesi dikkat çekicidir. 2000 milden uzun yolculuklar için gecikme süresine yolculuk gününün etkisi olmadığı görülmektedir. Benzer şekilde hafta sonu gerçekleşen diğer uçuş zamanlarındaki uçuşlarda gecikme süresine uzaklık değişkenin etkisi olmadığı gözlenmiştir. Bu bölümde AA için gecikme süresi değişkenin diğer değişkenlerce ne şekilde oluştuğunun yapısı gözlenmiş ve diğer havayolu şirketleri için genellenebilir bir çalışma yapılabileceği ortaya konmuştur. Her bir havayolu şirketi için herhangi bir gün herhangi bir zaman aralığındaki olası bir uçuş için olası gecikme süresinin ne kadar olacağına yönelik modelleme çalışması öncesinde koşullu çıkarım ağacı yaklaşımı ile veri seti içerisindeki yapının gözlenmesi modelleme öncesinde fikir verici olacaktır.

KAYNAKÇA

De Mauro, A., Greco, M., & Grimaldi, M. (2015, February). What is big data? A consensual definition and a review of key research topics. In AIP conference proceedings (Vol. 1644, No. 1, pp. 97-104). AIP.

Yengi, Y ., (2016). *Büyük Veride Duygu Analizine Dayalı Öneri Sistemleri*, Yüksek Lisans Tezi, Kocaeli Üniversitesi Fen Bilimleri Enstitüsü, Kocaeli.

Salur, M.U., (2016). *Büyük Veri Araçlarından Hadoop Kullanarak Veri Madenciliği*, Yüksek Lisans Tezi, Pamukkale Üniversitesi Fen Bilimleri Enstitüsü, Denizli.

Alpaydin, E. (2004). Introduction to Machine Learning, Massachusetts.

Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge university press.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). Foundations of machine learning. MIT press.

Zaharia, M. (2016). An architecture for fast and general data processing on large clusters. Morgan & Claypool.

Özdeş, M., (2017). *Büyük Veri Araçları Kullanarak Duygu Analizi Gerçekleştirimi*, Yüksek Lisans Tezi, Pamukkale Üniversitesi Fen Bilimleri Enstitüsü, Denizli.

Gonzalez, J. (2012). Parallel and Distributed Systems for Probabilistic Reasoning (No. CMU-ML-12-111). CARNEGIE-MELLON UNIV PITTSBURGH PA MACHINE LEARNING DEPT.

Akgün, B., (2016). *Apache Spark Tabanlı Destek Vektör Makineleri ile Akan Büyük Veri Sınıflandırma,* Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

Çetinkaya, S., (2016). *Hadoop/MapReduce Teknolojisi Kullanılarak Hızlı Tüketim Sektöründe Büyük Veri Analizi*, Yüksek Lisans Tezi, İstanbul Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

Hallaç, İ.R., (2014). *Büyük Veri Analizinde Dağıtık Makine Öğrenmesi Algoritmalarının Kullanılması*, Yüksek Lisans Tezi, Fırat Üniversitesi Fen Bilimleri Enstitüsü, Elazığ.

Moctezuma, L. E. G., Lobov, A., & Lastra, J. L. M. (2012, November). Decision making by using tree-like structures on industrial controllers. In ICT and Knowledge Engineering (ICT & Knowledge Engineering), 2012 10th International Conference on (pp. 77-83). IEEE.

Anderberg, M. R. (1973). Cluster analysis for applications (No. OAS-TR-73-9). Office of the Assistant for Study Support Kirtland AFB N MEX.

Bryan F.J., (1994). Multivariate Statistical Methods, Second Edition, Chapman – Hall, Londra.

Alpar, Reha., (2013). Uygulamalı Çok Değişkenli İstatistiksel Yöntemler, Syf:318-319.

Alpar, Reha., (2013). Uygulamalı Çok Değişkenli İstatistiksel Yöntemler.

Çokluk, Ö., Şekercioğlu G., Büyüköztürk, Ş., (2010). Sosyal Bilimler için Çok Değişkenli İstatistik Spss ve Lisrel Uygulamaları, Pegem Akademi, Ankara.

Data, G. O., Han, I., & Kamber, M. (2010). Data Mining: Concepts and Techniques. Morgan Kaufinann (2006).

MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. IEEE Transactions on neural networks, 16(3), 645-678.

Tan, P. N. (2006). Introduction to data mining. Pearson Education India.

Elektronik Kaynaklar

What is big data (SAS). (20 Kasım 2015). Retrieved from http://www.sas.com/en_us/insights/big-data/what-is- big-data.html.

Büyük Veri Analitiği Türleri(IBM), (03 Şubat 2018). Retrieved from https://www.ibm.com/;www-01.ibm.com/com/on/ssi/cgi%20bin/ssialias?infotype=SA%26subtype=WH%26htmlfid=TIW14162USEN.

Hadoop Offical Web Site, Apache Hadoop. (08 Şubat 2018). Retrieved from http://hadoop.apache.org.

Apache Spark Offical Web Site, Apache Spark. (08 Şubat 2018). Retrieved from http://spark.apache.org.

Apache Spark Official Page, SparkR. (08 Şubat 2018). Retrieved from https://spark.apache.org/docs/latest/sparkr.html.

The Comprehensive R Archive Network. (2018). Retrieved from https://cran.r-project.org/

Stackoverflow.com, MapReduce Word Count Proccess Figure. (08 Şubat 2018). Retrieved from

https://stackoverflow.com/questions/20317152/how-shuffling-is-done-in- mapreduce.

rstudio.com, sparklyr. (08 Şubat 2018). Retrieved from http://spark.rstudio.com.

Decision tree. (2014). Retrieved from

http://www.webcitation.org/query?url=http%3A%2F%2Fen.wikipedia.org%2Fwiki%2FDecision_tree&date=2014-06-02.

Decision tree learning.(2014). Retrieved from

http://www.webcitation.org/query?url=http%3A%2F%2Fen.wikipedia.org%2Fwiki%2FDecision_tree_learning&date=2014-05-28.

Karar Teorisi, Karar Ağacı ve Tıpta Uygulamaları.(2014). Retrieved from

http://www.webcitation.org/query?url=http%3A%2F%2Fwww.saglikekonomisi.com%2Fsed%2Findex.php%2Fdergiarsivi%2Fsay-2%2F62-karar-teorisi-karar-agaci-vetipta uygulamalar+&date=2014-05-28.