BURSA TEKNİK ÜNİVERSİTESİ
MÜHENDİSLİK VE DOĞA BİLİMLERİ FAKÜLTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ
2023-2024 BAHAR DÖNEMİ
VERİ MADENCİLİĞİNE GİRİŞ PROJE RAPORU

LSTM MODELİ İLE TRAFİK YOĞUNLUĞUNUN ZAMAN SERİSİ TAHMİNİ

FURKAN YAYLAZ

İÇİNDEKİLER

Giriş

Projenin amacı, Minneapolis ve St. Paul arasındaki belirli bir noktada Minnesota'daki trafik akışının hacmini öngörmektir. Bu projede, önceki veriler dikkate alınarak gelecekteki 200 saatlik bir zaman diliminin trafik yoğunluğunu tahmin eden bir LSTM modeli oluşturmak hedeflenmektedir.

Bu projede, LSTM modeliyle bir dizi adımda trafik yoğunluğunu tahmin etmek için zaman serisi verilerini kullanarak gelecekteki trafik durumunu öngörme yeteneği araştırılmıştır. Modelin doğruluğunu değerlendirmek için, eğitim ve test veri setleri kullanılarak modelin performansı incelenmiş ve sonuçlar raporda sunulmuştur.

Raporun devamında, kullanılan veri setinin tanıtımı, modelin tasarımı ve mimarisi, eğitim ve test süreçleri, elde edilen sonuçların analizi ve değerlendirmesi, karşılaşılan zorluklar gibi konular detaylı bir şekilde ele alınmıştır. Bu çalışma, trafik yoğunluğunu tahmin etmek için LSTM tabanlı modelin etkinliğini değerlendirmek için bir temel sağlamış ve benzer uygulamalar için bir çerçeve oluşturmuştur.

Veri Seti Tanımı

Metro Otoyolu Trafik Hacmi Veri Seti, Minnesota Ulaştırma Bakanlığı (DoT) ATR istasyonu 301'deki saatlik Otoyol 94 Batı yönündeki trafik hacmini içerir. Bu istasyon, Minneapolis ve St. Paul arasında yaklaşık olarak ortada yer alır. Veri seti, trafik hacmi üzerinde etkileri olan saatlik hava durumu özellikleri ve tatil bilgilerini de içerir.

Toplamda 48.204 kayıttan oluşan bu veri setinde 9 özellik bulunmaktadır:

1. holiday: Tatil günlerini temsil eder. Tatil günü değilse 'None' olarak belirtilir.
2. temp: Sıcaklığı Kelvin cinsinden ifade eder.
3. rain_1h: Son saat içindeki yağış miktarını milimetre cinsinden gösterir.
4. snow_1h: Son saat içindeki kar miktarını milimetre cinsinden gösterir.
5. clouds: Bulut örtüsünün yüzde olarak ifadesidir.
6. weather_main: Hakim hava koşullarını kısa bir açıklama ile özetler.
7. weather_description: Mevcut hava durumunu daha detaylı bir şekilde açıklar.
8. date_time: Zaman damgasını 'Y/m/d H:M:S' formatında gösterir.
9. traffic_volume: Son saat içinde geçen araçların sayısını ifade eder.
10. Bu veri setindeki hedef özellik, trafik hacminin saatlik tahminidir.

Keşifçi Veri Analizi (EDA)

Eksik değer analizi

Veri setindeki eksik değerlerin sayısını incelediğimizde, sadece 'holiday' adlı sütunda 48143 eksik değer bulunmaktadır. Diğer taraftan, 'temp', 'rain_1h', 'snow_1h', 'clouds_all', 'weather_main', 'weather_description' ve 'date_time' sütunlarında herhangi bir eksik değer bulunmamaktadır. 'traffic_volume' sütununda da eksik değer bulunmamaktadır.

Bu eksik değerlerin neden kaynaklandığını anlamak ve veri setinin doğruluğunu ve güvenilirliğini sağlamak için, eksik değerleri daha detaylı bir şekilde incelemek gerekmektedir. Eksik değerlerin yapısını anlamak ve uygun bir doldurma veya çıkarım stratejisi geliştirmek, modelin doğruluğunu ve tahmin performansını artırmaya yardımcı olacaktır.

	count	mean	std	min	25%	50%	75%	max
temp	48204.0	281.205870	13.338232	0.0	272.16	282.45	291.806	310.07
rain_1h	48204.0	0.334264	44.789133	0.0	0.00	0.00	0.000	9831.30
snow_1h	48204.0	0.000222	0.008168	0.0	0.00	0.00	0.000	0.51
clouds_all	48204.0	49.362231	39.015750	0.0	1.00	64.00	90.000	100.00
traffic_volume	48204.0	3259.818355	1986.860670	0.0	1193.00	3380.00	4933.000	7280.00

Şekil 1

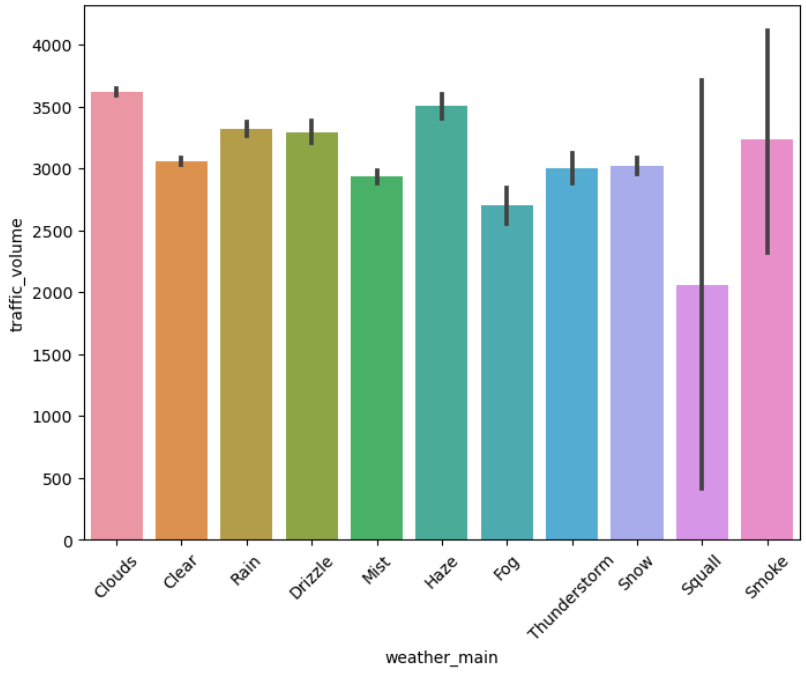
Şekil 1'de gösterilen sayısal değişkenlerin istatistiksel özetine göre, veri setindeki bazı ana özellikler aşağıdaki gibidir:

- **temp (Sıcaklık)**: Ortalama sıcaklık 281.21 Kelvin'dir ve standart sapması 13.34'tür. Minimum sıcaklık değeri 0.0 olarak ölçülmüştür.
- **rain_1h (Yağış Miktarı)**: Ortalama yağış miktarı 0.33 mm'dir ve standart sapması 44.79'dur. En yüksek yağış miktarı 9831.30 mm olarak kaydedilmiştir.
- **snow_1h (Kar Miktarı)**: Ortalama kar miktarı çok düşük düzeydedir, 0.0002 mm'dir. Standart sapması ise 0.008'dir.

- **clouds_all (Bulut Kaplama Yüzdesi)**: Ortalama bulut kaplama yüzdesi %49.36'dır. Minimum değer 0 (yani bulutsuz) iken, maksimum değer %100'dür.
- **traffic_volume (Trafik Hacmi)**: Ortalama trafik hacmi 3259.82'dir, ancak standart sapması oldukça yüksektir (1986.86). En yüksek trafik hacmi 7280 olarak kaydedilmiştir.

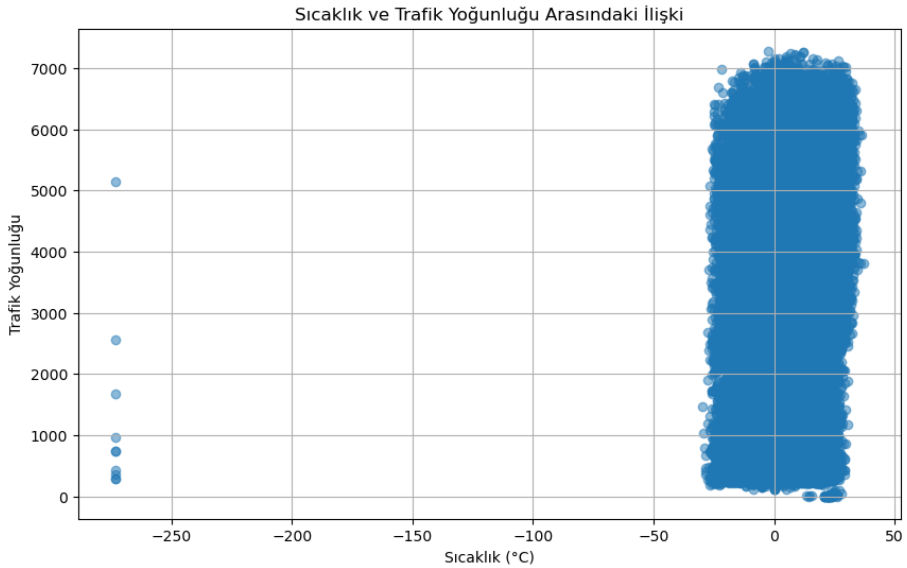
Veri setindeki analiz sırasında, 'weather_main' ve 'weather_description' değişkenlerinin benzer bilgileri içerdiği tespit edilmiştir. Dolayısıyla, aynı bilgileri iki farklı değişken içinde tutmanın gereksiz olduğu sonucuna varılarak, 'weather_description' değişkeni veri setinden kaldırılmıştır. Ayrica ‘temp’ değişkeninin birimi Kelvin’den Celciyus birimine çevrilerek anlaşılması daha kolay hale getirilmiştir.

Veri seti incelendiğinde, verilerin toplandığı zaman aralığının 2 Ekim 2012, 09:00 ile 30 Eylül 2018, 23:00 tarihleri arasında olduğu tespit edilmiştir.



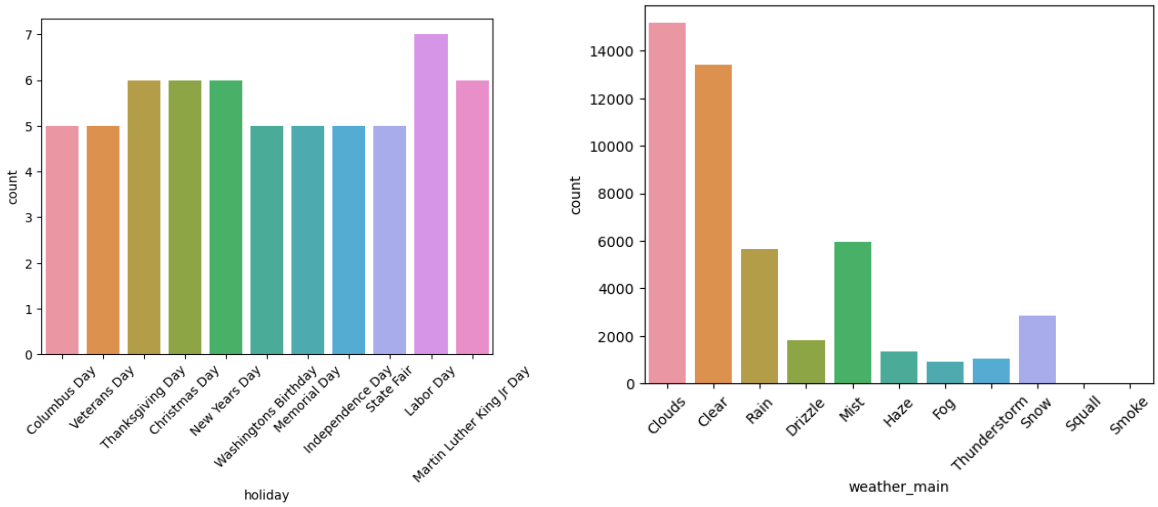
Şekil 2

Şekil 2’ye bakıldığında trafik yoğunluğunun hava durumu ile ilişkisi gösterilmektedir. Hava durumunun sisli, fırtınalı olduğundaki trafik ortalaması, diğer hava durumlarıyla kıyaslayınca daha düşük olduğu gözlemlenmiştir.



Şekil 3

Şekil 3'te sıcaklık ve trafik yoğunluğu arasındaki ilişkiye bakacak olursak sıcaklık değişkeninde aykırı değerler olduğu gözlemlenmektedir ve -40° ile +40° arasında dağılıma sahiptir.



Şekil 4

Şekil 4'te *grab_col_names()* fonksiyonu aracılığıyla tespit edilen kategorik değişkenlerin sayıları gösterilmektedir. Zaman aralığının yaklaşık 6 yıl olduğu için 'holiday' değişkenindeki sayılar gerçek değerleri yansıttığı söyenebilir. 'weather_main' değişkenine bakıldığında 2012-2018 yılları arasında I-94 otoyolunda havanın çoğunlukla bulutlu ve açık olduğu ve bu havalarda trafik hacminin daha fazla olduğu sonucu Şekil 2 grafiğine bakılarak çıkarılmaktadır.

Tekrarlayan Kayıtlar

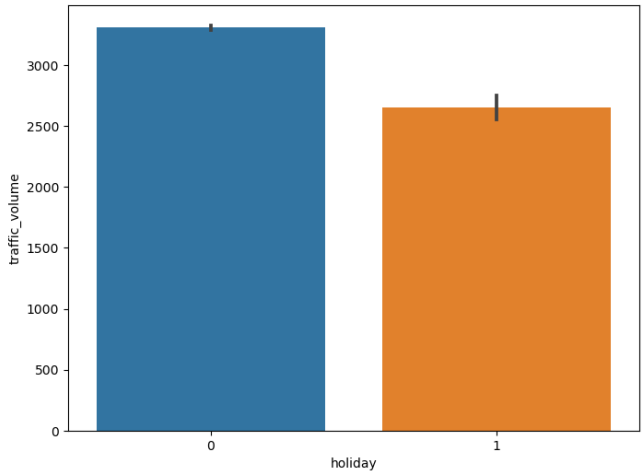
Veri setine eklenen kayıtlar 1 saatlik aralıklarla kaydedilmiştir, ancak daha detaylı bir inceleme sonucunda saatlik tekrar eden kayıtlar tespit edilmiştir. Toplamda 7629 tekrarlayan saatlik kayıt bulunmuştur. Bu veri tutarsızlığı, modelin doğruluğunu olumsuz etkilemiş ve verilerin güvenilirliği konusunda endişelere yol açmıştır. Bu nedenle, tekrarlanan saatlik kayıtların uygun bir şekilde işlenmesi ve veri tutarlılığının sağlanması amacıyla tekrarlayan kayıtlar veri setinden çıkarılarak daha genellenebilir tahminler yapması mümkün hale gelmiştir.

Özellik Çıkarımı

Veri setindeki 'date_time' sütunu kullanılarak yeni tarih özellikleri oluşturulmuştur. Bu özellikler, zaman serisi analizine uygun hale getirilerek modelin trafik hacmini daha doğru tahmin etmesine yardımcı olacaktır. Aşağıda, oluşturulan tarih özellikleri listelenmiştir:

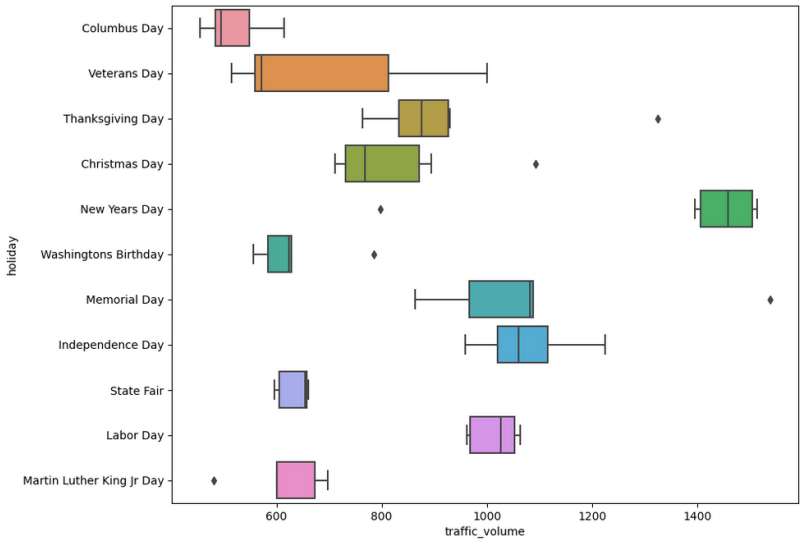
- **month**: Ay bilgisi.
- **day_of_month**: Ayın günü bilgisi.
- **day_of_year**: Yılın günü bilgisi.
- **day_of_week**: Haftanın günü bilgisi.
- **year**: Yıl bilgisi.
- **is_weekend**: Haftasonu olup olmadığını belirtir (1: Cumartesi veya Pazar, 0: Diğer günler).
- **is_month_start**: Ayın başlangıcı olup olmadığını belirtir (1: Evet, 0: Hayır).
- **is_month_end**: Ayın sonu olup olmadığını belirtir (1: Evet, 0: Hayır).

'Holiday' değişkeni, tatil günlerini temsil eden 1 ve tatil olmayan günleri temsil eden 0 değerlerine dönüştürülmüştür. Bu dönüşüm, modelin tatil günlerinin trafik hacmi üzerindeki etkisini daha iyi anlamasına olanak tanımaktadır.



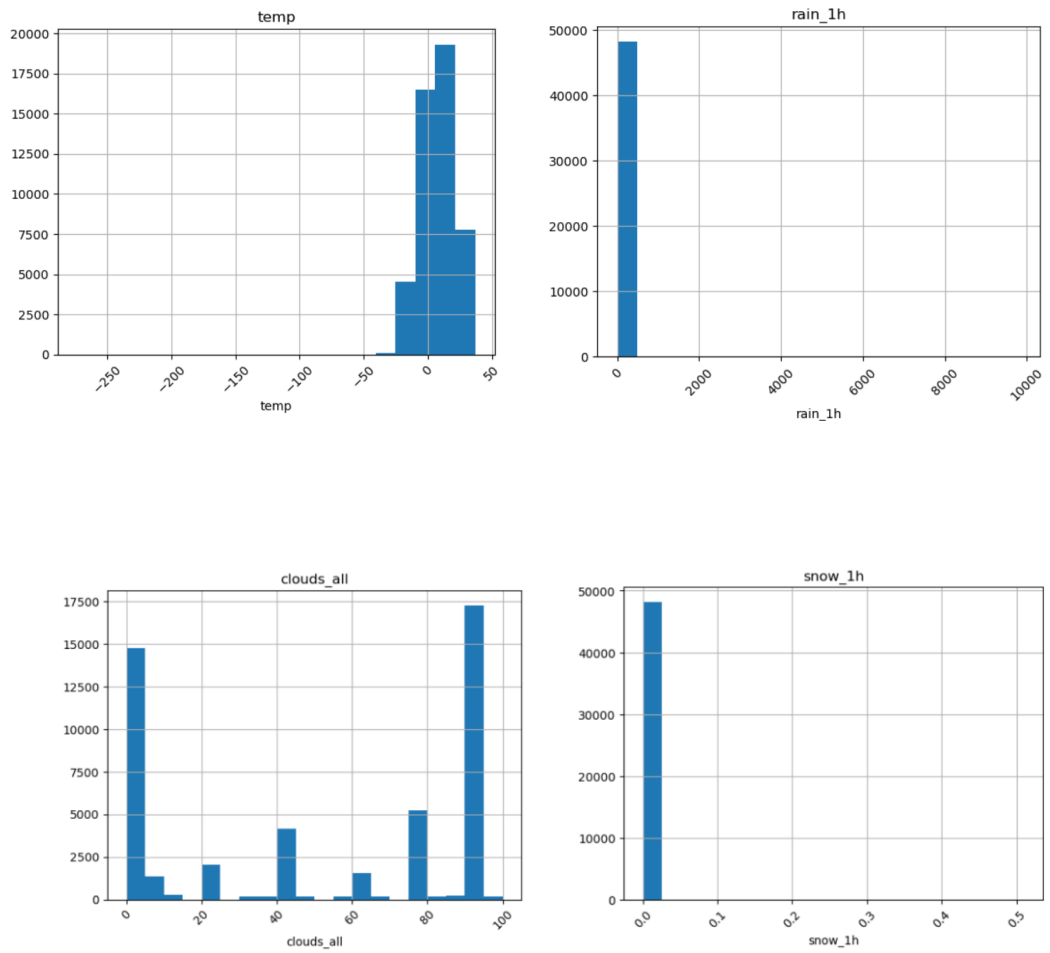
Şekil 5

'holiday' değişkeni değiştirildikten sonra resmi bayramlarda ve diğer günlerdeki trafik hacmine bakıldığında I-94 otoyolunun resmi günlerdeki ortalamasının daha düşük olduğu Şekil 5'te gösterilmiştir.



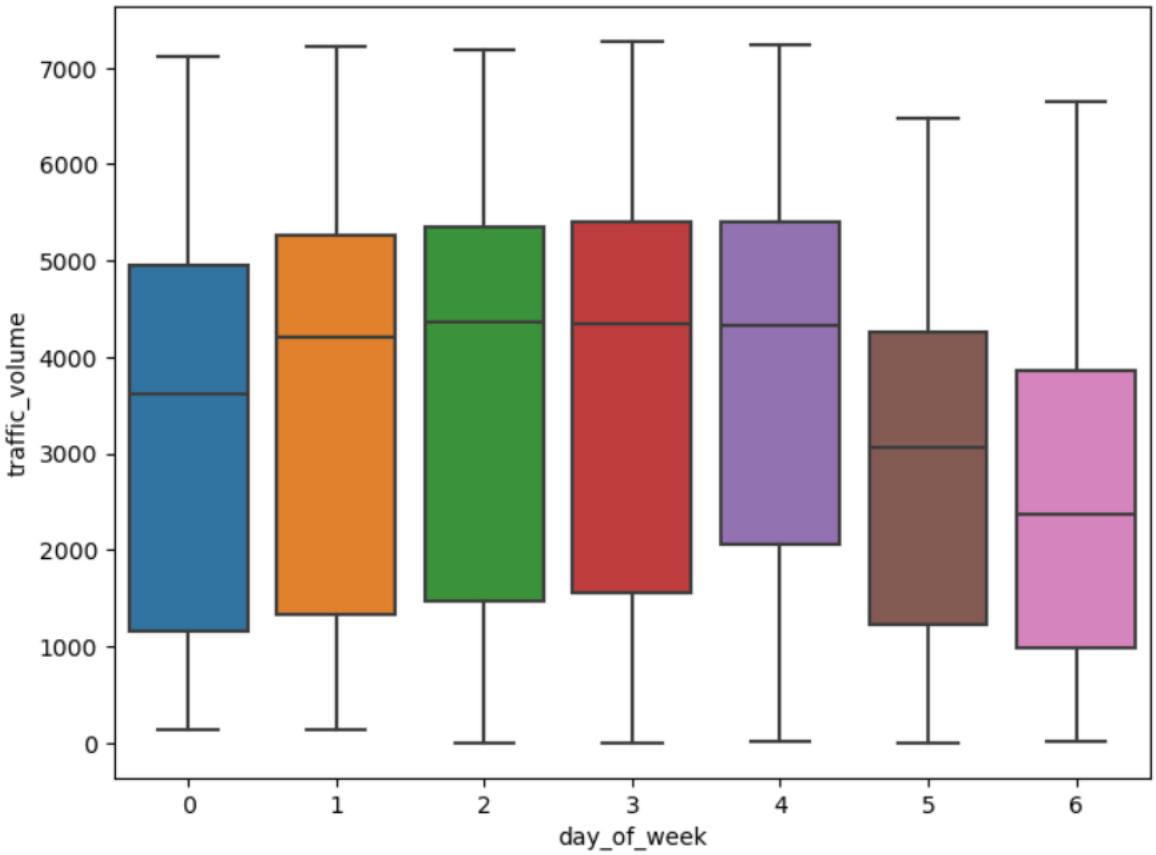
Şekil 6

Şekildeki grafik, tatil günlerinde trafik yoğunluğunu gözlemlemek için oluşturulmuştur. Kutu grafiği, tatil günlerinde trafik hacminin farklılaştığını açıkça göstermektedir. Özellikle New Year's Day ve Independence Day gibi bazı tatil günlerinde trafik hacmi oldukça yüksek ve değişkendir.



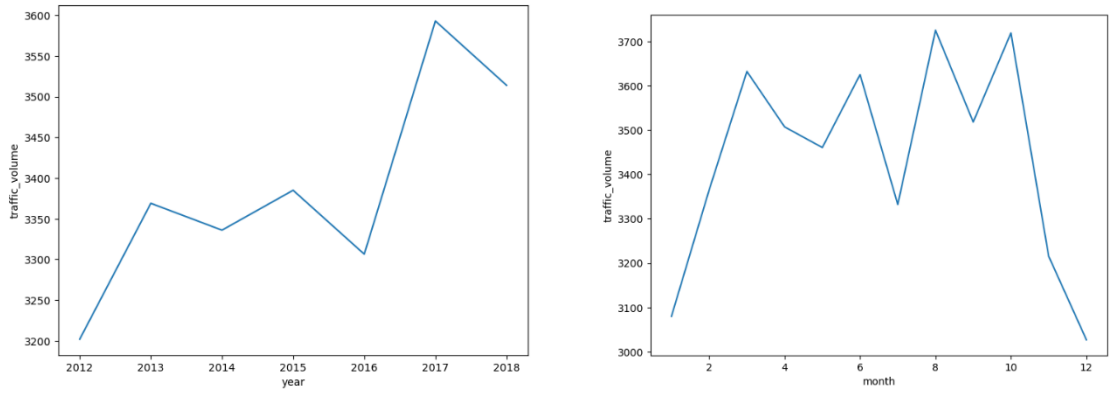
Şekil 7

Yukarıdaki 4 grafik (Şekil 7) sayısal değişkenlerin dağılımını ve adedini göstermektedir. Bu histogramlar, sıcaklık, yağmur, bulutluluk ve kar değişkenlerinin dağılımlarını göstermektedir. Sıcaklık genellikle -50 ile 50 arasında yoğunlaşmış olup, çoğu veri pozitif değerlerde yoğunlaşmıştır. Yağmur ve kar miktarları genellikle sıfırdır, bu da yağış olaylarının nadir olduğunu belirtir. Bulutluluk oranları ise %0 ve %100'de yoğunlaşarak ya tamamen açık ya da tamamen kapalı havaların sıkça yaşandığını gösterir. Bu dağılımlar, hava koşullarının genellikle ortalama ve değişken olduğunu, ancak yağış olaylarının nadir olduğunu ortaya koymaktadır.



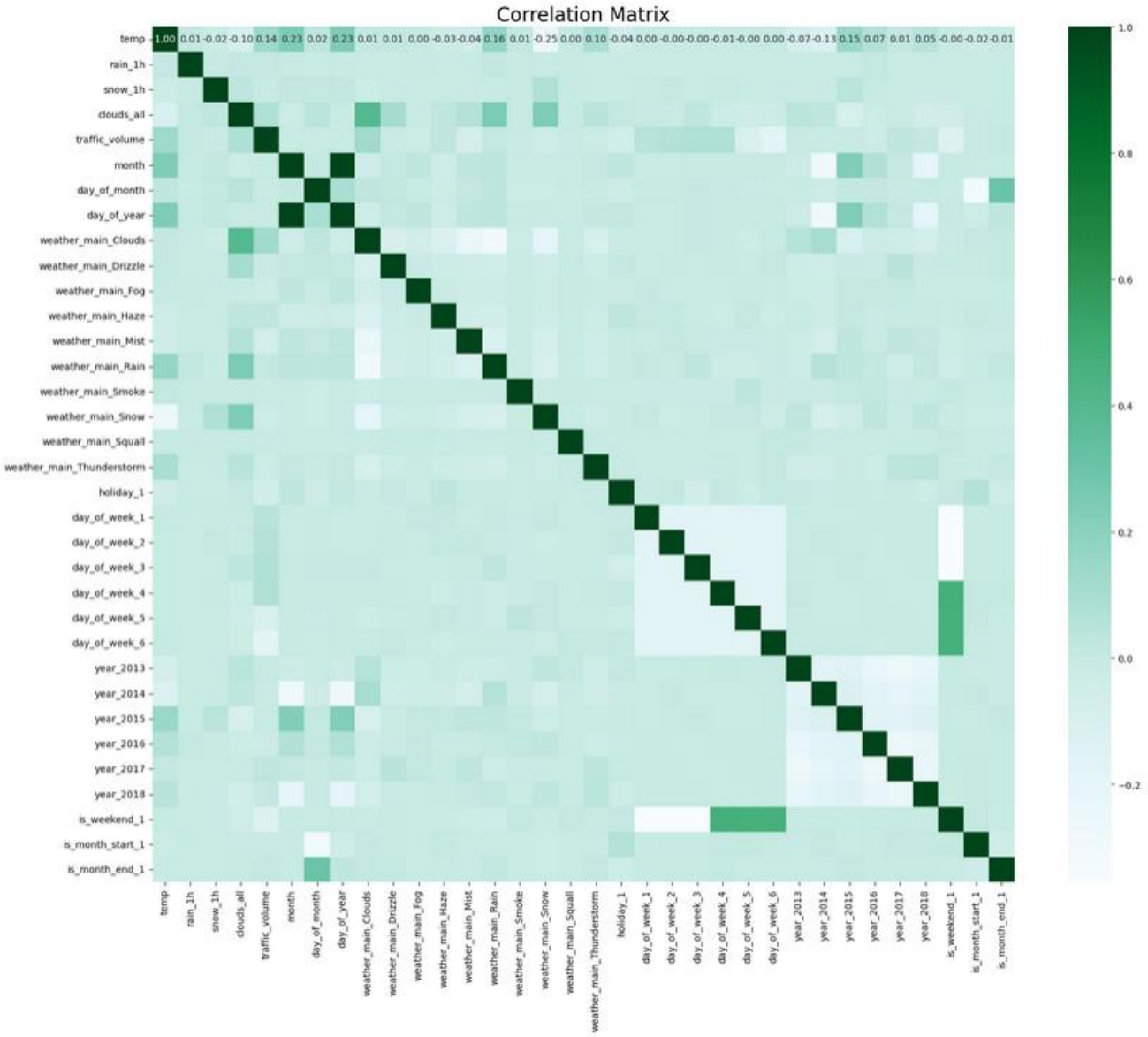
Şekil 8

Haftanın günlerine göre trafik yoğunluğuna box-plot yöntemi ile bakacak olursak gün bazında aykırı değer olmadığı gözlemlenmektedir. Hafta sonlarındaki trafik hacmi, hafta içi günlerine göre daha düşük olmuştur.



Şekil 9

Yukarıda gösterilen grafiklerden Şekil trafik hacminin o yıldaki değerlerinin medyan değerine göre gösterilmiştir. Şekil ise aylara göre trafik yoğunluğunu göstermektedir. Medyan değerinin seçilmesinin sebebi ortalamanın aykırı değerlerden fazla etkilenmesidir. Şekil verilere göre kış ayında I-94 otoyolunun kullanımı diğer aylara göre oldukça düşüktür.



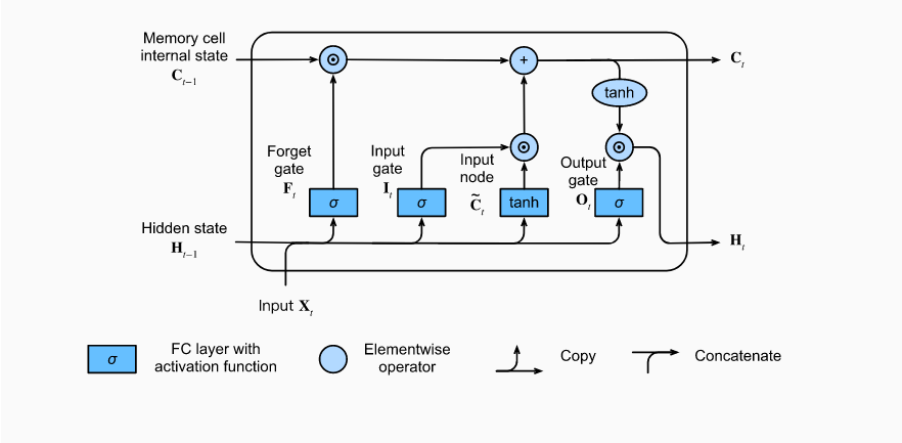
Şekil 10

Korelasyon matrisine bakıldığında ‘day_of_year’ ve ‘month’ değişkenleri arasında yüksek korelasyon olduğu Şekil görülmektedir. Aynı zamanda ‘is_weekend’ ile ‘day_of_week_5’, ‘day_of_week_6’ değişkenleri arasında da beklenildiği gibi yüksek korelasyon vardır.

Yapılan denemelerde, değişkenleri çıkarmadan modelin daha düşük MSE elde ettiği gözlemlenmiştir. Bu nedenle, değişkenleri çıkarmama kararı alınmıştır. Bu kararın verilmesinde, modelin daha fazla öğrenme yeteneğine sahip olması ve daha iyi performans göstermesi etkili olmuştur.

Veri Hazırlama ve Modelleme Süreci

Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN) mimarisinin bir türüdür ve özellikle zaman serisi verileri ve sıralı verilerin işlenmesi için kullanılır. RNN'lerin geleneksel yapısı, uzun dönemli bağımlılıkları öğrenme konusunda sınırlamalara sahip olduğu için LSTM'ler geliştirilmiştir. LSTM'ler, bu sınırlamaları aşarak daha uzun süreli bağımlılıkları ve bağlamları öğrenebilir.



Şekil 11 (LSTM Algoritmasının gösterimi)

LSTM'nin Temel Yapısı ve İşleyişi:

LSTM, hücre (cell), giriş kapısı (input gate), çıkış kapısı (output gate) ve unutma kapısı (forget gate) olmak üzere dört ana bileşenden oluşur:

1. **Hücre Durumu (Cell State):** LSTM'nin uzun süreli belleğini temsil eder. Bu durum, zaman adımları boyunca bilgiyi taşır ve LSTM'nin geçmiş bilgiyi hatırlamasını sağlar.
2. **Giriş Kapısı (Input Gate):** Yeni gelen bilginin hücre durumuna ne kadar ekleneceğini kontrol eder. Bu kapı, hangi bilgilerin güncelleneceğine karar verir.
3. **Çıkış Kapısı (Output Gate):** LSTM hücresinden çıkacak olan bilgiyi belirler. Bu kapı, hücre durumundan hangi bilgilerin çıktı olarak kullanılacağını kontrol eder.
4. **Unutma Kapısı (Forget Gate):** Hücre durumunda hangi bilgilerin unutulacağına karar verir. Bu kapı, hücre durumunu günceller ve gereksiz bilgileri siler.

Veri Ölçeklendirme

Veri setindeki değişkenler farklı ölçeklerde değerlere sahip olduğundan, verilerin ölçeklendirilmesi gereklidir. Bu süreçte, MinMaxScaler kullanılarak veriler 0 ile 1 arasında ölçeklendirilmiştir. Bu sayede, farklı ölçeklerdeki verilerin modelde düzgün bir şekilde işlenmesi sağlanmıştır.

Hedef ve Özellik Değişkenlerinin Ayrılması

Trafik hacmi (traffic_volume) hedef değişken olarak belirlenmiştir. Diğer tüm değişkenler ise özellik (feature) değişkenleri olarak kullanılmıştır. Bu ayrım, modelin sadece trafik hacmini tahmin etmeye odaklanmasını sağlamıştır.

Veri Setinin Eğitim ve Test Setlerine Bölünmesi

Veri seti, %80 eğitim ve %20 test oranında bölünmüştür. Bölme işlemi sırasında rastgelelik için random_state parametresi kullanılarak, verilerin rastgele seçimi sağlanmıştır. Bu şekilde, modelin performansının daha doğru bir şekilde değerlendirilmesi hedeflenmiştir.

Veri Şekillerinin Yeniden Düzenlenmesi

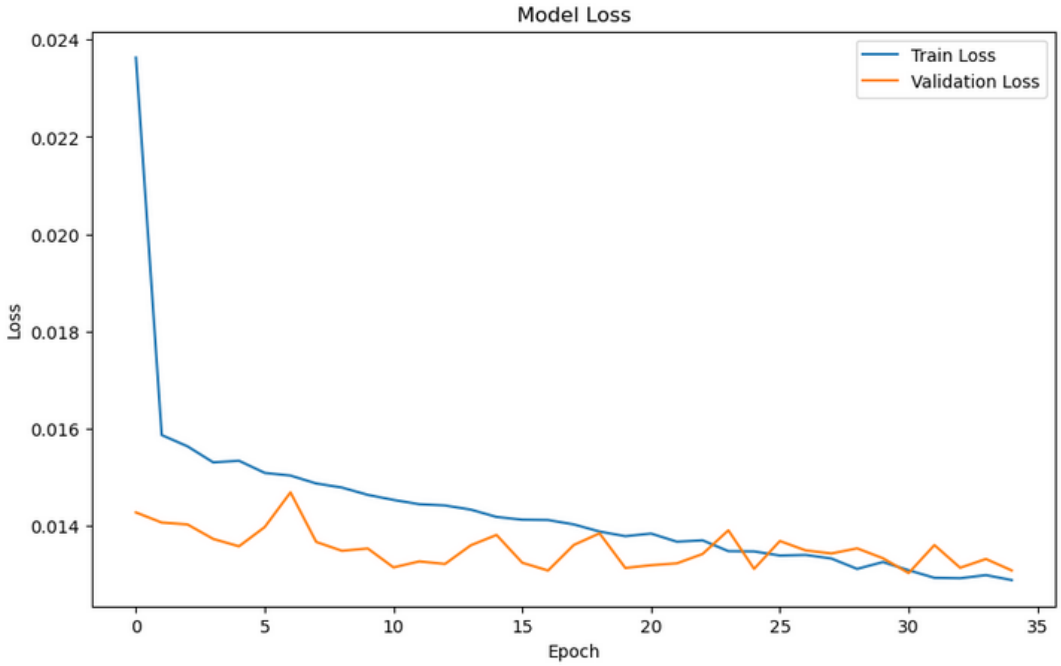
LSTM modelinin gereksinimlerine uygun olacak şekilde, veri setinin şekilleri yeniden düzenlenmiştir. Bu adım, verilerin modelin girişine uygun formatta olmasını sağlamıştır.

LSTM Modelinin Oluşturulması

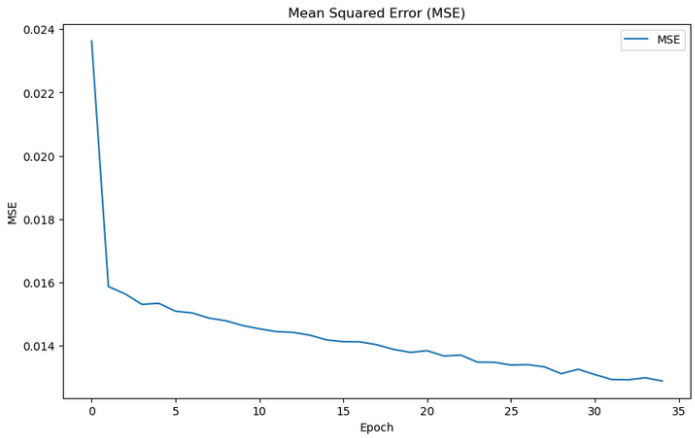
Trafik hacmini tahmin etmek için bir LSTM modeli oluşturulmuştur. Model, iki LSTM katmanı ve bir Dropout katmanı ile yapılandırılmıştır. Dropout katmanı, overfitting sorunlarını önlemek amacıyla eklenmiştir. Yoğun katman (Dense layer) ile birlikte, modelin çıktısı elde edilmiştir. Model, adam optimizer ve mean squared error (MSE) kayıp fonksiyonu ile derlenmiştir.

Modelin Eğitilmesi

Model, eğitim verileri ile 35 epoch boyunca eğitilmiştir. Eğitim süreci sırasında, modelin doğrulama performansı da gözlemlenmiştir. Bu, modelin doğrulama veri seti üzerindeki performansını izlememize olanak tanımıştır. Eğitim sürecinde, modelin overfitting yapmaması için Dropout katmanı kullanılmıştır. Modelin eğitimi sırasında doğrulama verileri üzerindeki performansı da dikkate alınarak, modelin genel başarımı değerlendirilmiştir.



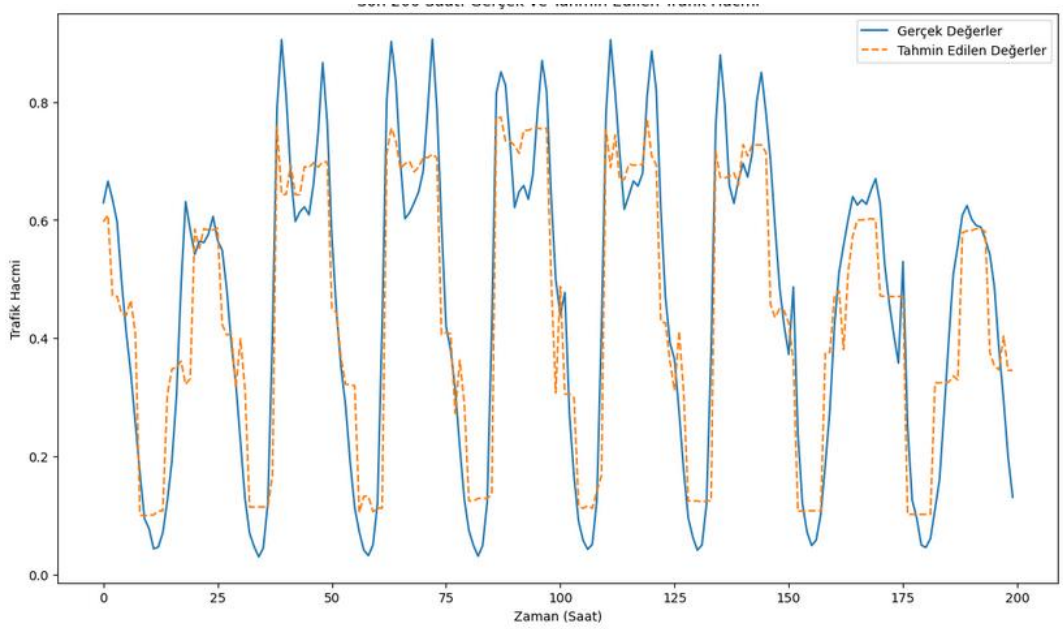
Şekil 12



Şekil 13 (Modelin mean squared error (MSE) grafiği)

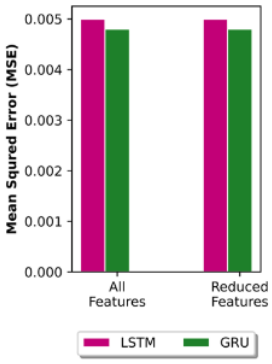
Şekil 13

Zaman Serisi Tahmini



Şekil 14 (Son 200 saatlik trafik hacminin tahmini ve gerçek değerlerle kıyaslanması)

Karşılaştırma



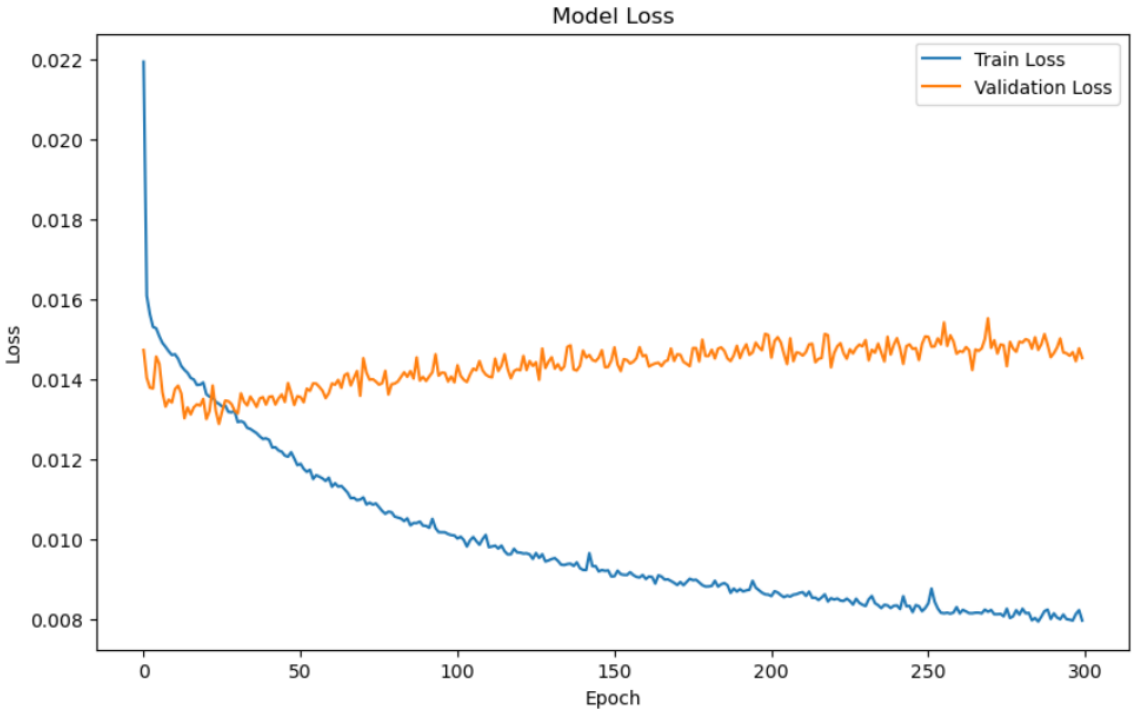
Neural Network Settings	
Input Shape	(# of Records, Time Step, # of Features)
Hidden Layers	4 Hidden Layers with 128, 64, 32, 16 units of neurons in each layer respectively
Activate Function	tanh
Batch Size	64
Learning Rate	0.0001 with decay rate 1e-5
Optimizer	Adam
Epochs	300

Şekil 16

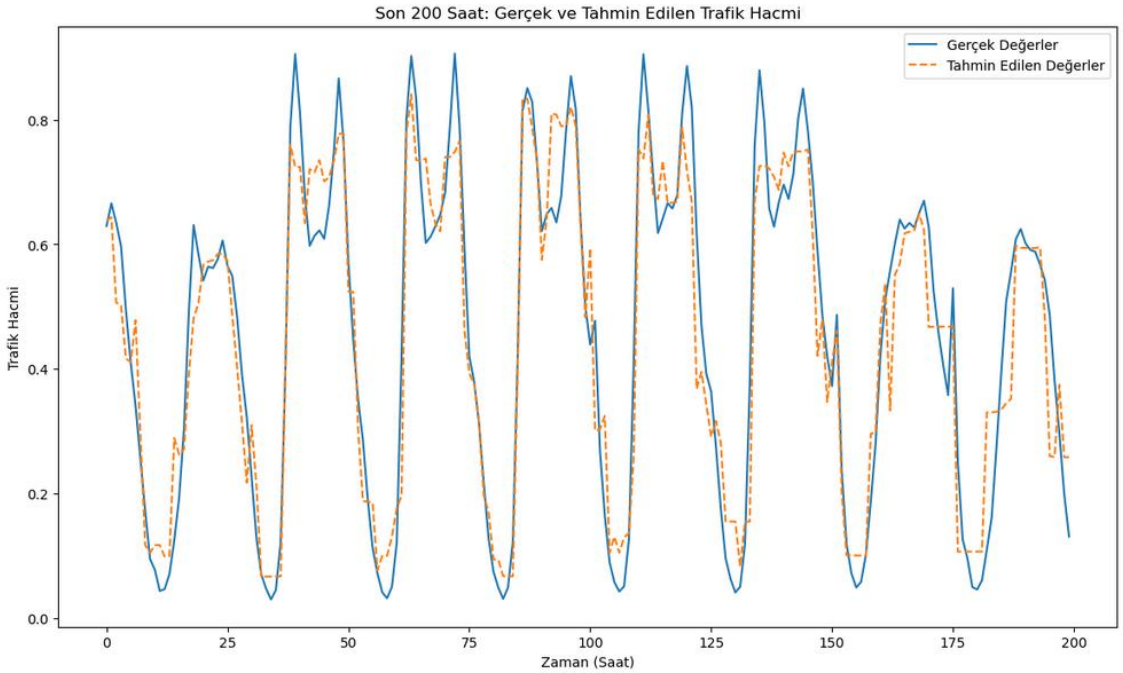
Şekil 15

Şekil 14'te, aynı veri seti kullanılarak yapılan çalışmanın sonuçları görülmektedir. Karşılaştırma yaptığım makalede LSTM modelinin parametreleri ve mimarisi Şekil 15'te gösterilmiştir. Yapılan eğitim sonucunda bu makalede MSE değeri yaklaşık olarak 0.05 bulunmuştur. Yapılan veri ön işleme, özellik çıkarımı gibi tekniklerden sonra LSTM mimarisi kullanarak yaptığım eğitimin sonucunda MSE değeri 0.0084 olarak bulunmuştur. (Şekil 16). Modelin tahmin performansı, Şekil 14'te görüldüğü üzere sadece 35 epoch boyunca

eğitilerek Şekil 18 ile çok yakın tahminlerde bulunmuştur. Bu sonuçlar da modelin aşırı öğrenmeden daha genellenebilir bir model olduğunu gösterir.



Şekil 17



Şekil 18

Kaynakça

-Gültekin, Y. B. (2021). *Traffic volume prediction using LSTM improved with rules obtained from anomalies* [M.S.- Master of Science]. Middle East Technical University. https://hdl.handle.net/11511/93167

- Das, L. C. (2023). *Traffic volume prediction using memory-based recurrent neural networks: A comparative analysis of LSTM and GRU*. arXiv. https://arxiv.org/abs/2303.12643