All Sciences Proceedings
http://as-proceeding.com/

2nd International Conference on Engineering, Natural and Social Sciences

April 4-6, 2023 : Konya, Turkey

https://www.icensos.com/

MAKİNE ÖĞRENMESİ YÖNTEMLERİ KULLANARAK HAVAYOLLARI BİRİM GELİR TAHMİNLEME

Beyza Nur Atalay[1*], Cemil Zalluhoğlu [2]

[1]*Veri ve Bilgi Mühendisliği / Bilişim Enstitüsü, Hacettepe Üniversitesi, Türkiye*
[2]*Bilgisayar Mühendisliği Bölümü / Hacettepe Üniversitesi, Türkiye*

[*]*(beyza.atalay@hacettepe.edu.tr)*

Özet **–** Bu proje havayollarının her bir yolcu başına elde ettiği geliri tahmin etmek amacıyla makine öğrenmesi uygulamalarından faydalanan çalışmayı içermektedir. Analiz için farklı git-gel uçuş rotaları seçilerek elde edilen sonuçların birbirleri ile olan ilişkisi gözlemlenmiştir. Her bir rotanın belirli süre zarfında yolcu sayısı ve birim gelir değişimi incelenmiş olup, seçilen birden fazla makine öğrenmesi modellemelerinden en başarılı sonuç Rassal Orman yaklaşımından elde edilmiştir.

Makine Öğrenmesi, Birim Gelir, Yolcu, Hat, Meydan

I. GİRİŞ

Bu projede yapılan çalışmanın amacı havayolu firmalarının hat bazında bir yolcu başına elde ettiği geliri (birim gelir) öngörebilmesi için makine öğrenmesi algoritmalarından faydalanılmasıdır.
Havayolu firmaları için üretimleri sonucunda elde edeceği toplam geliri öngörebilmesi önemli bir amaçtır. Bu amaç doğrultusunda her bir üretim için birim gelir tahmini yapabilen bir algoritmaya sahip olmak, her bir hat için birim gelir analizini daha doğru yapmasına sebep olacaktır ve bu minvalde üretimleri sonucunda daha karlı sonuç elde etmelerine olanak tanınacaktır. Birim gelirlerini önceden tahmin etmek, havayollarının üretkenliğini arttırmalarına, operasyonlarını buna göre planlamalarına, uçaklarının kapasitesini ve fiyatlarını belirlemelerine, pazarlama stratejilerini uygulamalarına ve nakit akışlarını yönetmelerine de yardımcı olur. Temelde proje kapsamında geliştirilen algoritma her bir hattın bir yıllık süre içerisinde birim gelir gelişim grafiğini yolcu sayısı ile orantılı bir şekilde takip ederek, gelir analizi yapmaya çalışmaktadır.

II. VERİ ANALİZ

A. *Analiz İçin Kullanılan Veri Seti*

Geliştirilecek olan modelleme için öncelikle belirli rotalar seçilmiştir. Bu rotalar sırasıyla FRA-SAW, SAW-FRA, SAW-CDG, CDG-SAW, ESB-ECN, ECN-ESB olarak çalışmada yer almaktadır (Rotalar havalimanı 3'lü kodları ile belirtilmektedir). Algoritmanın çalıştığı veri setini hazırlamak için öncelikle farklı metrikleri içeren ham bir veri setine ihtiyaç duyulmuştur. Bu veri seti rotalara ait bir senelik (2022 yılı) uçuş bilgisini 25 farklı sütunda gruplandırmış olup, bu sütunlar özetle şu bilgileri içermektedir;

- Kalkış meydanı, varış meydanı
- Yarı ve tam parkur bilgisi
- Uçuş tarihi, veri akış tarihi
- Uçak kapasitesi, biletli ve rezervasyonlu yolcu sayısı bilgisi
- Bilet sınıfları, kabin sınıfları
- Net gelir
- Ülke, şehir bilgileri

Niteliklerin birbiri ile ilişkisini gözlemlemek için korelasyon matrisi oluşturulmuş olup, elde edilen bağımlılık durumlarına göre nitelik seçimi yapılmıştır.

B. *Veri Ön İşleme Çalışmaları*

İlk olarak incelenen Sabiha Gökçen Havalimanı – Frankfurt Havalimanı (SAW-FRA / FRA-SAW) arasındaki uçuşlara ait bilgileri içeren veri setinde, ön işleme çalışmasından önce 237.589 satır ve 11 sütun bulunmaktadır.

Veri setinde boş olan satırlar elenerek Tablo 1' de gösterildiği üzere 11 sütunluk 77.065 satırdan oluşan veri seti elde edilmiştir.

Tablo 1- Kullanılan Veri Seti Bilgisi

Nitelikler	Data Tipi
TKTPSGR_L	float64
UNIT_L	float64
RES_L	float64
TKTPSGR_G	int64
RES_G	int64
UNIT_G	float64
MONTH	int64
WEEKNO	int64
DOW	int64
GDR_UNIT	float64
SEMIROUTE	int64

Aynı ön işleme çalışması Sabiha Gökçen Havalimanı- Paris Charles de Gaulle Havalimanı (SAW-CDG-SAW) ve Esenboğa Havalimanı – Ercan Havalimanı (ESB-ECN-ESB) parkurları içinde yapılmış olup veri seti boş değerlerden temizlendikten 81.381, 44.014 satırdan oluşan veri setleri elde edilmiştir. Projenin ilerleyen bölümlerinde aynı dağıtım noktasına (HUB) sahip iki rota birleştirilerek oluşturulan yeni veri seti üzerinden de yapılmış olan bir çalışma yer almaktadır. Amaç modelin elde ettiği çıktıların anlamlı olması durumunda her hat bazında tek tek değerlendirme yapılmadan, modelin tüm veri seti içerisinden her hatta özel, doğru tahminlemeyi yapmasıdır. Bu minvalde CDG ve FRA uçuşları için toplanan veriler birleştirilmiştir. Ardından oluşan yeni veri setinden boş değerler çıkarılarak veri seti modellemeye hazır hale getirilmiştir.

III. MODELLEME

A. *Makine Öğrenmesi Algoritmaları*

Birim gelir tahminlenmesi için tercih edilen algoritmalar sınıflandırıcı makine öğrenmesi modellemelerinden ziyade regresyon temelli makine öğrenmesi modelleri olmuştur [1].

Değişkenler arasındaki ilişkiyi belirlemek için regresyondan faydalanılmaktadır. Bu ilişki içerisindeki değişkenler bağımlı ve bağımsız değişken olarak adlandırılır ve bağımsız değişkenler kullanılarak bağımlı değişkenlerin sonucunun tahminlenmesi regresyon modelleri ile gerçekleştirilir. Çalışmada kullanılan modellerin değerlendirilmesi aşağıdaki kriterlere göre yapılmıştır.

- Ortalama Mutlak Hata (Mean Absolute Error – MAE)
- Ortalama Hata Karesi (Mean Squared Error – MSE)
- Ortalama Karekök Sapması (Root Mean Squared Error – RMSE)
- Accuracy

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

$n = Gözlem\ sayısı$
$x_i = Gerçek\ değer$
$y_i = Tahmin\ edilen\ değer$

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$n = Gözlem\ sayısı$
$Y_i = Gözlemlenen\ değer$
$\hat{Y}_i = Tahmin\ edilen\ değer$

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(Y_i - \hat{Y}_i)^2}{n}}$$

$n = Gözlem\ sayısı$
$Y_i = Gözlemlenen\ değer$
$\hat{Y}_i = Tahmin\ edilen\ değer$

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

$TP = Doğru\ Pozitif$
$TN = Doğru\ Negatif$
$FP = Yanlış\ Pozitif$
$FN = Yanlış\ Negatif$

A.1. Support Vector Regression(SVR) Modeli

Support Vector Machine(Destek Vektör Makinesi - SVM) modeli bir hiperdüzlemi kullanarak mümkün olan en yüksek sayıda eğitim verisini doğru bir şekilde sınıflandırmak ve hiperdüzlemi ayıran en büyük marjı bulmak için destek vektörlerini kullanır ve optimal hiperdüzlemi temsil eder [2]. SVM, doğrusal ve doğrusal olmayan olmak üzere iki ana kategoriye ayrılır. Doğrusal SVM, doğrusal olarak ayrılabilen veri kümeleri için kullanılır. Bu tür veri kümeleri, iki sınıfı tek bir düz çizgi ile ayırmak mümkün olduğunda doğrusal SVM sınıflandırıcısı olarak adlandırılır. Ancak, veri kümeleri tek bir düz çizgi ile sınıflandırılamazsa, bu veriler doğrusal olmayan veriler olarak adlandırılır. SVM algoritması, sınıfları n boyutlu uzayda ayırmak için birden fazla karar çizgisi kullanabilir. Ancak, veri noktalarını sınıflandırmaya en iyi yardımcı olan karar sınırının bulunması gereklidir. Bu en iyi sınır, SVM'nin hiperdüzlemi olarak bilinir. SVM algoritması, verileri sınıflandırmak için en uygun hiperdüzlemi bulmak için optimize edilir. SVM'nin genelleştirilmiş hali olan SVR, fonksiyonun etrafına epsilon duyarsız bir bölge ekleyerek gerçekleştirilir [3]. Bu epsilon tüpü, model karmaşıklığını ve tahmin hatasını dengeleyerek, sürekli değerli fonksiyona en iyi yaklaşan epsilon tüpünü bulmak için optimizasyon problemini yeniden formüle eder. SVR'nin önemli bir özelliği, hesaplama karmaşıklığının girdi verilerinin boyutlarına bağımlı olmamasıdır. Bu sayede, SVR veri seti büyüdükçe performansını kaybetmez.

Modelin veri setine uygulanması sürecinde 3 farklı çekirdek (kernel) kullanımı tercih edilmiştir. Bu çekirdekler sırasıyla "RBF", "Polynomial" ve "Linear" olmuştur. Bu değerlerin yanı sıra SVR skoru 'da test edilmiştir. FRA-SAW , SAW-FRA uçuşlarına ait veri setinden elde edilen model sonuçları Tablo 2'de sırasıyla gösterilmektedir.

Tablo 2- SAW-FRA-SAW Hattı Model Sonuçları

	RBF Kernel	Polynomial Kernel	Linear Kernel
MAE	25.392818	27.193136	25.385616
MSE	1427.195654	1599.375929	1439.118850
RMSE	37.778243	39.992198	37.935719

Elde edilen SVR çıktılarına göre model birim gelir tahmininde çok yeterli olmamıştır. Mutlak hata değerleri modelin tahmini değerlerinin gerçek değerlerden ne kadar farklı olduğunu gösterir ve bu minvalde değerlendirildiğinde hata payının veri setine göre yüksek olduğu görülmektedir. Aynı model CDG hattı içinde uygulanmış olup, elde edilen model çıktıları Tablo 3' de gösterilmektedir.

Tablo 3- SAW-CDG-SAW Hattı Model Sonuçları

	RBF Kernel	Polynomial Kernel	Linear Kernel
MAE	26.666909	32.660061	28.703805
MSE	1405.876783	1947.271483	1629.534549
RMSE	37.495023	44.127899	40.367493

En düşük SVR sonucu ise ESB-ECN-ESB uçuş hattında ortaya çıkmıştır. Bu noktadaki farklılığın temel nedeni Esenboğa Havalimanı'ndan gerçekleşen uçuşların Sabiha Gökçen Havalimanı'ndan gerçekleşen uçuşlara kıyasla daha farklı uçuş karakterine sahip olmasından dolayı farklılaşan veri yapısıdır. ESB-ECN-ESB hattının SVR çıktıları Tablo 4' de gösterilmektedir. Bütün hatlar için elde edilen SVR skorları ise Tablo 5'de verilmiştir. Modelin SVR skor değerleri de arzu edilenden düşük kalmıştır ve elde edilen bu sonuçlara göre SVR modelinin veri kümesinde anlamlı bir varyansı açıklayamadığı anlaşılmaktadır.

Tablo 4- ESB-ECN-ESB Hattı Model Sonuçları

	RBF Kernel	Polynomial Kernel	Linear Kernel
MAE	15.641678	16.272760	17.040647
MSE	458.268188	496.843355	527.468584
RMSE	21.407199	22.289983	22.966684

Tablo 5- SVR Skor

	FRA	CDG	ECN
SVR SKOR	0.668703	0.820335	0.266445

A.2. Lineer Regresyon(LR) Modeli

Lineer regresyon, istatistiksel veri analizinde sıkça kullanılan bir yöntemdir [4]. Bu yöntem doğrusal ve sürekli değişkenler için kullanılır. Olasılık dağılımı ile analiz yapılır. Değişkenler arasında doğrusal bir ilişki gözlenirse, LR modeli gelecekteki tahminleri yapmak, değişkenler arasındaki etkileşimleri incelemek ve çıkarımlar yapmak için kullanılır. Modelin FRA veri setine uygulanan LR sonucunda Tablo 6' da verilen çıktılar elde edilmiştir. Elde edilen sonuçlar değerlendirildiğinde SVR

modelinde de olduğu gibi arzu edilen düzeyde başarılı bir sonuç elde edilememiştir. Modelin tahmin değerleri ve gerçek değerleri arasındaki fark azımsanacak düzeyde olmadığı için LR yaklaşımı yapılan çalışmada yetersiz kalmıştır. CDG veri seti için de farklı bir sonuç gözlemlenmemiştir. ECN-ESB-ECN uçuş hattı için yapılan çalışmada ise en düşük doğruluk payı ve en başarısız LR sonucu elde edilmiştir.

Tablo 6- LR Sonuçları

	FRA	CDG	ECN
Accuracy	66.438392	79.323814	16.959445
MAE	25.564668	29.121440	17.153112
MSE	1429.017328	1570.237334	524.748323
RMSE	37.802345	39.626220	22.907385

Bu doğrultuda iki farklı makine öğrenmesi algoritması daha model üzerinde denenmiştir. Bunlar sırasıyla Random Forest ve XG Boost modelleridir.

A.3. Random Forest(Rassal Orman-RF) Modeli

Random Forest, denetimli öğrenme tekniğine ait popüler bir makine öğrenme algoritmasıdır [5]. Sınıflandırma ve regresyon problemleri için kullanılabilmektedir. Algoritma, birden fazla sınıflandırıcıyı birleştirme sürecine dayanır ve verilen veri setinin çeşitli alt kümelerinde bir dizi karar ağacını içerir. Her ağaçtan yapılan tahminlere dayalı olarak, tahminleri alır ve nihai çıktıyı tahmin eder. Rastgele ormanı oluşturmak için N karar ağacını birleştirilir ve sonrasında her ağaç için tahmin yapılarak sınıflandırma işlemi tamamlanır. Ormandaki ağaç sayısı arttıkça doğruluk artar ve fazla uyum sorununu önlenir. RF modeli geleneksel tahminlerden daha yüksek doğruluğa sahip olmasının yanı sıra mikro bilgi içeren veri setleriyle başa çıkmak için kullanılabilmektedir. Son yıllarda RF gelişme kaydederek birçok alanda yaygın olarak kullanılmaya başlanmıştır. RF uygulaması modelde denendikten sonra çıkan sonuçlara göre seçili veri kümesinde iyi performans gösterdiği anlaşılmıştır. Tablo 7' de sırasıyla FRA , CDG ve ECN için belirtilmekte olduğu gibi modelin yüksek doğruluk değerine sahip olmasının yanı sıra ortalama mutlak hata değerinin düşüklüğü modelin arzu edilen sonuçları verdiğini göstermektedir.

Tablo 7- RF Sonuçları

	FRA	CDG	ECN
Accuracy	99.591750	99.581258	88.221672
MAE	0.876579	1.530982	4.287663
MSE	17.276763	32.915985	74.275207
RMSE	4.156532	5.737245	8.618306

A.4. XG Boost Modeli

XGBoost, eXtreme Gradient Boosting olarak önerilen algoritma Chen ve Guestrin tarafından geliştirilmiştir [6]. Bu algoritma, öğrenmeyi iyileştirmek ve aşırı uydurmayı kontrol etmek için algoritmik yenilikler ve hiper parametreler içeren optimize edilmiş bir Gradient Boosting sistemi sağlamaktadır. XGBoost modeli de bir karar ağacı modellemesine dayandığından dolayı RF Modeli'nden daha iyi bir sonuç elde edip etmeyeceğini test etmek için veri setleri üzerinde denenmiştir. Elde edilen çıktılara göre XGBoost bu amaç doğrultusunda yetersiz kalmıştır. Tablo 8' de verilmekte olan çıktılar incelendiğinde modelin doğruluk değeri her ne kadar yüksek çıksa dahi tahmin değerleri gerçek değerlerden yüksek oranda farklı seyretmektedir. Bu minvalde birim gelir tahminlemesinde XGBoost yaklaşımı başarılı olmamaktadır.

Tablo 8- XG Boost Sonuçları

	FRA	CDG	ECN
Accuracy	99.591750	99.581258	88.221672
MAE	55.616198	66.753124	25.274790
MSE	4776.240253	7146.666367	1060.955686
RMSE	69.1110348	84.537958	32.572314

A.5. Birleştirilmiş Veri Set Çalışması

FRA ve CDG uçuşlarının veri setlerinin birleştirilmesiyle oluşturulan yeni veri setinde meydanlar ayrı ayrı değerlendirilirken, başarılı sonuç elde eden RF Modeli denenmiştir. Çıkan sonuçlar gözlemlendiğinde birbirinden farklı bu iki meydana gerçekleştirilen git-gel uçuşlarının birim gelirinin tahmin verilerinin, gerçekleşen verilere yakın olduğu gözlemlenmiştir. Tablo 8'de verilmiş olan çıktılar incelendiğinde RF yaklaşımının başarılı olduğu görülmektedir.

Tablo 8-FRA+CDG RF Sonuçları

	FRA + CDG
Accuracy	99.354627
MAE	1.839382
MSE	41.802441
RMSE	6.465480

IV. TARTIŞMA

Havayolları için birim gelir öngörme çalışması sonucunda elde edilen çıktılara göre ağaç algoritması yaklaşımıyla modelde daha etkili sonuçlar elde edilmesi beklenmektedir. SVR, LR ve XGBoost yaklaşımlarında elde edilen hata payının arzu edilen çok daha fazla oluşu bu modelleri yapılan çalışma için yetersiz kılmaktadır. RF modeli bu hata payını azımsanmayacak ölçüde düşürerek kabul edilebilir seviyeye indirmektedir. Modelin öğrenebilmesi ve anlamlı sonuçlar elde edebilmesi için birim gelirin uçuş gerçekleşene kadar geçen zaman içerisindeki değişiminin göz önünde bulundurulması önemlidir. Yapılan çalışma sonrasında modelin daha da geliştirilmesiyle proje hayata geçirilerek, yanlış öngörüden kaynaklanan gelir kayıplarına engel olabilir. Havayollarının doğru zamanda doğru ücrette biletlerini satmayı başarması önemli bir hedeftir. Yapılan analizler sonucunda kullanılacak olan makine öğrenmesi modellemeleri bu hedefe ulaşmayı daha mümkün kılacaktır.

V. SONUÇLAR

Proje kapsamında üç farklı ülkeye Sabiha Gökçen Havalimanı'nda gerçekleşen seferlerin birim gelirleri analiz edilerek, gelecek süreç için birim gelir öngörme çalışması yapılmıştır. RF yaklaşımı ile elde edilen sonuçlar başarılı olup, gelecek çalışmalarda modelin geliştirilip, hata payının daha fazla düşürülmesi doğrultusunda çalışma canlı ortama aktarılabilecek konuma gelebilecektir.İlerleyen çalışmalarda projede incelenen ülkelerin dışarısında kalan diğer pazar verileri de incelenerek modelin geliştirilmesi beklenmektedir. Farklı pazarların kendine özgü dinamikleri göz önünde bulundurulduğunda modelin tüm hatlar için tek bir seferde öngörü yapması yerine pazara göre gruplandırılarak denenmesi hedeflenmektedir.

KAYNAKLAR

[1] Kavitha S, Varuna S, & Ramya R. (2016). A comparative analysis on linear regression and support vector regression. *2016 Online International Conference on Green Engineering and Technologies (IC-GET).* https://doi.org/10.1109/get.2016.7916627

[2] Suthaharan, S. (2016). Support Vector Machine. *In Machine learning models and algorithms for Big Data Classification Thinking with examples for effective learning* (pp. 207–235). essay, Springer US.

[3] Özkaya, U., & Öztürk, Ş. (2022). Gaussian Regression Models for Day-Level Forecasting of COVID-19 in European Countries. Understanding COVID-19: The Role of Computational Intelligence, 339-356.

[4] Dar, U. (n.d.). *Doğrusal (lineer) regresyon.* Retrieved January 27, 2023, from https://bookdown.org/ugurdar/dogrusalregresyon/ (Erişim tarihi: 22 Ocak 2023)

[5] Liu, Y., Wang, Y., & Zhang, J. (2012). New Machine Learning Algorithm: Random forest. *Information Computing and Applications,* 246–252. https://doi.org/10.1007/978-3-642-34062-8_32

[6] Shi, X., Wong, Y. D., Li, M. Z.-F., Palanisamy, C., & Chai, C. (2019). A feature learning approach based on XGBoost for driving assessment and risk prediction. *Accident Analysis & Prevention,* 129, 170–179. https://doi.org/10.1016/j.aap.2019.05.005