

Toplu taşımada varış süresi tahmini için karar ağaçları tabanlı makine öğrenmesi modellerinin karşılaştırması

Turgay SARIOĞLU¹
Hakan Burak EMEKLİ^{*2}

Geliş tarihi / Received: 12.12.2025

Düzeltilerek geliş tarihi / Received in revised form: 13.12.2025

Kabul tarihi / Accepted: 23.02.2025

DOI: 10.17932/IAU.ABMYOD.2006.005/abmyod_v21i73005

Öz

Bu çalışma, kentsel toplu taşıma sistemlerinde otobüslerin duraklar arası tahmini varış sürelerini (Estimated Time of Arrival, ETA) daha yüksek doğrulukla öngörebilmek amacıyla karar ağaçları tabanlı makine öğrenmesi algoritmalarının kapsamlı bir karşılaştırmasını sunmaktadır. MBTA tarafından sağlanan “Bus Arrival/Departure Times (2024)” veri seti kullanılarak ham zaman damgaları, sefer kimlikleri ve durak sıraları çok aşamalı bir ön işleme sürecinden geçirilmiş; her seferdeki tüm $i < j$ durak kombinasyonlarını kapsayan durak çifti oluşturulmuştur. Bu yaklaşım, hem kısa mesafeli hem de uzun mesafeli segmentlerin aynı çerçevede modellenmesine olanak tanıyarak seyahat süresi dinamiklerinin daha ayrıntılı temsil edilmesini sağlamıştır.

Çalışmada Karar Ağacı, Rastgele Orman, XGBoost, LightGBM ve CatBoost algoritmaları aynı eğitim ve test ayrımı altında değerlendirilmiş; performansları MAE, RMSE ve R^2 metrikleri ile ölçülmüştür. Her hat için zaman serisi yapısına uygun 5 katlı çapraz doğrulama uygulanmış ve raporlanan tüm metrikler katmanlar üzerinden elde edilen ortalama değerler olarak sunulmuştur. Üç farklı hat üzerinde yapılan deneylerde CatBoost en yüksek doğruluğu sağlamıştır. Ortalama sonuçlara göre Hat 01’de $RMSE = 237.46$, $MAE = 157.08$ ve $R^2 = 0.86$; Hat 111’de $RMSE = 161.31$, $MAE = 98.31$ ve $R^2 = 0.85$; Hat CT2’de ise $RMSE = 260.98$,

¹ İstanbul Aydın Üniversitesi, Lisansüstü Eğitim Enstitüsü, turgaysarioğlu@stu.aydin.edu.tr, ORCID:0009-0007-3316-2371

^{2*} Sorumlu Yazar/ Corresponding Author: İstanbul Aydın Üniversitesi, hakanemekli@aydin.edu.tr, ORCID: 0000-0002-6503-1284

MAE = 156.52 ve $R^2 = 0.89$ elde edilmiştir. Eğitim süresi görece uzun olsa da CatBoost'un doğruluk ve istikrar açısından ETA tahmini için üstün bir performans sunduğu görülmüştür.

Anahtar Kelimeler: *Varış Süresi Tahmini, Karar Ağaçları, Rastgele Orman, LightGBM, CatBoost, XGBoost*

Comparison of decision tree-based machine learning models for arrival time prediction in public transportation

Abstract

This study presents a comprehensive comparison of decision tree-based machine learning algorithms to predict bus arrival times (ETAs) between stops with higher accuracy in urban public transportation systems. Using the “Bus Arrival/Departure Times (2024)” dataset provided by the MBTA, raw timestamps, trip IDs, and stop sequences were subjected to a multi-stage preprocessing process. A stop-pair structure encompassing all stop combinations $i < j$ on each trip was created. This approach allows for the modelling of both short-distance and long-distance segments within the same framework, providing a more detailed representation of travel time dynamics.

In this study, Decision Tree, Random Forest, XGBoost, LightGBM, and CatBoost algorithms were evaluated under the same training and testing split using MAE, RMSE, and R^2 metrics. A 5-fold cross-validation strategy consistent with the time-series structure was applied for each route, and all reported metrics represent average values across folds. Across three different routes, CatBoost achieved the highest accuracy. Based on average results, the model obtained $RMSE = 237.46$, $MAE = 157.08$, and $R^2 = 0.86$ for Route 01; $RMSE = 161.31$, $MAE = 98.31$, and $R^2 = 0.85$ for Route 111; and $RMSE = 260.98$, $MAE = 156.52$, and $R^2 = 0.89$ for Route CT2. Although its training time was longer, CatBoost demonstrated superior performance and stability for ETA prediction.

Keywords: *Time to Arrival Estimation, Decision Trees, Random Forest, LightGBM, CatBoost, XGBoost*

Giriş

Kentsel toplu taşıma sistemlerinde varış süresi tahminlerinin (ETA) doğruluğu, hem operatörler için operasyonel verimliliği artırmada hem de yolcular için bekleme süresi belirsizliğini azaltmada kritik bir rol oynamaktadır (Vlahogianni ve ark., 2014). GPS (Küresel Konumlama Sistemi) ve AVL (Otomatik Araç Konumlama) gibi otomatik konumlandırma teknolojilerinin yaygınlaşması, otobüs hareketlerine ilişkin yüksek hacimli verilerin düzenli olarak toplanmasını sağlayarak, istatistiksel ve makine öğrenimi yöntemleri kullanılarak seyahat süresinin daha ayrıntılı modellenmesini mümkün kılmıştır (Büchel ve Corman, 2020).

Son yıllarda yapılan çok sayıda çalışma, gerçek zamanlı AVL/GPS verilerinin toplu taşımada seyahat süresi ve varış süresi tahmininin doğruluğunu önemli ölçüde artırdığını göstermiştir. Çeşitli çalışmalar, GPS ile elde edilen araç konum verilerinin düzenli aralıklarla işlenmesiyle kısa vadeli seyahat süresi tahmininin iyileştirilebileceğini göstermiştir (Sun ve ark., 2007; Fan ve Gurmu, 2015). Ayrıca, düşük örnekleme frekanslı verilerin, uygun veri işleme teknikleri kullanılarak modellendiğinde kentsel ağlarda güvenilir seyahat süresi tahmini sağlayabileceği bildirilmiştir (Jenelius ve Koutsopoulos, 2013).

Akıllı şehirler bağlamında, IoT tabanlı sensörler, trafik kameraları ve GPS destekli araçlardan elde edilen verilerin uç bilişim altyapıları ile birlikte kullanılması, trafik ve ulaşım sistemlerinin daha gerçekçi ve anlık olarak modellenmesine olanak tanımaktadır. Bu tür sistemlerin, özellikle trafik yoğunluğu ve araç hareketlerinin tahmininde önemli avantajlar sunduğu vurgulanmaktadır (Okutucu ve Kaya, 2025). Benzer şekilde, IoT tabanlı sensörlerden elde edilen yüksek frekanslı operasyonel verilerin yapay zekâ algoritmaları ile entegre biçimde kullanılması, dinamik ve değişken sistemlerde tahmin doğruluğunu ve kaynak kullanım verimliliğini önemli ölçüde artırmaktadır. Jobani ve Kaya (2025), farklı iklim koşulları altında çalışan veri merkezleri için geliştirdikleri hibrit IoT ve yapay zekâ tabanlı modelde, gerçek zamanlı sensör verilerinin makine öğrenimi yöntemleriyle işlenmesinin sistem performansını anlamlı düzeyde iyileştirdiğini göstermiştir. Bu bulgular, kentsel toplu taşıma sistemlerinde AVL/GPS tabanlı gerçek zamanlı verilerin makine öğrenimi modelleri ile birlikte kullanılmasının, varış süresi (ETA) tahmin problemleri için de uygulanabilir ve genellenebilir bir yaklaşım sunduğunu desteklemektedir.

Makine öğrenimi alanında, karar ağacı tabanlı modeller, özellikle doğrusal olmayan ilişkileri yakalama ve açıklanabilirlik kabiliyetleri nedeniyle otobüs seyahat süresi tahmini literatüründe öne çıkmıştır. Çeşitli çalışmalar, Rastgele Orman ve modern gradyan artırma modelleri (XGBoost, LightGBM, CatBoost) gibi topluluk yöntemlerinin klasik yöntemlere göre daha yüksek doğruluk sağladığını göstermektedir (Moosavi ve ark., 2023). Ancak, bu algoritmaları aynı veri kümesi ve özellik mühendisliği süreci içinde sistematik olarak karşılaştıran çalışmalar nispeten azdır.

Bu çalışmada, Massachusetts Bay Transportation Authority (MBTA) tarafından sağlanan “Bus Arrival/Departure Times (2024)” veri seti kullanılarak beş karar ağacı tabanlı makine öğrenmesi algoritmasının (Karar Ağacı, Rastgele Orman, XGBoost, LightGBM ve CatBoost) kapsamlı bir karşılaştırmasını yürütmektedir (MBTA, 2024). Veri setindeki orijinal zaman damgaları ve durak dizileri, her rotadaki tüm ardışık durak çiftlerini içerecek şekilde durak çifti formatına dönüştürülmüş ve böylece mikro düzeyde seyahat süresi damgaları oluşturulmuştur. Modeller, aynı eğitim ve test bölümü altında MAE, RMSE ve R^2 performans ölçütleri kullanılarak değerlendirilmiştir.

Problemin tanımı

Şehir içi toplu taşıma ağlarındaki duraklara otobüslerin tahmini varış zamanını (ETA) yüksek doğrulukla tahmin etmek, hem yolcu bilgilendirme sistemlerinin etkinliği hem de filo yönetimi, hız optimizasyonu ve operasyonel planlama gibi idari süreçler için kritik bir gerekliliktir. Ancak, seyahat süreleri trafik yoğunluğu, sinyal döngüleri, istasyon giriş ve çıkış dinamikleri, sürücü davranışları ve hat özellikleri gibi çok boyutlu ve zamanla değişen faktörlerden etkilendiği için oldukça stokastik ve doğrusal değildir. Bu nedenle, ETA tahmin problemi klasik doğrusal yaklaşımlarla çözülemeyecek kadar karmaşıktır ve gelişmiş makine öğrenimi yöntemlerinin kullanılmasını gerektirmektedir.

Bu çalışmada ele alınan problem, yaygın olarak incelenen "ardışık duraklar arasındaki seyahat süresi tahmini" probleminin ötesine geçerek daha kapsamlı bir çerçeve sunmaktadır. Buna göre, her durak için yalnızca bir sonraki durak değil, aynı zamanda sonraki durakların tüm durak çifti kombinasyonları ($i < j$) da dikkate alınmakta ve bir başlangıç durağından herhangi bir varış durağına kadar olan seyahat süresi tahmin edilmektedir. Bu çerçeve, tüm güzergah boyunca hem kısa mesafeli hem de uzun mesafeli

segmentlerin aynı modelleme çerçevesine entegre edilmesini sağlayarak, durağa özgü, hat ölçeğinde ve zamana bağlı seyahat dinamiklerinin birlikte öğrenilmesini mümkün kılmaktadır.

MBTA'nın "Bus Arrival/Departure Times (2024)" veri seti kullanılarak her bir sefer için durak sırası, yön bilgisi ve gerçek zamanlı zaman damgaları işlenmiş; her durak-çifti için gerçek seyahat süresi şu şekilde hesaplanmıştır:

$$y_{ij} = \text{Timestamp}(j) - \text{Timestamp}(i), \quad 0 < i < j \leq N$$

Burada N , ilgili seferin toplam durak sayısını, y_{ij} ise başlangıç durağı i ile hedef durağı j arasındaki gerçek seyahat süresini (saniye cinsinden) ifade etmektedir.

Böylelikle problem, bağımsız değişkenler vektörü X_{ij} (durak sırası, yön, hat bilgisi, zaman bileşenleri ve sefer özellikleri) ile hedef değişken $\hat{y}_{ij}$ arasındaki fonksiyonel ilişkiyi modelleyen bir lineer olmayan regresyon problemi olarak tanımlanmaktadır:

$$\hat{y}_{ij} = f(X_{ij})$$

Amaç, tahmin edilen değer $\hat{y}_{ij}$ ile gerçek değer y_{ij} arasındaki hatayı minimize eden en uygun modeli $f(\cdot)$ elde etmektir. Bu doğrultuda çalışmanın temel araştırma sorusu; "Geniş kapsamlı durak-çifti seyahat süresi tahmin probleminde, karar ağaçları tabanlı makine öğrenmesi algoritmalarından hangisi en yüksek doğruluk ve genelleme performansını sunmaktadır?" şeklinde ifade edilebilir.

Seyahat sürelerindeki yüksek değişkenlik, duraklar arası mesafelerin heterojenliği ve veri setindeki kategorik değişkenlerin (ör. route_id, direction_id, stop_id) yoğunluğu göz önüne alındığında, bu problem oldukça karmaşık bir öğrenme yapısı sergilemektedir. Bu nedenle, karar ağaçlarına dayalı sağlam topluluk yöntemlerinin performansının sistematik olarak karşılaştırılması, hem literatürdeki boşlukları gidermek hem de pratik ETA tahmin uygulamalarına katkıda bulunmak için önemli bir araştırma ihtiyacını karşılamaktadır.

Materyal ve yöntem

Bu çalışmada, MBTA tarafından sağlanan "Bus Arrival/Departure Times

(2024)” veri seti kullanılarak otobüslerin duraklar arası seyahat sürelerini tahmin etmek amacıyla karar ağaçları tabanlı beş makine öğrenmesi modeli karşılaştırılmıştır. AVL/GPS zaman damgaları silinerek, seferler yön ve durak sırasına göre düzenlenmiş, ardından her seferdeki tüm ileri durak kombinasyonlarını içeren durak çifti tabanlı bir temsil oluşturulmuştur. Bu dönüşüm, aynı modelleme çerçevesi içerisinde hem kısa hem de uzun mesafeli durak segmentlerinin öğrenilmesine olanak sağlamıştır.

Model eğitim sürecinde zaman (ay, gün, saat), güzergâh ve konumsal (durak sırası, durak farkı) özellikler kullanılmıştır. Kategorik değişkenler uygun biçimde sayısallaştırılmış, tüm modeller aynı eğitim–test ayrımı altında uygulanmıştır. Karar Ağacı, Rastgele Orman, XGBoost, LightGBM ve CatBoost algoritmaları, MAE, RMSE ve R^2 performans ölçütleri ile değerlendirilmiştir. Böylece çeşitli ağaç tabanlı yöntemlerin ETA tahmini üzerindeki karşılaştırmalı performansı sistematik biçimde analiz edilmiştir.

Veri

Bu çalışmada kullanılan MBTA “Bus Arrival/Departure Times (2024)” veri seti, her bir otobüs seferine ait durak bazlı zaman damgalarını ve güzergâh bilgilerini içeren çok boyutlu bir yapı sunmaktadır. Veri setinde her satır, bir seferde belirli bir durağa gerçekleşen varış veya ayrılışı temsil etmektedir. Veri yapısı aşağıdaki alanlardan oluşmaktadır:

1. `service_date`: Seferin gerçekleştiği takvim günü. Bu alan, günlük trafik yoğunluğu, hafta içi/hafta sonu farklılıkları ve dönemsel etkilerin modellenmesi açısından önemlidir.
2. `route_id`: Otobüs hattının kimliği. Her hat farklı bir güzergâhı temsil ettiği için, rota bilgisi seyahat dinamiklerindeki yapısal değişkenliklerin yakalanmasında temel bir kategorik değişkendir.
3. `direction_id`: Sefer yönü (Inbound/Outbound). Aynı hat üzerinde karşıt yönlerde trafik koşulları ve durak dizilişleri farklılık gösterebildiği için model açısından kritik bir değişkendir.
4. `half_trip_id`: Bir seferi benzersiz olarak tanımlayan kimlik numarasıdır. Aynı hat ve aynı gün içinde gerçekleşen yüzlerce seferi ayırt etmek için kullanılır. Durak çifti dönüşümünde segmentlerin doğru sefer bağlamında gruplanması için zorunludur.
5. `stop_id`: Seferin uğradığı durağın benzersiz kodudur. Tahmin edilecek seyahat süresinin başlangıç ve hedef durağının tanımlanmasında temel rol oynar.

6. **time_point_id**: MBTA'nın dahili durak kodlama sistemine ait ek tanımlayıcıdır. Durak konumlarının dış kaynaklarla eşleştirilmesinde kullanılır.
7. **time_point_order**: Sefer içindeki durak sırasıdır. Bu alan sayesinde bir seferdeki duraklar arasında doğrusal bir ilerleyiş sağlanmakta ve durak çifti dönüşümünde ($i < j$) ilişkisi güvenilir biçimde oluşturulabilmektedir.
8. **point_type**: Duranın başlangıç (Startpoint), orta nokta (Midpoint) veya bitiş durağı (Endpoint) olup olmadığını belirtir.
9. **standard_type**: Seferin Schedule veya Headway standardına tabi olup olmadığını gösterir. Bu alan planlanan zamanların anlamlandırılması için gereklidir.
10. **scheduled**: Planlanan varış/ayrılış zaman damgasıdır. ETA tahmin modellerinde doğrudan kullanılsa da hat planlaması ve sefer gecikmelerinin analizinde bağlamsal bilgi sağlar.
11. **actual**: Gerçekleşen varış zamanı. Bu çalışma kapsamında seyahat süresi hesaplamalarının temel bileşeni olup tüm etiket (label) değerleri bu alandan türetilmiştir.
12. **scheduled_headway**: Bir önceki sefere göre planlanan zaman aralığını (headway) saniye cinsinden ifade eder.
13. **headway**: Sefer aralığı bilgisini temsil eder; veri setinde çoğu satırda boş olsa da bazı hatlarda başta kullanılan tamamlayıcı bir alandır.

Veri ön işleme

Veri seti doğrudan kullanılabilir nitelikte olmadığından, modelleme sürecine aktarılmadan önce kapsamlı bir yapılandırma aşamasından geçirilmiştir. İlk olarak, veri setinde yer alan gözlemler hat numarasına göre ayrıştırılmış ve her bir route_id değeri için bağımsız alt veri kümeleri oluşturulmuştur. Bu tercih durak dizilimleri, trafik yoğunluk profilleri ve operasyonel karakteristikleri nedeniyle heterojenliği azaltarak daha istikrarlı bir öğrenme ortamı sağlamaktadır. Benzer şekilde, gürültülü ve heterojen veri kaynaklarında modelleme öncesi gerçekleştirilen veri ön işleme ve ayrıştırma adımlarının, makine öğrenimi modellerinin tahmin performansını ve kararlılığını artırdığı literatürde vurgulanmaktadır (Kaya ve ark., 2021).

Büyük ölçekli ve zamana bağlı veri kümelerinde, gözlemler sıklıkla gürültü, eksik değerler ve yapısal düzensizlikler içerebilmekte; bu durum, makine öğrenimi modellerinin doğrudan uygulanması halinde tahmin

performansını olumsuz yönde etkileyebilmektedir. Bu nedenle, modelleme öncesinde gerçekleştirilen veri ön işleme, gürültü azaltma ve uygun özellik çıkarımı adımlarının, tahmin doğruluğunu artırmada kritik bir rol oynadığı vurgulanmaktadır. İşler ve ark. (2025), zaman serisi tabanlı tahmin problemlerinde uygulanan sistematik veri ön işleme yaklaşımlarının, model kararlılığını ve genellenabilirliğini önemli ölçüde iyileştirdiğini göstermiştir.

Her hat alt kümesinde seferler (half_trip_id) kronolojik olarak sıralanmış ve durakların gerçekleşme zamanına ($t^{(a)}$) göre yeniden düzenlenmesi sağlanmıştır. ISO-8601 formatındaki zaman damgaları, regresyon analizine uygunluk amacıyla epoch saniyesine dönüştürülmüş ve böylece zaman değişkeni sürekli, tek boyutlu ve aritmetik işlemlere elverişli bir forma kavuşturulmuştur. Bir seferdeki duraklar, $1, 2, 3, \dots, N$ şeklinde sıralanarak her durak için deterministik bir sıra indeksinin atanması sağlanmıştır.

Bu çalışmanın en kritik metodolojik bileşeni, klasik “ardışık durak farkı” yaklaşımının ötesine geçerek her seferdeki tüm durak çiftlerinin durak çifti uzayında modellenmesidir. Buna göre, bir seferdeki durakların oluşturduğu indeks kümesi üzerinde aşağıdaki tüm çiftler dikkate alınmıştır:

$$(i, j) \text{ öyle ki } 1 \leq i < j \leq N$$

Bu tanım sefer başına oluşan örnek sayısının doğrusal değil, karesel büyümesini sağlayan kombinyonel genişlemeye yol açmaktadır. Durak çifti sayısı:

$$\binom{N}{2} = \frac{N(N-1)}{2}$$

şeklinde olup, özellikle $\bar{N}$ değerinin büyük olduğu hatlarda veri hacminin asimptotik olarak $O(N^2)$ ölçeğinde genişlediğini göstermektedir. Bu genişleme, modelin yalnızca kısa menzilli segment dinamiklerini (ör. $i \rightarrow i + 1$) değil, aynı zamanda uzun menzilli kümülatif yolculuk sürelerini de (ör. $i \rightarrow j, j - i$ büyük) öğrenebilmesini mümkün kılarak daha kapsamlı bir tahmin uzayı sağlamaktadır. Her durak çifti için hedef değişken olan gerçek seyahat süresi şu şekilde tanımlanmıştır:

$$y_{i,j} = t_j - t_i$$

Bu ifade, seyahat süresinin deterministik bir zaman farkı olarak doğrudan ölçülmesine olanak tanır. Zaman damgası kaymalarından, GPS jitter

davranışından veya veri giriş hatalarından kaynaklanabilen $y_{i,j} < 0$ koşulunu sağlayan tüm gözlemler fiziksel olarak imkânsız kabul edilerek veri kümesinden çıkarılmıştır.

Gözlem vektörleri, yalnızca zaman farkına dayalı tekil bir ölçüm içermez; bunun ötesinde mekânsal, zamansal ve operasyonel bağlamı temsil eden çok boyutlu bağımsız değişkenlerle zenginleştirilmiştir. Her durak çifti gözlemi aşağıdaki özellik vektörü ile temsil edilmiştir:

$$X_{i,j} = [\text{route_id}, \text{direction_id}, \text{stopid}_i, \text{stopid}_j, i, j, (j - i), \text{month}, \text{day}, \text{hour}]$$

Bu özellik seti; hattın operasyonel kimliğini, durakların güzergâh üzerindeki göreceli konumlarını, duraksal mesafe derecesini $(j - i)$, ve seyahat süresi üzerindeki time-of-day etkilerini modellemeye imkân tanır. Böylece veri seti, hem düzensizlik ve gürültüden arındırılmış hem de seyahat dinamiklerinin çok boyutlu temsilini içeren yapısal bir forma ulaşmıştır.

Sonuç olarak, veri ön işleme aşaması; hat bazlı ayırtırmayı, zaman damgalarının sayısal dönüşümünü, durak çifti kombinasyonlarının karesel genişlemesini ve çok boyutlu özellik mühendisliğini içeren bütünleşik bir süreç olarak, karar ağaçları tabanlı makine öğrenmesi modellerine uygun yüksek çözünürlüklü bir veri uzayı oluşturmuştur.

Veri setinin bu şekilde yapılandırılmasının ardından, kullanılan bağımsız değişkenlerin hedef değişken olan seyahat süresi ile ilişkisini nicel olarak değerlendirmek amacıyla Cramér's V bağımlılık katsayısı hesaplanmıştır. Cramér's V, ki-kare bağımsızlık testine dayanan, iki kategorik değişken arasındaki ilişki gücünü 0 ile 1 arasında ölçen parametrik olmayan bir istatistiksel bağımlılık katsayısıdır (Cramér, 1946). Katsayının 1'e yaklaşması güçlü ilişkiye, 0'a yaklaşması ise bağımsızlığa işaret etmektedir. Bu analiz kapsamında *stop_id*, *order*, *stopid_target*, *order_target*, *month*, *day*, *hour* ve *order_diff* değişkenlerinin hedef değişken olan $y(i,j)$ üzerindeki etkisi incelenmiş; elde edilen sonuçlar Tablo 1'de sunulmuştur.

Tablo 1. Değişkenlerin Cramér's V değerleri

Değişken	Cramér's V	İlişki Düzeyi
stopid	0.42	Yüksek
order	0.42	Yüksek
stopid_target	0.22	Orta
order_target	0.22	Orta
month	0.17	Düşük
day	0.17	Düşük
hour	0.19	Düşük-Orta
order_diff	0.45	En Yüksek

Analiz sonuçları, seyahat süresi varyansının özellikle mekânsal konum bilgisi ve duraklar arasındaki segment mesafesini gösteren değişkenlerle daha güçlü ilişkili olduğunu ortaya koymaktadır. Buna karşılık *month* ve *day* gibi makro zaman bileşenleri, hedef değişkeni açıklamada oldukça sınırlı katkı sunmuştur. Bu durum, toplu taşımada seyahat süresi dinamiklerinin büyük ölçüde durak dizilimi, duraklar arası mesafe ve günün saatine bağlı olarak şekillendiğini gösteren literatürle tutarlıdır (Cats, 2011; Jenelius ve Koutsopoulos, 2013).

Modelleme yaklaşımı

Bu çalışmada, durak-çifti temelli seyahat süresi tahmini için karar ağaçları tabanlı beş farklı makine öğrenmesi modeli değerlendirilmiştir. Modeller, veri ön işleme aşamasında oluşturulan durak çifti yapısı üzerinden, her bir route_id hattı için ayrı ayrı eğitilmiştir. Böylelikle her hattın kendine özgü durak dizilimi, trafik akışı, güzergâh uzunluğu ve operasyonel değişkenliği modele daha tutarlı biçimde yansıtılmıştır. Ağaç tabanlı yöntemler, doğrusal olmayan ilişkileri yakalama kapasiteleri, büyük veriyle ölçeklenebilirlikleri ve kategorik değişkenleri etkili işleme yetenekleri sayesinde toplu taşımada ETA tahmini literatüründe yaygın olarak kullanılmaktadır (Vlahogianni ve ark., 2014). Bu kapsamda kullanılan modeller ve temel özellikleri aşağıda özetlenmiştir.

Karar ağacı (decision tree)

Karar ağacı modeli, veri uzayını art arda bölerek her bir alt bölgede sabit bir tahmin değeri üreten temel bir ağaç yapısıdır. Model, her düğümde belirli bir özellik X_k için bir eşik değeri c seçerek veri kümesini iki alt kümeye ayırır:

$$X_k \leq c \text{ ve } X_k > c$$

Regresyon problemlerinde her düğümdeki safsızlık, düğüm içindeki örneklerin ortalama kare hatası ile ölçülür. Bir düğümdeki örnek kümesi S ve hedef değerleri y_i olmak üzere safsızlık:

$$I(S) = \frac{1}{|S|} \sum_{i \in S} (y_i - \bar{y}_S)^2$$

şeklinde tanımlanır. Aday bölme (X_k, c) seçildiğinde veri $|S_L|$ ve $|S_R|$ alt kümelerine ayrılır. Karar ağacı, tüm olası bölmeler içinde ağırlıklı safsızlığı en aza indiren bölmeyi seçer:

$$\frac{|S_L|}{|S|} I(S_L) + \frac{|S_R|}{|S|} I(S_R)$$

Bu süreç tekrarlandıkça ağaç büyür ve her yaprakta sabit bir tahmin değeri elde edilir. Karar ağaçları yorumlanabilir olmaları nedeniyle sıklıkla taban model olarak kullanılsa da tek başlarına yüksek varyans eğilimlidir ve bu nedenle topluluk yöntemlerinin gerisinde kalırlar (Hastie, Tibshirani ve Friedman, 2009).

Rastgele orman (random forest)

Random Forest, çok sayıda karar ağacının bootstrap örneklemeleri üzerinde eğitilmesi ve tahminlerin ortalamasının alınması prensibine dayanır. Bu yöntem aşırı öğrenmeye karşı oldukça dirençlidir ve yüksek boyutlu, kategorik değişken yoğunluklu veri kümelerinde kararlı performans göstermektedir (Breiman, 2001). Bu nedenle toplu taşıma seyahat süresi literatüründe sıkça tercih edilen güçlü bir ansambl yöntemidir.

XGBoost

XGBoost, gradyan artırmalı karar ağaçlarının düzenlileştirme terimleriyle güçlendirilmiş optimize edilmiş bir sürümüdür. Ardışık her ağaç, bir önceki modelin hatasını azaltacak yönde üretilir. XGBoost, yüksek performansı, eksik veri toleransı ve L1–L2 düzenlileştirme mekanizmaları nedeniyle farklı ulaşım uygulamalarında başarılı sonuçlar vermiştir (Chen ve Guestrin, 2016). Büyük ve karmaşık veri setlerinde stabil bir öğrenme davranışı sunması, ETA tahmin problemleri için önemli bir avantajdır.

LightGBM

LightGBM, histogram tabanlı bölme algoritması ve yaprak odaklı (leaf-wise) büyüme stratejisiyle, boosting yöntemleri arasında en hızlı modellerden biridir. Özellikle çok geniş veri setlerinde hesaplama maliyetini önemli ölçüde azaltmaktadır (Ke ve ark., 2017). ETA tahmini gibi yüksek boyutlu ve çok sayıda durak çifti örneği içeren problemlerde bu hız ve ölçeklenebilirlik önemli bir üstünlük sağlamaktadır.

CatBoost

CatBoost, kategorik değişken yoğunluğu yüksek veri setleri için özel olarak tasarlanmıştır ve bu özelliğiyle toplu taşımada yaygın kullanılan stop_id, direction_id, route_id gibi alanlarda önemli performans kazancı sağlamaktadır. CatBoost'un en kritik bileşeni, kategorik değişkenlerin klasik hedef kodlama yöntemlerinden farklı olarak sözde zamanlı bir işlemle dönüştürülmesidir. Böylece hedef sızıntısı oluşmaz ve model kategorik sütunlardan daha güvenilir bilgi öğrenir (Prokhorenkova ve ark., 2018). Literatürde CatBoost'un ulaşım ve seyahat süresi tahmininde yüksek doğruluk ürettiği çeşitli çalışmalarda rapor edilmiştir (Moosavi ve ark., 2023).

Tüm modeller aynı eğitim–test stratejisi ile değerlendirilmiş ve zaman sırası korunarak veri %80 eğitim, %20 test olacak şekilde ayrılmıştır. Performans değerlendirmesinde Ortalama Mutlak Hata (MAE), Kök Ortalama Kare Hata (RMSE) ve R^2 metrikleri kullanılmıştır. Bu metrikler seyahat süresi tahmin çalışmalarında yaygın olarak tercih edilmekte olup RMSE'nin büyük hataları cezalandıran yapısı, ETA tahmininde önemli bir avantaj sunmaktadır (Vlahogianni ve ark., 2014).

DeneySEL Kurulum

Bu çalışmada karar ağaçları tabanlı makine öğrenmesi modellerinin performanslarını karşılaştırmak amacıyla tüm deneyler aynı donanım ve yazılım ortamında yürütülmüştür. Deneyler, Windows 11 işletim sistemi üzerinde çalışan Intel Core i7-14700 2.10 GHz işlemci ve 32 GB RAM donanımına sahip bir makinede gerçekleştirilmiştir. Tüm modeller Python 3.10.4 ortamında uygulanmış; XGBoost, LightGBM ve CatBoost kütüphaneleri güncel kararlı sürümleriyle kullanılmıştır. Eğitim süreçlerinde GPU kullanılmamış, tüm algoritmalar yalnızca CPU üzerinde değerlendirilmiştir.

Veri seti, her bir route_id için bağımsız olarak ele alınmış ve modelleme süreci hat bazında yürütülmüştür. Her hat, daha önce açıklanan durak çifti dönüşüm adımının ardından yüksek hacimli gözlem matrislerine ayrılmıştır. Veri setlerinin zaman boyutunda taşıdığı ardışıklık nedeniyle rastgele karıştırma uygulanmamış; zaman serisi tutarlılığını korumak amacıyla veriler kronolojik sıraya göre yapılandırılmıştır.

Model performanslarının daha güvenilir ve genellenebilir biçimde değerlendirilmesi amacıyla, her hat için zaman serisi yapısına uygun 5 katlı çapraz doğrulama (5-fold cross-validation) uygulanmıştır. Çapraz doğrulama sürecinde, her bir katman kronolojik bütünlük korunacak şekilde oluşturulmuş ve böylece ileriye dönük bilgi sızıntısının önüne geçilmiştir.

Çapraz doğrulama sonucunda elde edilen performans metrikleri katmanlar üzerinden aritmetik ortalama alınarak raporlanmış, katmanlar arasındaki değişkenlik ise standart sapma ile ifade edilmiştir. Ayrıca, modellerin gerçek operasyon koşullarındaki davranışlarını değerlendirmek amacıyla veri setinin son %20'lik bölümü bağımsız bir test kümesi olarak ayrılmıştır. Bu yaklaşım, hem model kararlılığının hem de zamansal genellenebilirliğin daha sağlıklı biçimde analiz edilmesini amaçlamaktadır.

Model hiperparametreleri

Hiperparametre ayarları, her model için literatürde yaygın olarak önerilen temel konfigürasyonlar dikkate alınarak belirlenmiş; aşırı uyum (overfitting) riskini azaltmak amacıyla ağaç derinlikleri sınırlandırılmış ve yaprak başına minimum örnek sayıları artırılmıştır. CatBoost modeli özelinde kategorik değişkenlerin doğal işlenmesi etkinleştirilmiş; böylece ek kodlama adımlarına ihtiyaç duyulmadan, durak kimlikleri ve zamansal kategoriler gibi ayrık değişkenlerin model tarafından doğrudan öğrenilmesi sağlanmıştır. Benzer şekilde LightGBM modelinde, büyük ölçekli veri kümeleri için bellek verimliliği ve eğitim hızı avantajı sunan histogram tabanlı bölünme algoritması tercih edilmiştir. XGBoost modeli için ise tree_method=hist seçeneği kullanılarak hem hesaplama maliyetleri azaltılmış hem de büyük veri hacimleri altında daha kararlı bir eğitim süreci hedeflenmiştir.

Model performanslarının değerlendirilmesinde MAE, RMSE ve R^2 metrikleri birlikte kullanılmıştır. Bu metriklerin tercih edilme nedeni; MAE'nin ortalama mutlak hata büyüklüğünü doğrudan yorumlanabilir

biçimde sunması, RMSE'nin büyük hatalara daha duyarlı olması sayesinde aşırı sapmaların etkisini ortaya koyabilmesi ve R^2 'nin modelin hedef değışkendeki varyansı ne ölçüde açıkladığını göstermesidir. Bu üç metriğin birlikte raporlanması, modellerin yalnızca ortalama performanslarını değil, hata dağılımlarını ve açıklayıcılık kapasitelerini de çok boyutlu olarak değerlendirmeye olanak tanımaktadır.

Tüm modeller, aynı eğitim–test ayrımı, aynı özellik seti ve aynı veri ön işleme prosedürü altında değerlendirilmiş olup, deneysel sonuçların karşılaştırılabilirliği titizlikle korunmuştur. Ayrıca, modeller arası adil bir karşılaştırma yapılabilmesi amacıyla, herhangi bir modele özgü avantaj sağlayabilecek özel veri manipölasyonlarından veya model dışı ek bilgi kullanımından kaçınılmıştır. Bu yaklaşım, gözlenen performans farklarının doğrudan algoritmik özellikler ve öğrenme stratejilerinden kaynaklandığının daha net biçimde yorumlanmasına imkân tanımaktadır. Buna ek olarak, çapraz doğrulama sürecinde elde edilen performans ölçütlerinin katmanlar arasındaki değışkenliği incelenerek, modellerin zamansal kararlılıkları ve genellenebilirlikleri de değerlendirilmiştir. Katmanlar arasında düşük varyans sergileyen modellerin, farklı zaman dilimlerinde daha tutarlı tahminler üretebildiği gözlemlenmiş; bu durum, özellikle operasyonel kullanıma yönelik ETA tahmin senaryolarında model seçimi açısından önemli bir kriter olarak ele alınmıştır. Böylece, yalnızca tek bir zaman dilimine özgü yüksek performans gösteren modeller yerine, farklı koşullar altında istikrarlı sonuçlar üretebilen yaklaşımların önemi vurgulanmıştır.

Tablo 2. Model hiperparametreleri

<i>Model</i>	<i>iterations/n_</i> <i>estimators</i>	<i>max_</i> <i>depth</i>	<i>learning_</i> <i>rate</i>	<i>Diğer Parametreler</i>
<i>Decision Tree</i>	-	6	-	Varsayılan parametreler
<i>Random Forest</i>	100	6	-	Varsayılan parametreler
<i>CatBoost</i>	100	6	0.1	loss_function=RMSE, verbose=100
<i>XGBoost</i>	100	6	0.1	objective=reg:squarederror, colsample_bytree=0.3, alpha=10
<i>LightGBM</i>	100	6	0.1	objective=regression

Bulgular

Bu bölümde, üç farklı güzergâh (Hat 01, Hat 111 ve Hat CT2) için durak çifti tabanlı seyahat süresi tahmin modellerinden elde edilen sonuçlar sunulmaktadır. Modeller; MAE, RMSE, R^2 ve eğitim süresi bakımından karşılaştırılmış ve her hat için performanslar ayrı tablolar halinde rapor edilmiştir.

Hat 01'e ait sonuçlar incelendiğinde (Tablo 3), ансамбл yöntemlerin tek bir karar ağacına göre belirgin biçimde daha iyi genelleme performansı gösterdiği görülmektedir. Özellikle CatBoost, RMSE ve MAE değerlerinde tüm modellerin önüne geçerek bu hat için en isabetli tahminleri üretmiştir. XGBoost ve LightGBM eğitim süresi açısından daha hızlı olmakla birlikte, doğruluk seviyesi bakımından CatBoost'un gerisinde kalmıştır. Rastgele Orman ise yüksek varyanslı hata davranışı ve daha uzun eğitim süresi nedeniyle diğer modellerden ayrılmaktadır.

Tablo 3. 01 Hattı için model performans karşılaştırması

	<i>RMSE</i>	<i>MAE</i>	<i>R²</i>	<i>Tekil Eğitim Süresi (s)</i>
<i>Catboost</i>	237.46 (std: 21.99)	157.08 (std: 13.49)	0.86 (std: 0.02)	0,7
<i>Xgboost</i>	239.65 (std: 22.75)	157.73 (std: 10.68)	0.86 (std: 0.02)	0.2
<i>Lightgbm</i>	242.16 (std: 24.00)	159.85 (std: 13.15)	0.85 (std: 0.02)	0.3
<i>Karar Ağacı</i>	249.55 (std: 29.86)	161.96 (std: 14.44)	0.85 (std: 0.02)	0.2
<i>Rastgele Orman</i>	247.87 (std: 30.63)	160.36 (std: 15.22)	0.85 (std: 0.02)	10.9

Hat 111'e ilişkin sonuçlar (Tablo 4), genel eğilimin korunduğunu göstermektedir. CatBoost, özellikle düşük MAE değeriyle kısa mesafeli durak çiftlerinde kararlı bir öğrenme performansı sergilemiştir. Bu hatta XGBoost, CatBoost'a göre daha düşük eğitim süresine sahip olmasına rağmen doğruluk metriklerinde aynı üstünlüğü sağlayamamıştır. LightGBM ise hız açısından avantajlı olsa da hata değerleri bakımından orta seviyede kalmıştır.

Tablo 4. 111 Hattı için model performans karşılaştırması

	RMSE	MAE	R ²	Tekil Eğitim Süresi (s)
Catboost	161.31 (std: 9.22)	98.31 (std: 3.28)	0.85 (std: 0.02)	1,3
Xgboost	162.53 (std: 9.23)	100.80 (std: 3.08)	0.84 (std: 0.02)	0.3
Lightgbm	164.94 (std: 11.19)	100.01 (std: 3.56)	0.84 (std: 0.02)	0.4
Karar Ağacı	170.87 (std: 8.15)	104.97 (std: 3.75)	0.82 (std: 0.01)	0.4
Rastgele Orman	169.77 (std: 8.20)	104.68 (std: 3.83)	0.82 (std: 0.01)	30.2

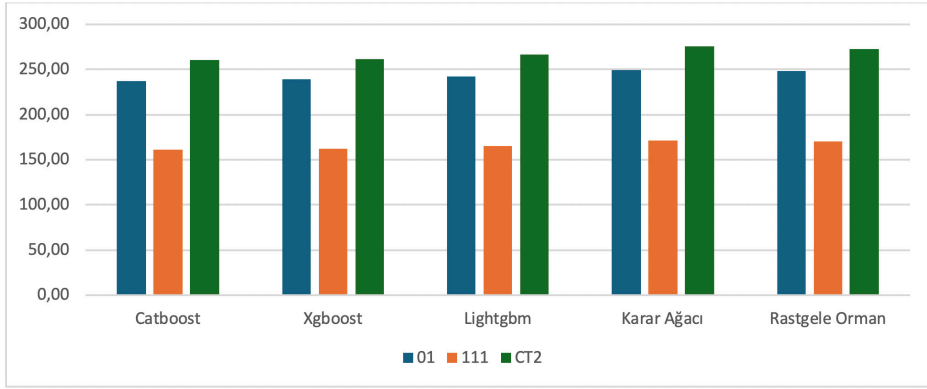
Hat CT2 için elde edilen bulgular (Tablo 5), duraklar arası mesafe değişkeninin ve trafik koşullarının daha heterojen olması nedeniyle hata metriklerinin mutlak değer olarak yükseldiğini göstermektedir. Buna rağmen CatBoost bu hatta da en yüksek R² değerini sağlayarak model kararlılığını korumuştur. XGBoost ve LightGBM yine daha kısa eğitim süreleri sunmakla birlikte tahmin doğruluğu bakımından CatBoost'un gerisinde kalmıştır. Tek bir karar ağacı ise her üç hatta da düşük doğruluk ve yüksek varyans nedeniyle tatmin edici performans sergileyememiştir.

Tablo 5. CT2 hattı için model performans karşılaştırması

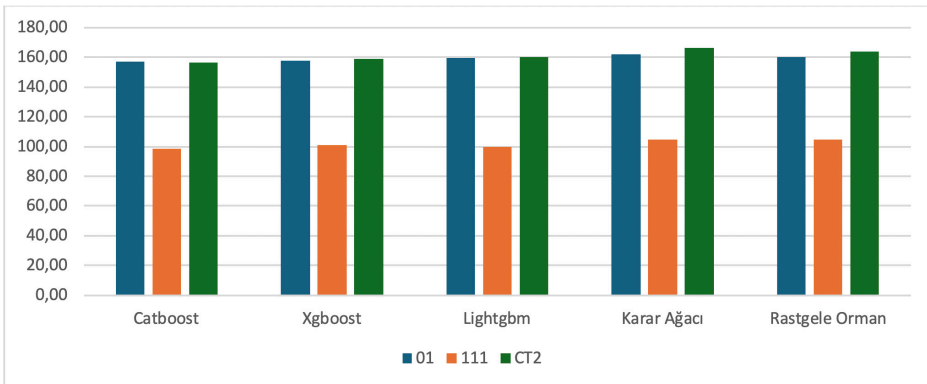
	RMSE	MAE	R ²	Tekil Eğitim Süresi (s)
Catboost	260.98 (std: 38.55)	156.52 (std: 15.81)	0.89 (std: 0.03)	0.3
Xgboost	261.94 (std: 37.95)	159.14 (std: 16.90)	0.88 (std: 0.03)	0.1
Lightgbm	266.27 (std: 31.92)	160.43 (std: 14.42)	0.88 (std: 0.02)	0.2
Karar Ağacı	275.43 (std: 14.76)	166.65 (std: 36.87)	0.87 (std: 0.03)	0.1
Rastgele Orman	272.33 (std: 15.10)	163.91 (std: 37.96)	0.88 (std: 0.03)	3.9

Şekil 1, Şekil 2 ve Şekil 3'te sunulan RMSE, MAE ve R² metrikleri birlikte değerlendirildiğinde, algoritmaların hatlar arası performanslarının yalnızca mutlak hata seviyeleri açısından değil, aynı zamanda kararlılık ve genellenebilirlik açısından da farklılık gösterdiği görülmektedir. Özellikle hatlar arasında gözlenen hata dağılımlarının değişkenliği, modellerin veri yapısındaki heterojenliğe verdikleri tepkilerin birbirinden ayrıştığını ortaya koymaktadır.

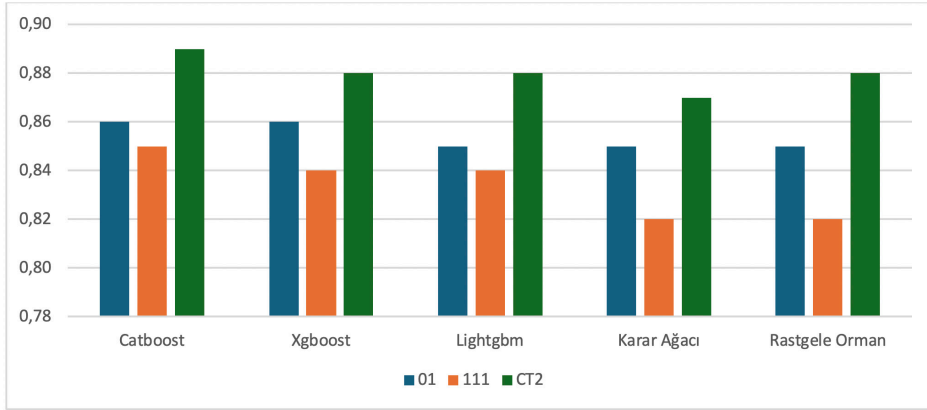
Gradyan artırma tabanlı yöntemler, tüm hatlarda daha dengeli hata profilleri ve daha yüksek açıklama gücü sergileyerek, doğrusal olmayan ilişkileri ve değişkenler arası karmaşık etkileşimleri daha etkin biçimde modelleyebilmiştir. Çapraz doğrulama sonuçları, bu yöntemlerin veri dağılımındaki değişimlere karşı daha dayanıklı olduğunu ve zamansal yapıdan kaynaklanan belirsizlikleri daha etkin biçimde yakalayabildiğini ortaya koymaktadır. Bu bağlamda, algoritmalar arasındaki performans farklarının yalnızca model karmaşıklığıyla açıklanamayacağı; kullanılan öğrenme stratejileri, hata fonksiyonları ve veri yapısına uyum yetenekleriyle doğrudan ilişkili olduğu anlaşılmaktadır.



Şekil 1. Modellerin hat bazında RMSE karşılaştırması



Şekil 2. Modellerin hat bazında MAE karşılaştırması



Şekil 3. Modellerin hat bazında R^2 karşılaştırması

Genel olarak tüm hatlarda gözlenen sonuçlar, CatBoost'un kategorik değişkenleri doğal işleme kabiliyeti, sıralı artırmalı öğrenme yaklaşımı ve gürültülü/veri dengesizliği içeren veri kümelerinde sağladığı stabilizasyon sayesinde ETA tahmininde en yüksek doğruluk seviyesini sunduğunu göstermektedir.

Sonuç

Bu çalışma, kentsel toplu taşıma sistemlerinde duraklar arası varış süresi tahmini (ETA) için karar ağaçları tabanlı makine öğrenmesi modellerinin sistematik bir karşılaştırmasını sunmuş ve MBTA'nın "Bus Arrival/Departure Times (2024)" veri seti durak çifti temelli bir yapıya dönüştürülerek yüksek çözünürlüklü bir tahmin uzayı oluşturulmuştur. Bu dönüşüm, ardışık duraklar yerine tüm durak çiftlerinin modele dahil edilmesi sayesinde hem kısa menzilli hem de uzun menzilli seyahat dinamiklerinin bütüncül biçimde öğrenilmesini sağlamıştır. Beş farklı modelin üç ayrı güzergâh üzerinde değerlendirilmesi sonucunda, ансамбль yöntemlerinin tekil karar ağacı modeline kıyasla belirgin üstünlük sağladığı görülmüştür. Özellikle CatBoost, kategorik değişkenleri doğal olarak işleme kapasitesi, sıralı artırmalı öğrenme mekanizması ve aşırı öğrenme azaltan düzenleme özellikleri sayesinde tüm güzergâhlarda en yüksek doğruluk seviyesini elde etmiştir. R^2 değerlerinin 0.85–0.89 aralığında gerçekleşmesi, modelin hem güçlü hem de istikrarlı bir açıklayıcılık sunduğunu göstermektedir. Eğitim süresinin XGBoost ve LightGBM'e kıyasla daha uzun olmasına rağmen, CatBoost'un doğruluk-hız dengesi açısından pratik uygulamalara uygun bir çözüm sunduğu belirlenmiştir.

Elde edilen bulgular, durak çifti tabanlı karar ağacı modellerinin, özellikle CatBoost mimarisinin, gerçek zamanlı ETA tahmini yapan toplu taşıma sistemlerinde etkin biçimde kullanılabileceğini göstermektedir. Gelecek çalışmalarda trafik yoğunluğu, hava durumu ve yolcu yükü gibi ek değişkenlerin modele entegrasyonu ile tahmin doğruluğunun daha da artırılması mümkündür. Ayrıca diğer kentlerdeki veri setlerinin test edilmesi, modelin genelleme kapasitesinin değerlendirilmesine katkı sağlayacaktır.

Çıkar çatışması beyanı

Yazarlar arasında herhangi bir çıkar çatışması bulunmamaktadır.

Etik kurul beyanı

Bu araştırma, yalnızca MBTA tarafından sağlanan açık erişimli “Bus Arrival/Departure Times (2024)” veri seti kullanılarak gerçekleştirilmiştir. Çalışmada insan veya hayvan denekleri üzerinde herhangi bir deneysel uygulama yapılmamış, kişisel veri toplanmamış ve müdahale içeren bir süreç yürütülmemiştir. Bu nedenle etik kurul onayı gerekmemektedir. Araştırma ve yayın etiği ilkelerine uygun olarak yürütülmüştür.

Kaynaklar

- [1] Büchel, B. ve Corman, F. (2020). Review on statistical modeling of travel time variability for road-based public transport. *Frontiers in Built Environment*, 6, Article 70.
- [2] Vlahogianni, E. I., Karlaftis, M. G. ve Golias, J. C. (2014). Short-term traffic forecasting: Where we are and where we’re going. *Transportation Research Part C*, 43, 3–19.
- [3] Jenelius, E. ve Koutsopoulos, H. N. (2013). Travel time estimation for urban road networks using low-frequency probe vehicle data. *Transportation Research Part B: Methodological*, 53, 64–81.
- [4] Sun, D., Luo, H., Fu, L., Liu, W., Liao, X. ve Zhao, M. (2007). Predicting bus arrival time on the basis of global positioning system data. *Transportation Research Record*, 2034(1), 62–72.
- [5] Fan, W. ve Gurmu, Z. (2015). Dynamic travel time prediction models for buses using GPS data. *Transportation Research Part C*, 59, 18–32.
- [6] Moosavi, S. M. H., Aghaabbasi, M., Yuen, C. W. ve Jahed Armaghani, D. (2023). Evaluation of applicability and accuracy of bus travel time

- prediction in high and low-frequency bus routes using tree-based machine learning techniques. *Journal of Soft Computing in Civil Engineering*, 7(2), 74–97.
- [7] Hastie, T., Tibshirani, R. ve Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). *Springer*.
- [8] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- [9] Chen, T. ve Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM.
- [10] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. ve Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *In Advances in Neural Information Processing Systems (NeurIPS 2017)* (pp. 3149–3157).
- [11] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. ve Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *In Advances in Neural Information Processing Systems (NeurIPS 2018)* (pp. 6638–6648).
- [12] Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.
- [13] Cats, O. (2011). *Dynamic Modelling of Transit Operations and Passenger Decisions*. TU Delft.
- [14] Okutucu, H. ve Kaya, Ş. M. (2025). Traffic Density Estimation with IoT Edge Computing In Smart City: A Case Study. *Anadolu Bil Meslek Yüksekokulu Dergisi*, 20(72), 189-215.
- [15] Esmaili Jobani, A. ve Kaya, Ş. M. (2025). Hybrid IoT and AI-based Solution for Energy Management in Data Centres under Various Climate Conditions. *Anadolu Bil Meslek Yüksekokulu Dergisi*, 20(72), 107-124.
- [16] Kaya, Ş. M., Güneş, A. ve Erdem A. (2021). "A Smart Data Pre-Processing Approach by Using ML Algorithms on IoT Edges: A Case Study," *2021 International Conference on Artificial Intelligence of Things (ICAIoT)*, Nicosia, Turkey, 2021, pp. 36-42

- [17] İşler, B., Kaya, Ş.M. ve Kılıç, F.R. (2025). Fog-Enabled Machine Learning Approaches for Weather Prediction in IoT Systems: A Case Study. *Sensors* 2025, 25, 4070.

İnternet Kaynakları

Massachusetts Bay Transportation Authority. (2024). Bus Arrival/Departure Times Dataset. Retrieved from <https://gis.data.mass.gov/datasets/96c77138c3144906bce93d0257531b6a/abo>