**T.C. İSTANBUL KÜLTÜR ÜNİVERSİTESİ**

**LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ**

# ÖZELLİK ÖNEMİNE GÖRE OTOMATİK TANIMLA SİSTEMİ VERİLERİNDEKİ EKSİK KALIPLARI YÜKLEME

**YÜKSEK LİSANS TEZİ**

**Ecem Nilay BAZMAN**

**2000006292**

**Anabilim Dalı: Bilgisayar Mühendisliği**

**Programı: Bilgisayar Mühendisliği**

**Tez Danışmanı: Doç.Dr. Fatma PATLAR AKBULUT**

**TEMMUZ 2023**

**T.C. İSTANBUL KÜLTÜR ÜNİVERSİTESİ**

**LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ**

# ÖZELLİK ÖNEMİNE GÖRE OTOMATİK TANIMLA SİSTEMİ VERİLERİNDEKİ EKSİK KALIPLARI YÜKLEME

**YÜKSEK LİSANS TEZİ**

**Ecem Nilay BAZMAN**

**2000006292**

**Anabilim Dalı: Bilgisayar Mühendisliği**

**Programı: Bilgisayar Mühendisliği**

**Tez Danışmanı: Doç.Dr. Fatma PATLAR AKBULUT**
**Jüri Üyeleri: Doç.Dr. Akhan AKBULUT**
**Prof.Dr.Özgür Koray ŞAHİNGÖZ**

**TEMMUZ 2023**

## GENEL BİLGİLER


| | | |
|---|---|---|
| Üniversite | : | T.C. İstanbul Kültür Üniversitesi |
| Enstitüsü | : | Lisansüstü Eğitim Enstitüsü |
| Dalı | : | Bilgisayar Mühendisliği |
| Programı | : | Bilgisayar Mühendisliği |
| Tez Danışmanı | : | Doç. Dr. Fatma PATLAR AKBULUT |
| Tez Türü ve Tarihi | : | Yüksek lisans – Temmuz 2023 |


## KISA ÖZET

Özellik Önemine Göre Otomatik Tanımlama Sistemi Verilerindeki Eksik Kalıpları Yükleme


**Ecem Nilay BAZMAN**



Denizcilik sektöründe, AIS (Otomatik Tanımlama Sistemi) verileri, deniz güvenliği, deniz trafiği yönetimi, liman operasyonları, deniz araştırmaları ve çevre izleme, deniz ticareti ve lojistik gibi birçok alanda önemli bir rol oynamaktadır. AIS, gemilerin konum, hız, rotasyon ve diğer ilgili bilgilerini gerçek zamanlı olarak ileten bir sistemdir. Ancak, AIS verilerinin toplandığı süreçte veya iletim sırasında eksik verilerin ortaya çıkması oldukça yaygın bir durumdur. Eksik AIS verilerinin oluşması, gemi sınıflandırması ve diğer denizcilik uygulamaları için önemli bir sorun oluşturur. Özellikle gemi sınıflandırma modelleri, gemilerin tipini doğru bir şekilde tahmin etmek için çeşitli veri özelliklerine ihtiyaç duyar. Statik verilerdeki eksiklikler, modelin doğruluğunu ve performansını olumsuz yönde etkileyebilir. Bu çalışma, gerçek AIS verileri kullanılarak gemi sınıflandırması yapan bir modelin girdileri olan statik verilerdeki eksik verilerin özellik önemine göre tamamlanması sağlayarak modele etkileri paylaşılmıştır. Eksik AIS verilerinin tamamlanması, aynı zamanda veri setindeki azınlık sınıflarının model tarafından daha iyi öğrenilmesini de hedeflemektedir. Elde edilen sonuçlar, eksik verilerin tahmine dayalı bir şekilde tamamlanması yaklaşımının modelin doğruluğunu ve performansını artırabildiğini göstermektedir. Bu çalışma, eksik verilerin tamamlanması için kullanılan yaklaşımın uygulanabilir olduğunu ve açıklayıcı bir şekilde sunulabileceğini göstermektedir.



Anahtar Kelimeler: AIS, Gemi Sınıflandırması, Eksik Verilerin Tamamlanması, Özellik Önemi.

## GENERAL INFORMATION


| | | |
|---|---|---|
| University | : | T.C. İstanbul Kültür University |
| Institute | : | Institute of Graduate Education |
| Department | : | Computer Engineering |
| Program | : | Computer Engineering |
| Supervisor | : | Assoc. Prof. Dr. Fatma PATLAR AKBULUT |
| Degree Awarded and Date | : | Master of Science – July 2023 |


## ABSTRACT

Imputing Missing Values in Automatic Identification System Data by Feature Importance


**Ecem Nilay BAZMAN**



In the maritime industry, AIS (Automatic Identification System) data plays an important role in many areas such as maritime safety, maritime traffic management, port operations, marine research and environmental monitoring, maritime trade, and logistics. AIS is a system that transmits ships' position, speed, rotation, and other relevant information in real time. However, it is quite common for missing data to appear during the collection or transmission of AIS data. The generation of missing AIS data poses a significant problem for ship classification and other marine applications. Ship classification models, in particular, require a variety of data features to accurately predict the type of ships. Missing static data can adversely affect model accuracy and performance. In this study, the effects on the model were shared by imputing the missing data in the static data, which are the inputs of a model that makes ship classification using real AIS data, according to the feature importance. Imputing the missing AIS data also aims to better learn the minority classes in the data set by the model. The results show that the predictive imputation of the missing data approach can improve the accuracy and performance of the model. This study shows that the approach used to impute the missing data is applicable and can be presented in an explanatory way.



Keywords: AIS, Ship Classification, Imputation of Missing Values, Feature Importance

# ÖNSÖZ

"Özellik Önemine Göre Otomatik Tanımlama Sistemi Verilerindeki Eksik Kalıpları Yükleme" adlı tez çalışmam süresince bilgi ve deneyimi ile destek veren değerli tez danışmanım Doç. Dr. Fatma PATLAR AKBULUT'a, sevgili eşim Ahmet BAZMAN'a ve aileme sonsuz teşekkürlerimi sunarım.

## İÇİNDEKİLER

**Sayfa No**

**ÖZET……..…….……i**

**ABSTRACT……...…….....ii**

**ÖNSÖZ…….…...……...…….....iii**

**İÇİNDEKİLER…...……...…….....iv**

**KISALTMALAR LİSTESİ……..……..……..vi**

**ŞEKİLLER LİSTESİ……..…….……vii**

**TABLOLAR LİSTESİ ……..……..……viii**

**FORMÜLLER LİSTESİ…….……..…..ix**

**1. GİRİŞ .....1**

**2. BACKGROUND .....4**

2.1. AIS Verisi Türleri ..... 4

2.2 AIS Verisindeki Kayıp Verilerin Sınıflandırılması..... 5

2.3. AIS Verileri ile Yapılan Çalışmalarda En Çok Karşılaşılan Sorunlar ..... 6

**3. METHOD.....8**

3.1. Kullanılan Veri Setine Bakış ..... 8

3.2. Veri Seti Ön İşleme ..... 11

3.2.1. Veri Temizleme..... 12

3.3. Eksik Verilerin Tamamlanması ..... 18

3.3.1 Rastgele Orman Regresyonu ..... 19

3.3.2 K En Yakın Komşu Regresyonu ..... 20

3.4. Sınıflandırma Modelinin Oluşturulması ..... 21

3.5 Model Performansı Değerlendirme Yöntemleri..... 22

**4. BULGULAR ..... 24**

4.1. AIS Verileri İçin Farklı Veri Tamamlama Tekniklerinin Karşılaştırılması ..... 24

4.2. Gemi Sınıflandırmasında Özellik Önemi Analizi ..... 25

4.3. Gemi Sınıflandırma Modeli Performans Analizi ..... 28

4.4 Veri Yükleme Sonrası Gemi Sınıflandırma Modeli Performans Analizi ..... 30

**5. TARTIŞMA ..... 33**

**6. SONUÇ ..... 34**

**7. SINIRLAMALAR VE GELECEK ÇALIŞMA ..... 36**

**8. BİLGİLENDİRME ..... 37**

**9. KAYNAKLAR ..... 38**

**10. EKLER ..... 41**

## KISALTMALAR

| | | |
|---|---|---|
| AIS | : | Otomatik Tanımlama Sistemi (Automatic Identification System) |
| ANN | : | Yapay Sinir Ağları (Artificial Neural Network) |
| CNN | : | Evrişimli Sinir Ağları (Convolutional Neural Network) |
| COG | : | Yere Göre Yön (Course Over Ground) |
| DNN | : | Derin Sinir Ağları (Deep Neural Netework) |
| HDG | : | Yön Açısı (Heading) |
| IMO | : | Uluslararası Denizcilik Örgütü (International Maritime Organization) |
| K-NNR | : | K-N Yakın Komşu Regresyonu (K-Nearest Neighbors Regressor) |
| MFELCM | : | Çoklu Ölçüt Farklılaşması Yöntemi (Multiple Factor Evaluation of Land Cover Models) |
| MMSI | : | Deniz Hareketli Hizmet Kimliği (Maritime Mobile Service Identity) |
| NOAA | : | Ulusal Okyanus ve Atmosfer Dairesi (National Oceanic and Atmospheric Administration) |
| RFR | : | Rastgele Orman Regresyonu (Random Forest Regressor) |
| SAR | : | Yapay Açıklıklı Radar (Synthetic Aperture Radar) |
| SHAP | : | Shapley Additive Explanations |
| SOG | : | Yere Göre Sürat (Speed Over Ground), |
| SOLAS | : | Denizde Can Emniyeti (Safety of Life at Sea) |
| S-AIS | : | Uydu AIS (Satellite AIS) |
| T-AIS | : | Karasal AIS (Terrestrial AIS) |
| UTC | : | Eşgüdümlü Evrensel Zaman (Universal Time Coordinate) |

## ŞEKİLLER LİSTESİ

Şekil 3-1. Tüm ABD Kıyı Bölgesi ..... 8
Şekil 3-2. Veri Setindeki Eksik Verilerin Dağılımı ..... 9
Şekil 3-3. Veri Temizleme Sonrası Verilerin Dağılımı ..... 13
Şekil 3-4. Limitli Temizleme Yöntemi-Özet İstatistikler ..... 14
Şekil 3-5. Dört Çeyrekler Yöntemi-Özet İstatistikler ..... 14
Şekil 3-6. Statik Özniteliklerin Korelasyonu ..... 19
Şekil 3-7. RFR Modeli ..... 20
Şekil 3-8. K-NNR Modeli ..... 21
Şekil 4-1. Gemi Sınıflandırma Modeli Senaryo1 Veri Seti Özellik Önemi ..... 26
Şekil 4-2. Gemi Sınıflandırma Modeli Senaryo2 Veri Seti Özellik Önemi ..... 26
Şekil 4-3. Gemi Sınıflandırma Modeli Senaryo3 Veri Seti Özellik Önemi ..... 27
Şekil 4-4. Senaryo1 Verisi ile Sınıflandırma Modeli ..... 28
Şekil 4-5. Senaryo2 Verisi ile Sınıflandırma Modeli ..... 29
Şekil 4-6. Senaryo3 Verisi ile Sınıflandırma Modeli ..... 29
Şekil 4-7. Veri Yükleme Sonrası Senaryo2 Verisi ile Model Doğruluğu ..... 30
Şekil 4-8. Veri Yükleme Sonrası Senaryo3 Verisi ile Model Doğruluğu ..... 31

## TABLOLAR LİSTESİ

Tablo 3-1. Türetilmiş Özellikler ..... 15
Tablo 3-2. Yük Türü (Cargo) ..... 16
Tablo 3-3. Gemi Türü Kodu ..... 17
Tablo 3-4. Gemi Türlerine Göre Veri Dağılımı ..... 18
Tablo 4-1. Regresyon Model Performans Metrikleri ..... 24
Tablo 4-2. Veri Yükleme Öncesi ve Sonrası Performans Metrikleri ..... 32

## FORMÜLLER LİSTESİ

Bulunmamaktadır.

# 1. GİRİŞ

Otomatik Tanımlama Sistemi (AIS), yerleşik alıcı-verici ve karasal ve/veya uydu baz istasyonları tarafından gemi hareketini izler. AIS tarafından toplanan veriler, uluslararası standarda[1] uygun olarak statik ve kinetik bilgiler içerir. AIS tarafından sağlanan veriler, deniz emniyeti ve güvenliği, gemi trafiğinin kontrolü ve çevre koruma dâhil ancak bunlarla sınırlı olmamak üzere çeşitli amaçlar için gereklidir. Ancak, AIS tarafından sağlanan veriler her zaman kapsamlı veya doğru değildir ve belirli verilerin bulunmaması, ilgili tüm bilgilere erişime dayalı gemi sınıflandırma algoritmalarının verimliliği üzerinde önemli bir etkiye sahip olabilir. Bu veri türleri, deniz istihbaratında kilit teknikler olan deniz anormallik tespiti ve gemi rotası tahmini için kullanışlıdır.

Uluslararası Denizcilik Kurumu (IMO), 2004'ten beri otomatik tanımlama sisteminin (AIS) gemide taşınmasını zorunlu kılmıştır. Bir gemide bulunan AIS, birkaç istisna dışında kapatılamaz. Karar A.917(22)[2] tarafından sağlanan IMO yönergelerine göre, gemiler yoldayken veya demirdeyken AIS her zaman çalışır durumda olmalıdır. Aksi takdirde, bir geminin yasa dışı faaliyetleri gizlemek için konumunu ve kimliğini gizlediği düşünülebilir. Kasıtlı ve kasıtlı olmayan AIS açma-kapama anahtarlama anomalisini çok sınıflı bir yapay sinir ağı (ANN) tabanlı bir anormallik tespiti ile dinamik veriler kullanılarak anlaşılmaya çalışılmıştır [1].

AIS cihazının kapalı olmamasına rağmen kötü hava koşulları, uydu tarafından kaydedilen sinyallerin kaybolmasına neden olabileceğinden AIS veri setinde eksik veriler oluşabilir. Bu eksik veriler, AIS verilerin madenciliğinin doğruluğunu etkiler [2], [3]. Bu eksik verilerin tamamlanması ile ilgili çalışmaların çoğu, eksik rota

1. International Standard IEC 61993-2, 2001.
2. Guidelines For The Onboard Opearational Use Of Shipborne Automatic Identification System (AIS) 2002.

bilgilerini kurtarmak [4], [5], eksik AIS veri analizi yoluyla gemi davranış modellerini belirlemek [6], [7], gemi trafiği emisyon tahmini [8] veya filonun enerji verimliliği [9] içindi. Eksik verilerin düşük veri kalitesi gibi sorunlara yol açabileceğine dair çalışmalar da mevcuttur [10]. AIS statik verilerin tamamlanması ile ilgili çalışmalar nadirdir. AIS statik verileri kullanılarak gemi sınıflandırılması yapan çalışmalar mevcuttur [11].

Son zamanlarda sadece SAR (Synthetic Aperture Radar) görüntü verileri ile gemi sınıflandırılması çalışmaları [12] olsa da SAR verileri genellikle AIS verileri ile birlikte tamamlayıcı bilgiler olarak kullanılmıştır [7], [13]. Burada amaç, sadece görüntü verilerinin kullanılmasında oluşabilecek yanlış pozitif ve yanlış negatiflerin azaltılması olmuştur [14].

Bazı araştırmalarda gemilerin sınıf sayısı azdır. Bu nedenle Wang ve arkadaşları [15] dört ana gemi kategorisi dedikleri yolcu, tanker, balıkçı ve kargo gemileri için sınıflandırma yaparken, Zhong ve arkadaşları 3 ana gemi kategorisi dedikleri kargo, tanker ve balıkçı gemilerini sınıflandırmaya çalışılmıştır [7]. Bu çalışmada, eksik statik verilerin tamamlanması ile veri setinin dengesiz dağılımına çözüm bulunmak istenmiştir ve tüm gemi sınıflarına ait bir sınıflandırma çözümü sunulmaya çalışılmıştır.

Yukarıda belirtilen bilgiler ışığında, bu çalışmanın amacı, kaybolan AIS verilerinin özellik önemini analiz etmek ve kayıp verilerin kurtarılmasının kurulan sınıflandırma modellerinin performansını nasıl etkilediğini araştırmaktır. Son 30 yılda yapılan eksik veri tamamlama çalışmalarını derleyen çalışmada [16] görüleceği üzere veri kurtarmanın modelin performansı üzerindeki etkilerini inceliyoruz ve eksik veriler için atamanın önemine dair bazı iç görüler sağlıyoruz.

Çalışmanın literatüre katkıları aşağıda listelenmiştir:

- ✓ T-AIS ve S-AIS arasındaki farklar ve bu farklılığın neden olduğu durumların öneminin anlaşılması için bilgiler paylaşılmıştır.

- ✓ Eksik AIS verilerini özellik öneme göre tamamlayarak, özellikle statik veriler arasında batma mesafesi olarak bilinen “Draft” özelliğinin model iyileşmesindeki etkisi paylaşılmıştır.
- ✓ AIS verisindeki aykırı değerlerin temizlenmesinde, çeyrekler veya yöntemindense bilinen minimum ve maksimum uzunluk, genişlik ve suya batma verisi değerlerine göre daha spesifik bir yöntem kullanılması.
- ✓ Gemi sınıflandırmasında kullanılan türetilmiş ve statik verilerin özellik önemi çıkarılmıştır.
- ✓ Gemi sınıflandırmasında gemi sınıfına göre veri dağılımının modele olan etkisi paylaşılmıştır.

Bu çalışmanın sonuçları, liman yetkilileri, denizcilik düzenleyicileri, denizcilik şirketleri ve diğerleri dâhil olmak üzere denizcilik işine dâhil olan çeşitli taraflar için yansımalara sahiptir. Eksik AIS verilerini doğru bir şekilde belirleme yeteneği, gemi sınıflandırma modellerinin güvenilirliğini ve etkinliğini artırabilir ve bu da deniz emniyetini, güvenliğini ve çevre korumasını artırabilir. AIS, otomatik tanımlama sistemi anlamına gelir. Ek olarak, bu çalışmadan elde edilen iç görüler, kısmi veya hatalı AIS verilerinin sunduğu sorunları ele alabilecek yeni metodolojilerin ve teknolojilerin geliştirilmesi için bir temel olarak kullanılabilir.

# 2. BACKGROUND

AIS verisi kullanılarak gemi sınıflandırması yapılırken en etkili yöntem ve özellik önemine göre girdilerin seçilmesi için birkaç yöntem ve girdi kombinasyonu denenmiştir. Kullanılan yöntemlere ve yapılmış çalışmalara ait literatür taraması önemli anahtar kelimeler üzerinden araştırılmıştır.

## 2.1. AIS Verisi Türleri

AIS, temel olarak iki tür veri kaynağına dayalı olarak sınıflandırılır: Karasal AIS (T-AIS) ve Uydu AIS (S-AIS). T-AIS, yerleşik alıcı-verici ve kara baz istasyonları tarafından toplanan AIS verilerine dayanırken, S-AIS uydu bazlı AIS verilerine dayanır. Bununla birlikte, farklı AIS veri kaynakları ve yöntemleri de kullanılabilmektedir. Örneğin, bazı ülkelerde AIS verileri, gemi trafik yönetimi ve deniz güvenliği amaçları için özel olarak toplanmaktadır. Bu nedenle, bazı durumlarda farklı AIS türleri veya kaynakları kullanılabilmektedir, ancak genel olarak T-AIS ve S-AIS en yaygın olanlarıdır.

T-AIS verisi gemilerin kıyıya yakın bölgelerdeki hareketlerini izlemek için kullanıldığı için, bazı çalışmalarda "Kıyı Tabanlı AIS Verisi" olarak adlandırılırken [5], [17], [18] S-AIS bazı çalışmalarda "Uzay Tabanlı AIS Verisi" olarak adlandırılmıştır [6], [7], [15]. Uzay tabanlı AIS verisi, kapsama alanı daha geniş olduğu için daha kapsamlı bir veri seti sağlar. Ancak, bu verilerin bazıları gürültülü olabilir ve veri doğruluğu konusunda sorunlar yaşanabilir. Diğer yandan, kıyı tabanlı AIS verisi, daha doğru ve güvenilir bir veri sağlar, ancak kapsama alanı daha sınırlıdır.

Bu iki tür AIS verisinin elde edilme yöntemlerinin farklı olmasından kaynaklı benzer statik ve dinamik verileri içerseler dahi veri işleme yöntemleri farklılaştığı görülmüştür. Örneğin, başka bir gemi sınıflandırılması çalışmasında [15] kullanılan

verinin doğruluğundan da emin olmak için statik ve dinamik veriler, 4 farklı seriye (statik özellik, dinamik özellik dağıtım, zaman serisi ve zaman serisi özellik örneklerine) bölünerek MFELCM bir yöntem kullanılarak gemi sınıflandırılması yapılmıştır. Ancak kıyı tabanlı AIS verisi kullanan çalışmalarında ise AIS verisinin sadece statik verileri ve bu statik verilerden türetilmiş yeni özellikleri kullanılarak gemi sınıflandırılması yapılabilmiştir [7]. Eksik statik verileri tamamlamak yerine sıfır ile doldurmuştur.

### 2.2 AIS Verisindeki Kayıp Verilerin Sınıflandırılması

Gemi karakteristiğini ortaya koyan gemi uzunluğu ve genişliği verilerinin eksikliği, rastgele eksik (MAR) verilere bir örnektir. Rastgele kayıp verilerde, bir veri noktasının eksik olma olasılığı gözlenen değerlere bağlıdır, ancak kayıp değerlere bağlı değildir. AIS verileri söz konusu olduğunda, eksik gemi boyu verileri, AIS sistemindeki teknik zorluklardan veya sınırlamalardan kaynaklanabilir veya bilgiler ilk etapta kaydedilmediği için eksik olabilir.

Kayıp veriler üç farklı türe ayrılabilir: tamamen rastgele eksik (MCAR), rastgele eksik (MAR) ve kayıp nedenine göre rastgele eksik (MNAR) [1] [19]. MCAR verileri, gözlemlenen veya eksik değerlerle hiçbir ilişkisi olmaksızın tamamen şans eseri eksiktir. MNAR verileri hem gözlemlenen hem de eksik değerlere bağlı olarak bir veri noktasının eksik olma olasılığı ile sistematik bir şekilde eksiktir.

Kayıp verilerin diğer veriler üzerindeki potansiyel etkisi göz önüne alındığında, verilerin neden eksik olduğunun altında yatan nedenleri dikkate almak önemlidir [19]. Eksik veriler ile ilgili yapılan çalışmaları, Scopus veri tabanı kullanılarak 60 yıllık (1960-2019) çalışmaları incelendiğinde, kayıp verilerde ciddi bir araştırma artışının 2016'da gerçekleştiğini göstermiştir [20]. Bu çalışma aynı zamanda kayıp veriler araştırmasında en çok ele alınan konu alanlarından birinin tıp ve sağlıkla ilgili olduğunu ortaya koymuştur. Yine benzer bir çalışma ile (1991-2021) 30 yıllık çalışmalar incelendiğinde ise bugüne kadar eksik veri atama araştırmalarında yükselen bir trend olduğu ve en çok tercih edilen iki değerlendirme yöntemi (rastgele orman ve en yakın komşu yöntemleri) olduğu anlaşılmıştır [16]. Bu çalışmada ise batma

mesafesi (draft) eksik verileri Rastgele Orman Regresyonu (RFR) ve K En Yakın Komşu Regresyon (K-NNR) modelleri kıyaslanarak en iyi sonucu verdiği görülen K-NNR ile tamamlanmıştır (bakınız 3.3. Eksik Verilerin Tamamlanması).

### 2.3. AIS Verileri ile Yapılan Çalışmalarda En Çok Karşılaşılan Sorunlar

AIS verileri ile yapılan çalışmalarda en çok karşılaşılan sorunlar, kuşkusuz ki verinin temizlenmesi, işlenmesi ve veri kullanım amacına göre seçilecek modellerde etkili olmuştur. Bu konuyu beş maddede listeleyebiliriz:

1. Veri eksikliği: Gemilerin AIS verilerinin sürekli olarak toplanması mümkün değildir ve bu nedenle veri eksikliği sorunu oluşabilir. Özellikle açık denizlerde ve az trafikli bölgelerde, gemilerin AIS sinyallerinin alınmaması veya eksik alınması mümkündür.
2. Veri doğruluğu: AIS verileri, gemilerin kendi AIS cihazları tarafından yayınlanır ve bazen bu cihazlarda teknik sorunlar meydana gelebilir. Bu durumda, veriler yanlış veya eksik olabilir. Ayrıca, bazı gemi sahipleri, gemilerinin gerçek konumunu gizlemek veya yanıltmak için AIS cihazlarını kapatabilir veya sahte veriler yayınlayabilir.
3. Veri boyutu: AIS verileri, yüksek hızda ve sürekli olarak toplandığı için çok büyük boyutlara ulaşabilir. Bu nedenle, verilerin saklanması, işlenmesi ve analiz edilmesi zor olabilir.
4. Veri bütünlüğü: AIS verilerinin doğruluğu ve eksiksizliği, verilerin doğru bir şekilde toplanması, aktarılması ve depolanması ile doğrudan ilişkilidir. Ancak, bu süreçlerde oluşan hatalar, veri bütünlüğünü olumsuz etkileyebilir.
5. Veri güvenliği: AIS verileri, gemilerin seyrüsefer bilgilerini içerdiği için, güvenliği tehdit edebilir. Bu nedenle, AIS verilerinin güvenliği, siber saldırılara karşı korunması gereken bir konudur.

Bu çalışmada kullanılan veri setini, veri doğruluğu, eksikliği ve boyutu ile ele alacak olursak;

1. Veri eksikliği ile başa çıkmak için çoğu çalışmada veri setinden eksik verilerin atılması veya sıfır ile doldurulduğu görülürken bu çalışmada eksik veriler tamamlanmıştır.

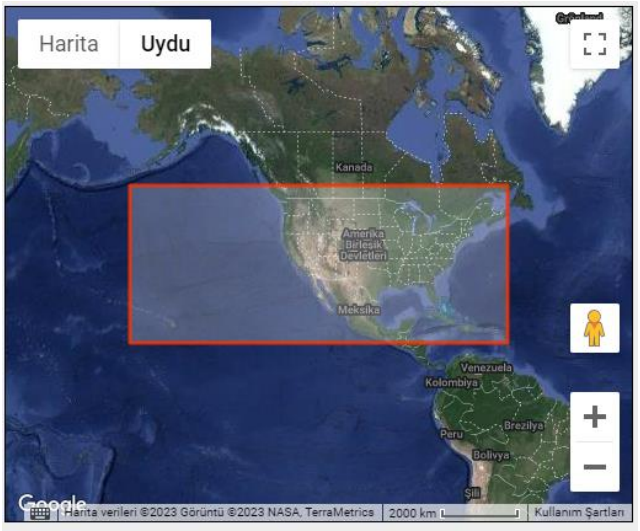
2. Veri setinin doğruluğu, uzay tabanlı yerine kıyı tabanlı AIS verisi kullanıldığından daha doğruluk payı yüksek bir veri seti elimizde olduğunu varsayarak insani hataları en aza indirgemek için de limitli veri temizleme işlemi yapılarak veri setinin doğruluğu artırılmaya çalışmıştır (bakınız 3.2.1. Veri Temizleme).
3. Veri boyutu için tercih ettiğimiz veri setinde 7801106 verinin MMSI ve IMO numaraları temel alınarak verinin tekrarının engellenmesi ile 4745 veri kalmıştır. Böylece modelde sızıntının engellenmesi sağlanmıştır.

# 3. METHOD

Bu çalışmanın amacı, AIS verisindeki eksik verilerin gemi sınıflandırılmasına etkisini araştırmaktır. Bu kapsamda, eksik verilerin tamamlandığı ve tamamlanmadığı iki veri seti, bir gemi sınıflandırması modeline girdi olarak verilmesi sonucu modelin doğruluğu karşılaştırılmıştır. Gemi sınıflandırılması için kurulan basit katmanlı derin öğrenme modelinin optimizasyonu yapıldıktan sonra iki veri ile denenerek sonuçları paylaşılmıştır.

## 3.1. Kullanılan Veri Setine Bakış

Bu çalışmada ise kullanılan veri kıyı tabanlı AIS verisi olup, NOAA tarafından sağlanan Şekil 2.1'de görüldüğü üzere tüm ABD Sahil Güvenlik karasal alıcılarından alınan kayıtlarının bir dakikalık örnekleme hızına göre filtrelenmiş ve konum verilerindeki eksikliklerin gemi rotalarına göre tamamlanmış bir veri setidir.

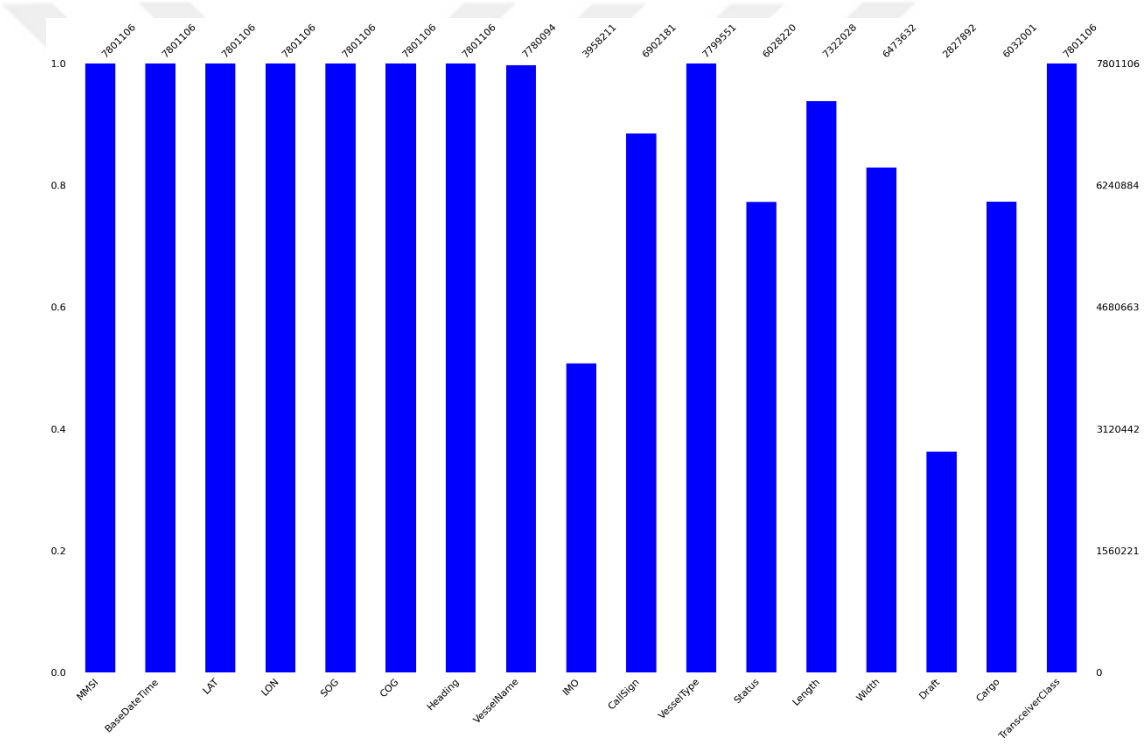



**Şekil 3-1.** Tüm ABD Kıyı Bölgesi

AIS veri setinde, statik ve dinamik veriler bulunmaktadır. Statik veriler; gemi adı, MMSI numarası, geminin IMO numarası, çağrı işareti, geminin uzunluğu,

genişliği ve batma mesafesi (Draft), gemi tipi, rota planı, AIS tipidir. Kinetik veriler ise; gemi pozisyonu, UTC saati ve rota bilgilerinden (yere göre yön (COG), yere göre sürat (SOG), yön açısı (HDG), dönüş hızı (ROT), meyil açıları) oluşur. Bu araştırmada, gemi sınıflandırmasında büyük rol oynayan statik verilerde bulunan eksik bilgilerin önem özellikleri yardımıyla tamamlanmasını içerecektir.

Kullanılan veri setinde, konum verilerindeki eksikliklerin gemi rotalarına göre tamamlanmış bir veri seti oluğundan Şekil 2.2'de görüleceği üzere dinamik verilerde eksiklik bulunmamaktadır. Belirli bir alandan elde edilmiş AIS verisindeki dinamik (kinetik) verilerin gemi sınıflandırılmasında girdi olarak dâhil etmeye ihtiyaç duyulmamıştır.




**Şekil 3-2.** Veri Setindeki Eksik Verilerin Dağılımı

Statik verilerden MMSI numarası, AIS ve diğer bazı ekipmanlar tarafından bir gemiyi veya sahil radyo istasyonunu benzersiz bir şekilde tanımlamak için kullanılan dokuz haneli bir numaradır. Bir diğer statik özellik olan IMO, üretici şirketin her gemiye verdiği, önünde IMO harfleri bulunan yedi haneli benzersiz gemi numarasıdır. Bu numara, hurdaya çıkarılana kadar aynı kalır ve geminin sahibi, kayıtlı olduğu ülke veya adı ne olursa olsun asla değişmez. MMSI ve IMO numaralarının benzersiz olması nedeniyle gemi sınıflandırılmasında kullanılmaya elverişli özellikler olarak görülmemiştir. Ancak veri sızıntısı engellemek amacıyla veri filtrelemek için IMO ve

MMSI numaraları kullanılmıştır. Ayrıca IMO numarasının AIS verisinde eksik olması, bir geminin IMO numarası bilinçli olarak gizli tutması veya yayınlamamasından kaynaklanıyor ise bu anomali çalışmalarında kullanılabilir.

Statik veriler arasında en çok kayıp verinin olduğu batma mesafesi (draft) verisi, bir geminin su hattının su seviyesine olan derinliğini temsil eder ve geminin yüzdüğü suyun ne kadar derin olduğunu belirtir. Bazı kaynaklarda batma mesafesi olarak da adlandırılabilir. Ancak, batma mesafesi bilgisi bazı durumlarda dinamik olarak güncellenebilir. Özellikle geminin yük değiştirdiği, su seviyesinin değiştiği veya geminin trim (denizcilikte teknenin kıç ve baş taraflar arasındaki farka verilen isim) veya stabilizesi gibi faktörlerin etkisiyle batma mesafesi veya aynı anlama gelen "su çekimi" değeri değişebilir. Bu durumda batma mesafesi bilgisi, geminin anlık durumunu yansıtan dinamik bir veri olarak kabul edilebilir. Sonuç olarak, batma mesafesi bilgisi genellikle geminin statik bilgileri arasında yer alsa da bazı durumlarda geminin anlık durumuna bağlı olarak dinamik olarak güncellenebilir. Bu sayede geminin yük alıp almadığına dair bir bilgiye ulaşmamızı sağlayan bu veri gemi sınıflandırmasında önemli bir etkiye sahipken yapılan çoğu çalışmada model girdi olarak kullanılmadığı görülmüştür. Bunun nedeni, AIS verilerinin genelinde batma mesafesi (draft) verisinde bulunan eksikliğin fazla olması sebebiyle göz ardı edilmesi daha uygun bulunmuş olabilir.

Statik veriler arasında ikinci en çok kayıp verinin Yük Türü (Cargo) olduğu gözlemlenmektedir. Bir geminin taşıdığı yük türü gemi sınıflandırmasına katkı sağlayamazken yük türü belirli bir geminin taşıyıcı ve büyük bir gemi olma olasılığının yüksek olması, gemi sınıflandırılmasında kullanılabilir.

Eksik veri bulunmayan bir özellik olan AIS Alıcı Sınıfı (AIS TransceiverClass) gemi sınıflandırılmasında kullanılabilecek bir özelliktir. Alıcı Sınıfı A, geminin donanımının daha gelişmiş bir AIS alıcı (transceiver) olduğunu ve veri yayını, alımı ve işlemesi için daha kapsamlı yeteneklere sahip olduğunu gösterir. AIS Alıcı Sınıfı A cihazları, daha yüksek güç çıkışına sahip olabilir, daha geniş bir veri yayın kapasitesine sahip olabilir ve daha karmaşık işlevlere sahip olabilir. Bu tür cihazlar, büyük ölçekli ticari gemilerde ve deniz trafiği yoğun bölgelerde kullanılmaktadır. AIS Alıcı Sınıfı B

ise daha düşük güç çıkışına ve daha sınırlı veri yayın kapasitesine sahip olan AIS cihazlarını temsil eder. AIS Alıcı Sınıfı B cihazları genellikle küçük ölçekli ticari gemiler, özel tekne sahipleri ve yatlar gibi daha küçük deniz araçlarında kullanılır. Bu tür cihazlar, AIS Alıcı Sınıfı A cihazlarına kıyasla daha az karmaşık işlevlere sahip olabilir ve daha düşük bir maliyetle sağlanabilir.

AIS statik verileri arasında yer alan uzunluk ve genişlik bilgileri, gemi tipinin tahmininde önemli bir gösterge olarak kullanılır. Bu bilgiler, gemi tipleri arasında belirgin boyut farklılıklarının olduğunu gösterir ve gemi tipini doğru bir şekilde tahmin etmek için değerli bir referans sağlar. Örneğin, daha küçük bir uzunluk ve genişliğe sahip bir gemi muhtemelen bir yat veya özel tekne olabilirken, daha büyük bir uzunluk ve genişlik, bir kargo gemisi veya yolcu gemisi olabileceğini düşündürebilir. Bu nedenle uzunluk ve genişlik bilgilerinden yeni özellikler türetilerek gemi hakkında daha fazla bilgi edinilmesi de amaçlanabilmektedir.

### 3.2. Veri Seti Ön İşleme

AIS statik verilerinden uzunluk (Length), genişlik (Width), yük türü (Cargo), suya batma mesafesi (Draft) ve AIS cihaz türü (TransceiverClass) gemi sınıflandırma modeli girdisi olarak değerlendirilmektedir. Gemi sınıflandırması çalışmalarında, geminin karakteristiğini ortaya koyan verilerin statik verilerden genişlik ve uzunluk verileri olduğu görülmüştür [7], [11]. Çoğu çalışmada batma mesafesi verisi veri setinden çıkartılmıştır. AIS verisi kullanılarak yapılan gemi sınıflandırılmasında uzunluk ve genişlik gibi özellikler tek başına yeterli olmayabilir. Bu nedenle, gemi boyutu hakkında daha ayrıntılı bir bilgi sağlamak için bu özelliklerden ek özellikler türetilir. Türetilen uzunluk ile genişliğin çarpılması (f4) gibi türetilmiş özellikler, geminin alanını temsil eder ve gemi sınıflandırılmasında kullanışlı bir özellik olarak kabul edilerek çalışmalarda kullanıldığı görülmüştür [7], [13].

Bu bilgiler ışığında bu çalışmada, statik verilerden AIS cihaz türü, yük türü, uzunluk, genişlik ve türetilmiş verilerin girdi olarak değerlendireceği *senaryo1* veri seti; AIS cihaz türü, yük türü, uzunluk, genişlik ve suya batma mesafesi verilerinin girdi olarak değerlendireceği *senaryo2* veri seti; AIS cihaz türü, yük türü, suya batma

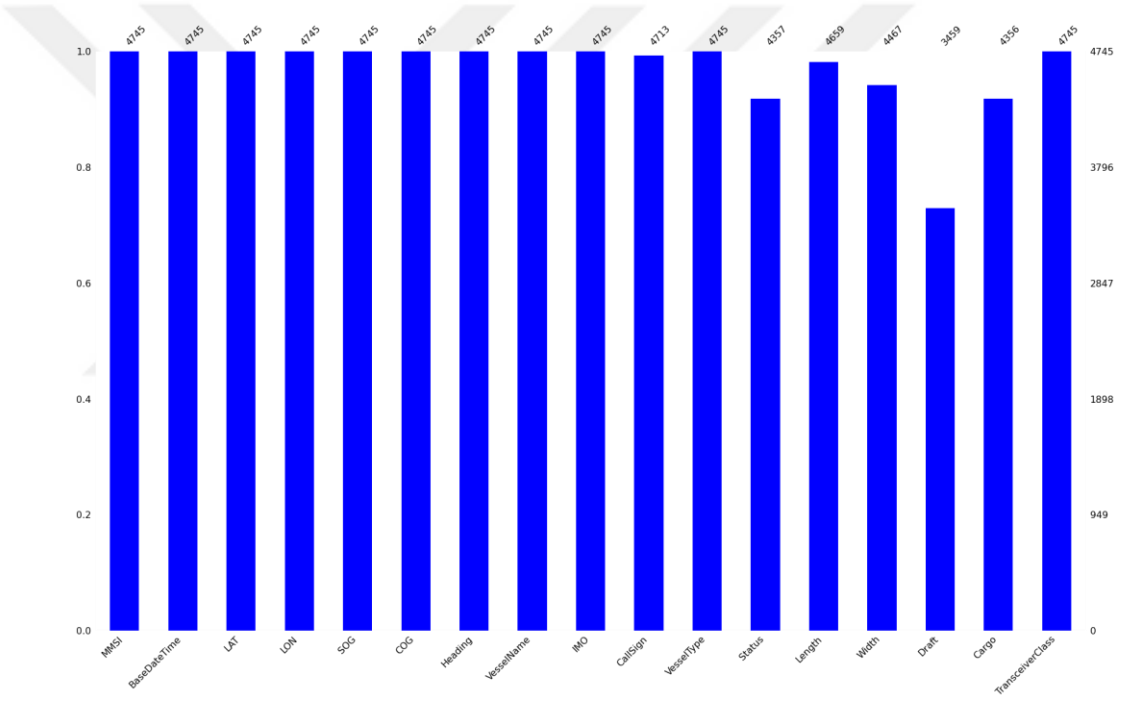
mesafesi, uzunluk, genişlik ve türetilmiş verilerin girdi olarak değerlendireceği *senaryo3* veri setleri oluşturulmuştur. Bu senaryo veri setlerinin oluşturulması, statik verilerin özellik önemlerinin ortaya çıkaracaktır. Bütün veri setlerinde aynı yaklaşımlar kullanılarak aykırı veriler temizlenmiştir.

Normalleştirme, özellikler arasındaki farklılıkları azaltarak modelin daha iyi genelleştirme yeteneğine sahip olmasını sağlar. Önce statik verileri normalize edip sonra öznitelikleri türetmek, normalize edilmemiş uzunluk ve genişlik değerleriyle türetilmiş özniteliklerin hesaplanmasını içerir. Bu durumda, türetilmiş öznitelikler normalize edilmemiş verilere dayanır ve bu verilerin aralıklarına bağlı olarak geniş farklılıklar gösterebilir. Örneğin, uzunluk ve genişlik değerleri 1 ila 100 arasında değişiyorsa ve bu değerler normalize edilmeden önce kullanılarak öznitelikler türetilirse, türetilmiş öznitelikler büyük değer aralıklarına sahip olabilir. Bu durumda, modelin türetilmiş öznitelikleri ağırlıklarını ölçeklendirme konusunda zorluk yaşayabilir. Öte yandan, önce uzunluk ve genişliği normalize edip sonra öznitelikleri türetmek, normalize edilmiş uzunluk ve genişlik değerlerine dayalı olarak türetilmiş özniteliklerin hesaplanmasını içerir. Bu yöntem, türetilmiş özniteliklerin değer aralığını daha tutarlı hale getirir ve modelin bu özniteliklere daha dengeli bir şekilde tepki vermesini sağlayacağı için statik ve türetilmiş veriler, bütün denemelerde normalleştirilerek modele girdi olarak verilmiştir.

### 3.2.1. Veri Temizleme

Bir gemiye ait benzersiz verilerin tespit edilmesi için aynı gemiye ait verilerin tespitinde MMSI ve IMO numaraları kullanılmaktadır. MMSI numarası, gemi ve diğer deniz araçlarının tanımlanmasında kullanılan benzersiz bir kimlik numarasıdır. Her bir gemi veya deniz aracı için farklı bir MMSI numarası bulunur. Ancak, farklı zaman dilimlerinde aynı MMSI numarasına sahip farklı deniz araçları da olabilir. Örneğin, bir gemi satıldığında veya yeniden adlandırıldığında MMSI numarası değişebilir. Bu nedenle, MMSI numarasına dayalı filtreleme yaparken dikkatli olmanız ve verileri dikkatlice incelemesi önemlidir. Öte yandan, IMO numarası, geminin uluslararası denizcilik standartlarına uygun olarak tescil edildiği ve kaydedildiği zaman verilir. Bu numara, geminin kimlik bilgileri, teknik özellikleri ve sahiplik durumu gibi önemli

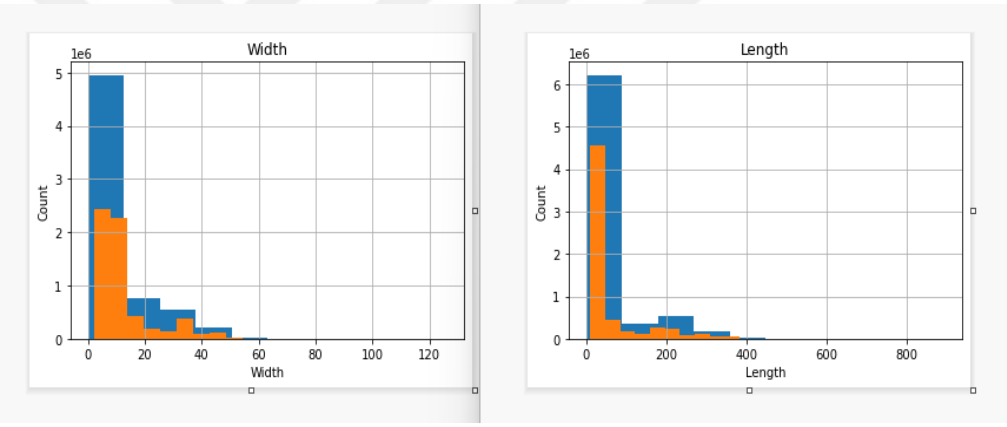
bilgileri içerir. Bu nedenle, bir gemi satıldığında veya yeniden adlandırıldığında bile IMO numarası değişmez ve geminin benzersiz kimliği korunur. Sadece IMO numarasına göre veri filtrelendiğinde aslında benzersiz verilere teoride ulaşılabilir. Ancak insani hatalara açık olan AIS hem IMO hem de MMSI numarasına göre filtreleme, daha kesin bir filtreleme sağlayarak daha spesifik verilere erişmenizi sağlar. Bu nedenle aynı MMSI ve IMO numarasına ait verilerden sadece bir tanesi kullanılarak verinin doğruluğu artırılırken veri sızıntısı da engellenmeye çalışılmıştır. IMO numaralarının oldukça eksik olması nedeniyle ve aynı gemi tarafından tekrarlayan verilerin atılmasıyla Şekil 3-3'deki veri seti dağılımı elde edilmiştir. Bu şekilden yapılan filtrelemenin veri boyutuna olan etkisi anlaşılmaktadır.



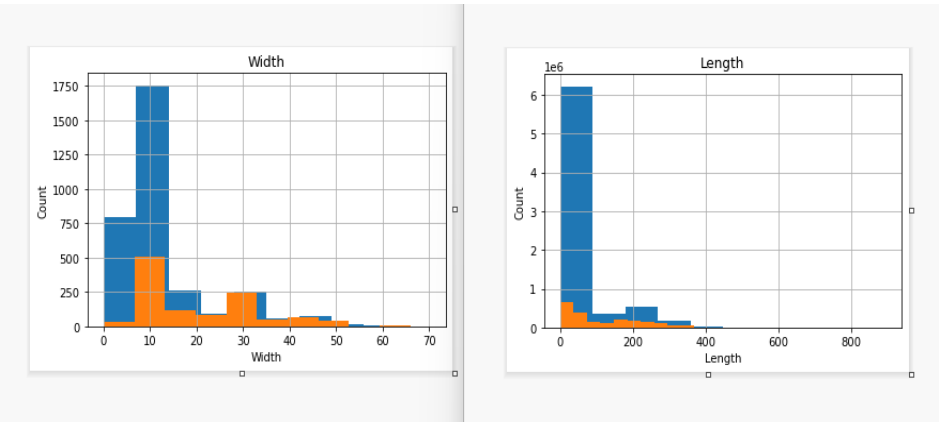

**Şekil 3-3.** Veri Temizleme Sonrası Verilerin Dağılımı

Veri temizleme yöntemleri, veri setlerindeki gürültüyü azaltmak ve hatalı, eksik veya aykırı değerleri düzeltmek için kullanılan tekniklerdir. Eksik verilerle başa çıkmanın önemini ele alacağımız bu çalışmada, başlangıç modelin girdileri oluşturulurken eksik veriler ve aykırı değerler veri setinden atılmıştır. Böylece eksik verileri hiçe sayan bir veri seti ile eksik verilerin tamamlandığı bir veri seti arasındaki farklar daha iyi vurgulanabilecektir.

AIS statik verilerinden uzunluk ve genişlik için veri temizleme ihtiyacı, bu verilerin doğruluğuna duyulan şüphe nedeniyle oluşmaktadır. Bu verileri temizleme yöntemi olarak daha önce yapılan çalışmalarda, veri setindeki aykırı değerleri tespit etmek için Dört Çeyrekler Yöntemi (Quartile Method) kullanıldığı görülmüştür. Ancak AIS statik verilerinden geminin uzunluğu ve genişliğinin değerleri, aslında minimum ve maksimum değerleri olabileceğinden belirli aralıklar limit kabul edilerek temizleme yapılmıştır. Gemilerin uzunluğu ile genişliği arasında genellikle 10:1 veya 5:1 oran bulunmaktadır [21]. Ayrıca hurdaya ayrılmamış en uzun geminin 400 metre olduğu bilinmektedir [22]. Bu bilgiler ışığında, geminin uzunluğu minimum 10 maksimum 400 metre iken genişliği minimum 2 maksimum 60 metre olacak şekilde aykırı verilerin temizlenmesi Şekil 3-4'deki gibi sağlanmıştır. Bunun yerine Dört Çeyrekler Yöntemi kullanıldığında verilerin dağılımı Şekil 3-5'teki gibi olmaktadır.




**Şekil 3-4.** Limitli Temizleme Yöntemi-Özet İstatistikler




**Şekil 3-5.** Dört Çeyrekler Yöntemi-Özet İstatistikler

Özet istatistiklerin görsellerin kanıtladığı gibi, aykırı verilerin Dört Çeyrekler Yöntemi ile temizlenmesi, belirli bir özellik için verilerin dağılımını değiştirirken, belirli limit aralıklarına göre aykırı değerlerin temizlenmesi verilerin dağılımı daha az etkilemiştir. Yapılan temizleme sonrası %1.49 veri uzunluk ve genişliği aykırı değere sahip olduğu tespit edildiğinden veri setinden çıkarılmıştır.

Gemi sınıflandırılmasında önemli rol oynayan AIS verisindeki statik verileri ve Tablo *3-1*'deki gibi statik verilerden yeni özellikler türetilmiştir. Türetilmiş her yeni özellik gemi sınıflandırma modeline normalleştirilmiş birer girdi olarak verilmiştir. Türetilmiş özelliklerin özellik öneminin bulunması, yapılacak diğer çalışmalara katkı sağlayacaktır.

**Tablo 3-1.** Türetilmiş Özellikler

| Fonksiyon | Uzunluk (L) ve Genişlik (W) Cinsinden Değeri |
|---|---|
| **f1** | L |
| **f2** | W |
| **f3** | 2*(L+W) |
| **f4** | L*W |
| **f5** | L/W |
| **f6** | W/L |
| **f7** | $(L+W)^2 / (L*W)$ |
| **f8** | $W^2 / (L^2 + W^2)$ |
| **f9** | (L-W) / (L+W) |
| **f10** | L/ (L+W) |
| **f11** | W/ (L+W) |

AIS statik verilerinden AIS alıcı türü (TransceiverClass) verisi, AIS cihazlarının alıcı (transceiver) özelliklerini ve yeteneklerini belirtmek için kullanılan standart sınıflandırmalardır, karakter olarak A ve B değerlerini almaktadır. Bu kategorik veriyi sayısal uyumluluk sağlamak amacıyla A yerine "0" ve B yerine "1" atadıktan sonra tek başına kodlama (one-hot encoding) yöntemi kullanılarak modele girdi olarak verilmiştir.

**Tablo 3-2.** Yük Türü (Cargo)

| Kod | Yük Türü |
|---|---|
| **10** | Kuru yük (genel) |
| **11** | Buğday |
| **12** | Mısır |
| **13** | Diğer tahıllar |
| **14** | Soya fasulyesi |
| **15** | Diğer yağlı tohumlar |
| **16** | Un |
| **17** | Şeker |
| **18** | Makarna |
| **19** | Diğer işlenmiş gıdalar |
| **20** | Kömür |
| **21** | Petrol ürünleri |
| **22** | Gaz yağı |
| **23** | LPG |
| **24** | LNG |
| **25** | Diğer sıvılaştırılmış gazlar |
| **26** | Kimyasallar |
| **27** | Gübre |
| **28** | Havaî yakıt |
| **29** | Diğer petrol ürünleri |
| **30** | İnşaat malzemeleri |
| **31** | Demir cevheri |
| **32** | Kömür külü |
| **33** | Agrega |
| **34** | Demir ve çelik |
| **35** | Diğer metaller |
| **36** | Makine ve araçlar |
| **37** | Elektronik ekipman |
| **38** | Diğer imalat malları |
| **39** | Diğer genel kargo |
| **40** | Tehlikeli yük |
| **41** | Patlayıcı madde |
| **42** | Gaz |
| **43** | Yanıcı sıvı |
| **44** | Yanıcı katı |
| **45** | Oksitleyici madde |
| **46** | Zehirli ve enfeksiyon yapıcı madde |
| **47** | Diğer tehlikeli yükler |

AIS statik verilerinden sayısal bir veri olan yük türü (Cargo), Tablo ***3-2***'de görüldüğü üzere taşınan yük türünü ifade etmektedir. Yük türü, geminin bir taşıyıcı gemi olup olmadığı bilgisini içerdiği için gemi sınıflandırmasına katkıda bulunması hedeflenmiştir. Bu nedenle tabloya göre yük türü verisi, minimum ve maksimum değerleri arasında kalan "1" yani taşıyıcı gemi; dışında kalan değerler için "0" atanarak taşıyıcı olmayan gemi bilgisine çevrilmiştir. Ayrıca bu özellik, gemi sınıflandırma modeline tek başına kodlama yöntemi kullanılarak girdi olarak verilmiştir.

Gemi sınıflandırılmasında çıktı olarak elde edilecek gemi türü sayısal değerleri Tablo 3-3'teki gemi türü koduna göre önce sayısal değerlere dönüştürülmüş ardından tek başına kodlama yöntemi uygulanmıştır.

**Tablo 3-3.** Gemi Türü Kodu

| Aralık | Gemi Tipi | Veri Setinde Kullanılan Gemi Türü |
|---|---|---|
| **80-89** | Tanker | 1 |
| **70-79** | Kargo Gemisi | 2 |
| **60-69** | Yolcu Gemisi | 3 |
| **52** | Çekme Gemisi | 4 |
| **51** | Kurtarma Gemisi | 11 |
| **50** | Pilot Gemisi | 5 |
| **40-49** | Yüksek Hızlı Tekne | 6 |
| **36-37** | Keyif Teknesi | 7 |
| **35** | Askeri Gemi | 8 |
| **31-32** | Çekme Gemisi | 4 |
| **30** | Balıkçı Gemisi | 9 |
| **20-29** | Su Üstü Kanadı | 10 |
| **21- 22** | Çekme Gemisi | 4 |
| **Aralık Dışı Bütün Değerler** | Diğer | 0 |

Veri temizleme sonrası veriye genel bir bakış yapıldığında, her bir gemi tipinden kaç adet veri olduğuna ilişkin dağılımın dengesiz olduğu Tablo 3-4'te görülmektedir.

**Tablo 3-4.** Gemi Türlerine Göre Veri Dağılımı

| Gemi Tipi | Veri Setinde Kullanılan Gemi Türü | Senaryo1 Gemi Türü Veri Sayısı | Senaryo2 ve Senaryo3 Gemi Türü Veri Sayısı | Yükleme Sonrası Gemi Türü Veri Sayısı |
|---|---|---|---|---|
| Tanker | 1 | 513 | 508 | 513 |
| Kargo Gemisi | 2 | 1117 | 1108 | 1117 |
| Yolcu Gemisi | 3 | 377 | 263 | 377 |
| Çekme Gemisi | 4 | 1057 | 690 | 1057 |
| Kurtarma Gemisi | 11 | 0 | 0 | 0 |
| Pilot Gemisi | 5 | 2 | 2 | 2 |
| Yüksek Hızlı Tekne | 6 | 6 | 6 | 6 |
| Keyif Teknesi | 7 | 214 | 203 | 214 |
| Askeri Gemi | 8 | 4 | 4 | 4 |
| Balıkçı Gemisi | 9 | 133 | 48 | 133 |
| Su Üstü Kanadı | 10 | 7 | 7 | 7 |
| Diğer | 0 | 595 | 526 | 595 |
| **TOPLAM:** | | 4025 | 3365 | 4025 |

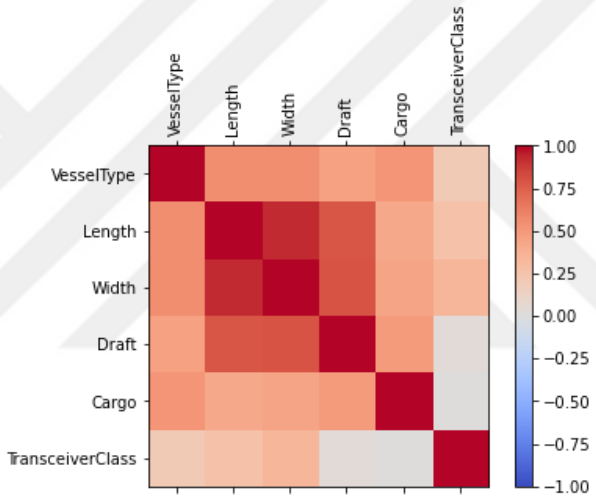
Başlangıç veri seti, 4745 veriden oluşmaktadır. Veri temizlenmesi ardından, çoğu gemi sınıflandırma çalışmalarında yapıldığı gibi eksik olan verilerin hepsi çıkarıldığında senayo1 veri setinde 4025 veri kalırken, senaryo2 ve senaryo3 veri setinde 3365 veri kalmaktadır. Veri setleri arasında bu farkın oluşmasının en önemli nedeni, suya batma verisinin büyük çoğunluğunun eksik olmasıdır. Suya batma verisindeki eksik verilerin veri setinden atılması, özellikle senaryo2 ve senaryo3 veri setlerinde başlangıç veri setine kıyasla, verinin yaklaşık %27'sinin kaybına neden olmaktadır. Bu durum, suya batma verisindeki eksik verilerin tamamlanmasının gemi sınıflandırmasını etkileyebileceğini göstermektedir.

### 3.3. Eksik Verilerin Tamamlanması

Gemi sınıflandırma modeline girdi olarak verilen statik verilerden oluşturulan senaryolu veri setleri ile yapılan denemelerde suya batma verisindeki eksik verilerin tamamlanmasının modele katkı sağlayacağı öngörülmüştür. Öte yandan suya batma verisinin yüzde %27'si eksik olduğundan, veri setinde en çok eksikliğe sahip öznitelik olarak dikkat çekmektedir.

Suya batma özelliğin tamamlanmasının  Şekil 3-6'de görüldüğü üzere diğer veriler ile ilişkili olması nedeniyle suya batma verisi dışındaki diğer statik verilerin ile eğitilen model ile suya batma verilerinde eksiklikler tamamlanmıştır. Bu nedenle yükleme sonrası veri setinin sayısı ve gemi türü dağılımı, senaryo1 veri setindeki ile aynı olduğu Tablo 3-4'te görülmektedir.

Suya batma verisindeki eksikliklerin tamamlanması sonrası suya batma verisinin de gemi sınıflandırmaya girdi olarak verilebilmesine imkân verecek veri seti dağılımında Tablo 3-4'te görüldüğü üzere, balıkçı gemilerinde yaklaşık 2,7 kat artış, çekme gemilerinde %53 artış ve yolcu gemilerinde ise %43 veri artış meydana gelmiştir.



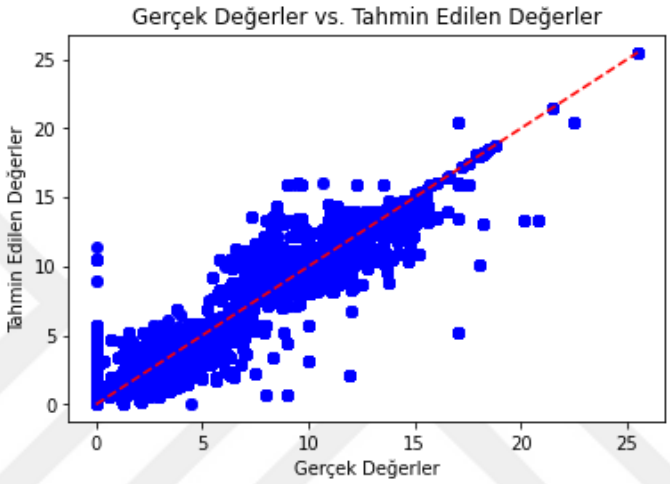

**Şekil 3-6.** Statik Özniteliklerin Korelasyonu

Son 30 yılda yapılan eksik veri tamamlama çalışmalarını derleyen çalışmada [12] görüleceği üzere, eksik veri tamamlanması için en çok tercih edilen iki yöntem Rastgele Orman Regresyonu (RFR) ve K En Yakın Komşu Regresyonu (K-NNR) yöntemleri olduğu anlaşılmıştır. Bu nedenle bu çalışmada iki yöntem de denenmiştir. Bu denemelere ait sonuç ve karşılaştırmalar 4.1. AIS Verileri İçin Farklı Veri Tamamlama Tekniklerinin Karşılaştırılması bölümünde paylaşılmıştır.

### 3.3.1 Rastgele Orman Regresyonu

Modelin parametreleri şu şekildedir: maksimum derinlik (max_depth) değeri 30, bir yaprak düğümünde bulunması gereken minimum örnek sayısı (min_samples_leaf)

5, bir düğümün dallanması için gereken minimum örnek sayısı (min_samples_split) değeri 10 ve oluşturulacak ağaç sayısın (n_estimators) 30. Bu parametreler, modelin karmaşıklık düzeyini, veri bölünmesini ve ağaç sayısını kontrol etmek için belirlendi. Bu şekilde belirlenen parametreler, modelin en iyi performansı elde etmek için optimize edilmiştir. Her bir parametre, modelin karmaşıklık düzeyini, veri bölünmesini ve ağaç sayısını kontrol ederek veri setindeki ilişkileri doğru bir şekilde öğrenmesini sağlar. Böylece, daha doğru sınıflandırma sonuçları elde edilmiştir.



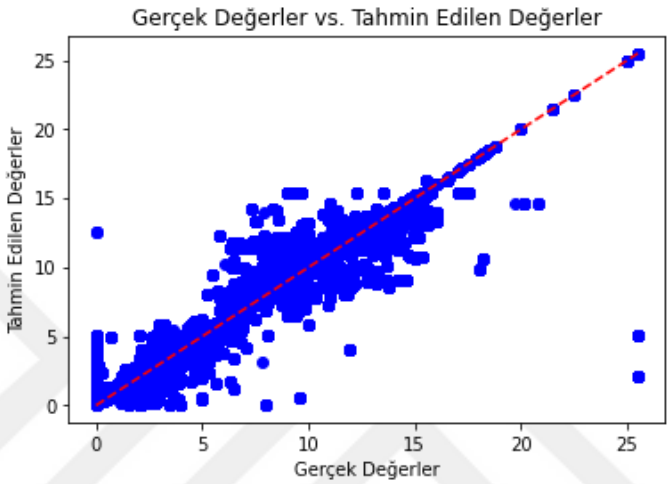

**Şekil 3-7.** RFR Modeli

Gerçek değerler ile tahmin edilen değerler arasındaki ilişkiyi gösteren Şekil 3-7'de görüldüğü üzere tahmin edilen mavi noktaların genellikle gerçek değerleri ifade eden kırmızı doğru etrafında yoğunlaştığı görülmektedir. Ancak bu değerlerin bazıları kırmızı çizgiden oldukça uzaktadır, yani hatalı tahmin edilmişlerdir.

### 3.3.2 K En Yakın Komşu Regresyonu

Model, tahminler yaparken komşu sayısını dikkate alır. Bu durumda, komşu sayısı 10 olarak belirlenmiştir. Yani, her bir tahminde 10 komşu veri noktası göz önünde bulundurulacaktır. Ayrıca, komşu noktaların ağırlıkları uzaklıklarına bağlı olarak hesaplanır. 'distance' seçeneğiyle belirlenen ağırlıklandırma yöntemi, daha yakın komşulara daha büyük ağırlıklar verir. Model, veriye hızlı erişim sağlamak için top ağacı (ball tree) algoritmasını kullanır. Bu algoritma, veri noktalarını kümelere ayırarak verimli komşu aramaları gerçekleştirir. Yaprak düğüm boyutu, 20 olarak

belirlenmiştir. Minkowski uzaklık metriği kullanılırken p değeri 1 olarak belirlenmiştir. Bu durumda, Manhattan uzaklık metriği kullanılır. Bu şekilde yapılandırılan model, veri setine dayalı olarak komşu tabanlı bir regresyon gerçekleştirir. Tahminler yapmak için, komşu sayısı ve veri noktaları arasındaki uzaklıkların ağırlıklı bir ortalaması hesaplanır. Bu model, regresyon problemlerinde kullanılan etkili bir yaklaşımdır.




**Şekil 3-8.** K-NNR Modeli

Gerçek değerler ile tahmin edilen değerler arasındaki ilişkiyi gösteren Şekil 3-8'de görüldüğü üzere tahmin edilen mavi noktaların genellikle gerçek değerleri ifade eden kırmızı doğru etrafında yoğunlaştığı görülmektedir. Bu değerlerin çoğunun RFR modeline göre kırmızı çizgiye daha yakın olduğu görülmüştür.

### 3.4. Sınıflandırma Modelinin Oluşturulması

Gemi sınıflandırma çalışmalarında [7] rastgele orman yöntemi sıkça kullanılan bir yöntemdir. Rastgele orman yöntemi, birçok karar ağacının bir araya getirilmesiyle oluşan bir topluluk öğrenmesi yöntemidir. Bu yöntem, bir veri kümesi üzerinde çoklu ağaçların eğitilmesi ile her bir ağacın sonuçlarından elde edilen sonuçların ortalamasının alınmasıyla bir sonuç elde eder. Ancak, bu yöntem, sınıflandırma yapmak için kullanılan değişken sayısı arttıkça daha karmaşık hale gelir ve hesaplama maliyeti artabilir.

Öte yandan, Sequential sınıfı, derin öğrenme algoritmalarını uygulamak için kullanılan bir modüldür. Bu sınıf, yapay sinir ağları oluşturmak için kullanılır. AIS verisi için yapay sinir ağı oluşturulması, veri setinin boyutu büyüdükçe daha iyi sonuçlar vereceği için bu çalışmada kullanılmıştır. Ayrıca bu yöntemde ağın doğru bir şekilde eğitilmesi ve aşırı uydurma (overfitting) sorununun önlenmesi için daha fazla özen gösterilmesi gerekmektedir. Bu sorunların giderilmesi, çalışmanın amacına uygun bulunduğundan bu çalışmada, Sequential sınıfı kullanılarak yoğun katmanlı derin öğrenme modeli oluşturulmuştur.

Modelin amacı, veri setindeki özelliklere dayanarak giriş verilerini doğru şekilde sınıflandırmaktır. Model, ReLU ve sigmoid gibi aktivasyon fonksiyonlarını kullanarak non-lineer ilişkileri yakalamaya çalışır. Son olarak, softmax fonksiyonu, çıktıları sınıf olasılıklarına dönüştürerek sınıflandırma yapılmasını sağlar.

### 3.5 Model Performansı Değerlendirme Yöntemleri

Doğruluk, sınıflandırma problemlerinde sıklıkla kullanılan bir metriktir ve modelin doğru sınıflandırma oranını yüzde olarak gösterir. Ancak, doğruluk değeri tek başına bazı durumlarda yeterli bir performans değerlendirme metriği olmayabilir. Özellikle dengesiz sınıf dağılımlarına sahip veri setlerinde, doğruluk değeri yanıltıcı olabilir. Bu durumlarda, hassasiyet (precision), geri çağırma (recall), F1 skoru gibi diğer metrikler de değerlendirilmelidir.

- Hassasiyet (Precision): Hassasiyet, doğru pozitif tahminlerin tüm pozitif tahminlere oranını temsil eder. Yani, bir sınıfı tahmin edildiğinde ne kadarının gerçekte o sınıfa ait olduğunu gösterir. Hassasiyet yüksek olduğunda, yanlış pozitiflerin sayısı düşüktür ve modelin pozitif tahminleri doğru bir şekilde yapma yeteneği artar.
- Geri Çağırma (Recall): Geri çağırma, gerçek pozitif tahminlerin tüm gerçek pozitiflere oranını temsil eder. Yani, bir sınıfın ne kadarının doğru bir şekilde tahmin edildiğini gösterir. Geri çağırma yüksek olduğunda, yanlış negatiflerin sayısı düşüktür ve modelin gerçek pozitifleri kaçırmama yeteneği artar.
- F1 Skoru: F1 skoru, hassasiyet ve geri çağırma değerlerinin harmonik ortalamasını temsil eder. F1 skoru hem hassasiyeti hem de geri çağırma oranını

dikkate alarak bir modelin performansını ölçer. Yani, hem doğru pozitif tahminlerin yüksek olmasını hem de gerçek pozitiflerin yüksek olmasını isteyen bir metriktir. F1 skoru yüksek olduğunda, modelin hem hassasiyeti hem de geri çağırma oranı iyi bir şekilde dengelenmiştir.

Bu metrikler, sınıflandırma modellerinin performansını değerlendirmede kullanılan önemli ölçümlerdir. Hassasiyet, geri çağırma ve F1 skoru değerlerinin yüksek olması, modelin iyi bir performans gösterdiğini ve doğru tahminlerde bulunduğunu gösterir.

Bu çalışmadaki veri setinde gemi türlerinin dengesiz dağılıma sahip olması nedeniyle model performans değerlendirmeleri bu metrikler üzerinden 4.4 Veri Yükleme Sonrası Gemi Sınıflandırma Modeli Performans Analizi başlığı altında paylaşılmıştır.

# 4. BULGULAR

## 4.1. AIS Verileri İçin Farklı Veri Tamamlama Tekniklerinin Karşılaştırılması

K-NNR ve RFR modelleri, eksik verilerin tamamlanması için uygulanmıştır. Karşılaştırmayı yaparken, hata metriklerinin düşük olması ve R-kare ($R^2$) değeri ile Açıklanan Varyans Skoru yüksek olması aranan özelliklerdir.

**Tablo 4-1.** Regresyon Model Performans Metrikleri

| Model | MAE | MSE | RMSE | $R^2$ Skoru | EVS |
|---|---|---|---|---|---|
| K-NNR | 0.3732 | 0.8514 | 0.9227 | 0.9535 | 0.9535 |
| RFR | 0.9883 | 2.3167 | 1.5221 | 0.8779 | 0.8779 |

Tablo 4-1'teki performans metrikleri karşılaştırıldığında, K-NNR modelinin daha iyi performans gösterdiği söylenebilir. K-NNR modelinin Ortalama Mutlak Hata, Ortalama Kare Hata ve Kök Ortalama Kare Hata değerleri RFR modelinden daha düşüktür. Ayrıca, K-NNR modelinin R-kare ($R^2$) değeri ve Açıklanan Varyans Skoru daha yüksektir. Bu da K-NNR modelinin bağımlı değişkeni daha iyi açıkladığını gösterir.

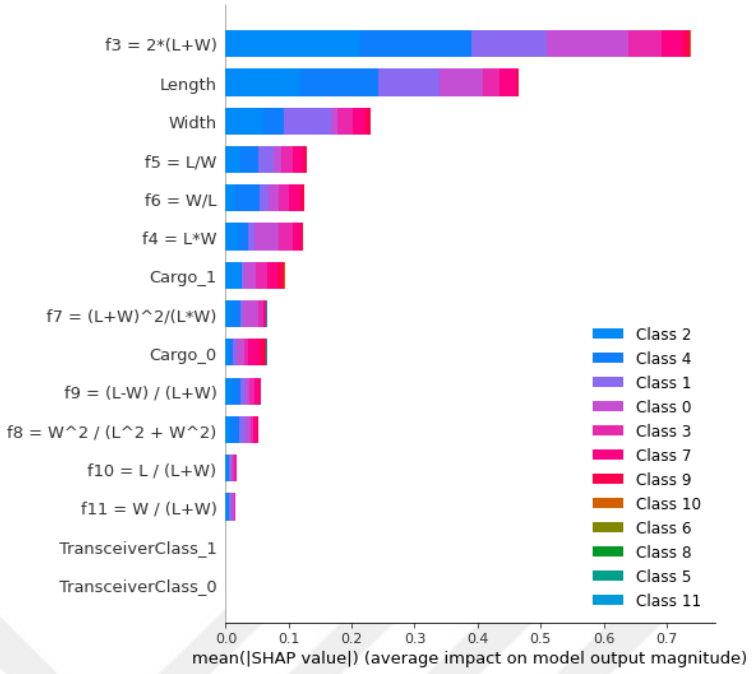
Sonuç olarak, K-NNR modeli daha iyi bir performans sergilediği için suya batma verisini tamamlamak için K-NNR modeli kullanılmıştır. Suya batma verilerinin eksiklik oranı yüzde %27'den %15'e düşürülmüştür. Veri setindeki bütün suya batma verisindeki eksikliklerin tamamlanamama sebebi, suya batma verisini tamamlayabilmek için diğer statik verilerde de eksiklik olmaması gerekliliğidir.

### 4.2. Gemi Sınıflandırmasında Özellik Önemi Analizi

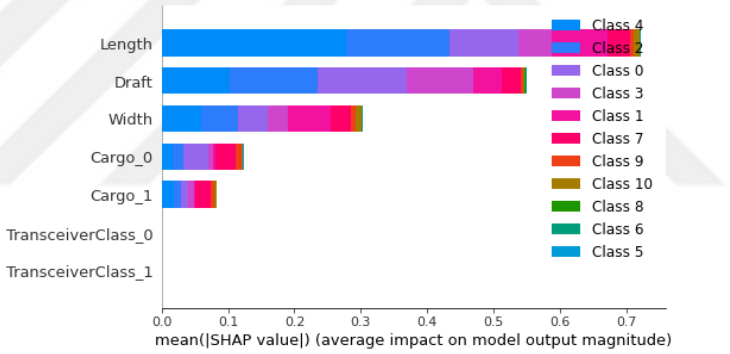
Tüm senaryolar göz önünde bulundurulduğunda Şekil 4-1, Şekil 4-2 ve Şekil 4-3'de görüldüğü üzere, AIS cihaz tipini yansıtan "AIS Alıcı Sınıfı" verisinin aslında gemi sınıflandırmaya bir katkısı olmadığı da görülmüştür. Öte yandan, geminin bir taşıma işlemi yapıp yapmadığını yorumlayabildiğimiz "Kargo" verisinin gemi sınıflandırmasına katkıda bulunması oldukça mantıklıdır. Özellikle bir taşıma işlemi belirttiği durumdaki katkısı daha fazla olduğu görülmüştür.

Bilinen ve bu zamana kadar yapılmış çalışmalarda da ispatlandığı üzere geminin uzunluğu ve genişliği modeli etkileyen önemli öznitelikler olsa da kendilerinden türetilmiş veriler aslında gemiye dair daha derin bilgileri sağlamış olduğu görülmüştür. Ancak bu zamana kadar çok fazla eksik olmasından kaynaklı kullanılması tercih edilmeyen suya batma verisinin de senaryo2 ve senaryo3'te modeli en fazla etkileyen öznitelik olduğu görülmüştür.

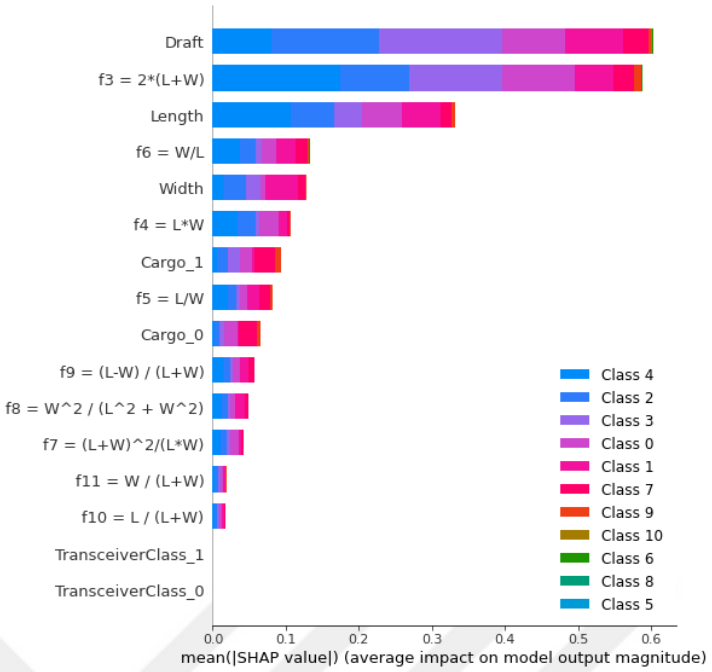
Gemi sınıflandırmasındaki özellik önemlerin çıkarılması için SHAP (Shapley Additive Explanations) değerleri hesaplanıp görselleştirilmiştir. Değişken "explainer" ile bir KernelExplainer nesnesi oluşturarak kullanılan eğitim veri setinden farklı alt kümeler ile SHAP değerleri hesaplanmıştır. SHAP değerlerini hesaplarken kullanılan eğitim ve test veri setlerinin boyutu önemlidir. Gemi sınıflandırma modellerinde, veri boyutu 4025 olan veri setinin %20'si test %80'i eğitim verisi olarak kullanılmıştır. SHAP değerleri hesaplanırken veri setinden seçilen örneklemler sırasıyla ilk 100, 200, 500 ve 1000 olarak denenmiştir. Bütün denemelerde özellik önem sıralamasının değişmediği görülmüştür. 1000 örnekli deneme, bütün senaryolardaki özellik önemlerin kıyaslanması için kullanılmıştır.




**Şekil 4-1.** Gemi Sınıflandırma Modeli Senaryo1 Veri Seti Özellik Önemi




**Şekil 4-2.** Gemi Sınıflandırma Modeli Senaryo2 Veri Seti Özellik Önemi




**Şekil 4-3.** Gemi Sınıflandırma Modeli Senaryo3 Veri Seti Özellik Önemi

Türetilmiş verilerden özellik önemi en fazla olan ise f3 formülüdür. Bu formül, gemi boyutunu basit bir şekilde hesaplamak için yaygın olarak kullanılan bir yöntemdir. Gemi boyutu, gemilerin tasarım ve performans özelliklerinin belirlenmesinde önemli bir faktördür. Bu nedenle, AIS verilerinde gemi boyutu gibi türetilmiş veriler, gemilerin fiziksel özelliklerini daha iyi tanımlamak ve analiz etmek için kullanılabilir.

Türetilmiş verilerden özellik önemi yüksek çıkan bir diğer özelliğin f4 olduğu görülmüştür. Bu hesaplama, geminin yüzey alanını veya bazen de tonajını temsil etmek amacıyla kullanılır. Yine oldukça etkili diğer iki özellik ise geminin uzunluğu ile genişliği arasındaki oranları yansıtan f5 ve f6'dır. Bu oranlar, genellikle geminin yapısal özelliklerini ve şeklini yansıtır. Bu oranlar, geminin stabilizesini, hızını, manevra kabiliyetini ve diğer performans özelliklerini etkileyebilir. Örneğin, uzun ve dar gemiler genellikle daha hızlı olma eğilimindeyken, kısa ve geniş gemiler daha istikrarlı olabilir.

Modeli etkilediği görülen bir diğer özellik f9, geminin şeklini ve asimetrisini belirlemek için kullanılır. Bu ifade, geminin asimetrisini veya şeklindeki değişiklikleri analiz etmek için kullanılır. Negatif bir değer, genellikle geminin daha çok uzun

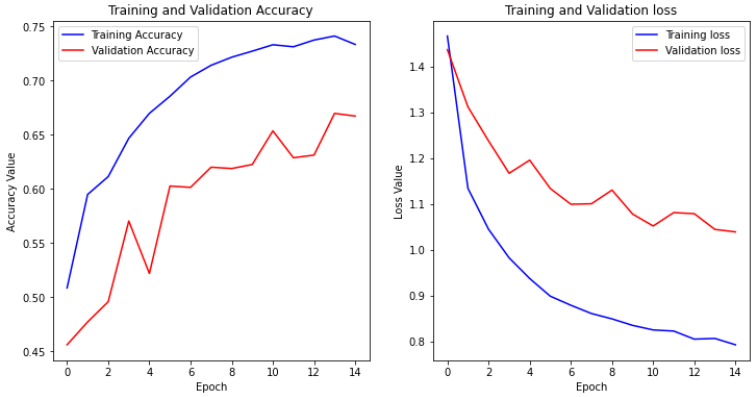
olduğunu ve pozitif bir değer, geminin daha çok geniş olduğunu gösterir. Sıfır değeri, geminin simetrik olduğunu gösterir.

### 4.3. Gemi Sınıflandırma Modeli Performans Analizi

3.4. Sınıflandırma Modelinin Oluşturulması bölümünde oluşturulan modele, 3.2. Veri Seti Ön İşleme bölümünde bahsedilen senaryo1, senaryo2 ve senaryo3 veri setlerinin ayrı ayrı girdi olarak verilerek denemeler yapılmıştır. Denemeler esnasında aynı modele aynı eğitim oranı ile veri setleri girdi olarak verilmiştir. Yapılan denemeler ile gözlemlenmek istenilen durumlar şu şekildedir:

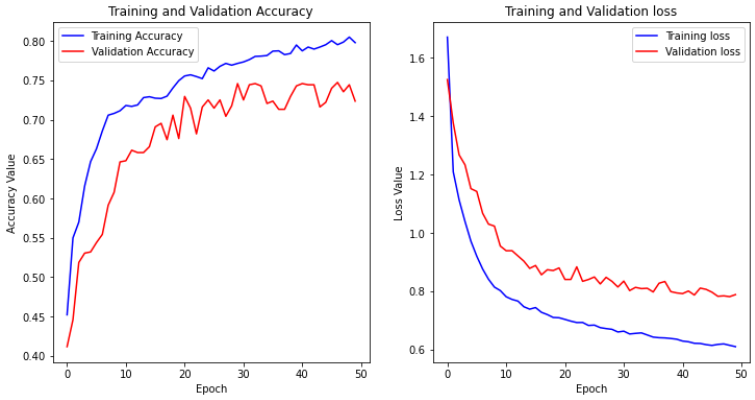
1. Senaryo1 veri seti ile suya batma verisinin olmadığı ancak türetilmiş verilerin olduğu durumda gemi sınıflandırma modeli doğruluk oranı ve performansı
2. Senaryo2 ile türetilmiş verilerin olmadığı ancak suya batma verisinin olduğu durumda gemi sınıflandırma modeli doğruluk oranı ve performansı
3. Senaryo3 ile hem suya batma verisinin hem de türetilmiş verilerin olduğu durum da gemi sınıflandırma modeli doğruluk oranı ve performansı

Senaryo1 veri setiyle gemi sınıflandırılması doğruluk oranının %73 olduğu gözlemlenmiştir. Senaryo1 veri setiyle gemi sınıflandırılması modelinin performansı Şekil 4-4'da verilmiştir. Modelin senaryo1 özellik önemine bakıldığında ise daha çok uzunluk ve genişlikten türetilmiş verilerin baskın olduğu Şekil 4-1'de görülmektedir.



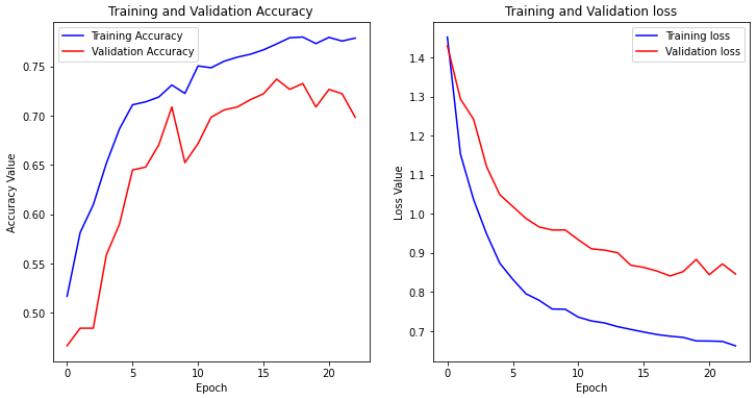

**Şekil 4-4.** Senaryo1 Verisi ile Sınıflandırma Modeli

Gemi sınıflandırılması modelinin Senaryo2 veri setiyle %78 doğruluk ile çalıştığı gözlemlenmiştir ve model performansı Şekil 4-5'de verilmiştir.




**Şekil 4-5.** Senaryo2 Verisi ile Sınıflandırma Modeli

Gemi sınıflandırılması modelinin Senaryo3 veri setiyle %74 doğruluk ile çalıştığı gözlemlenmiştir. Senaryo2 ile senaryo3 arasındaki tek fark, genişlik ve uzunluk verilerden türetilmiş verilerin senaryo3'te yer almasıdır. Dolayısıyla türetilmiş verilerin kullanılması, Şekil 4-6'de görüldüğü gibi doğruluğun azalmasına neden olduğu gözlemlenmiştir.



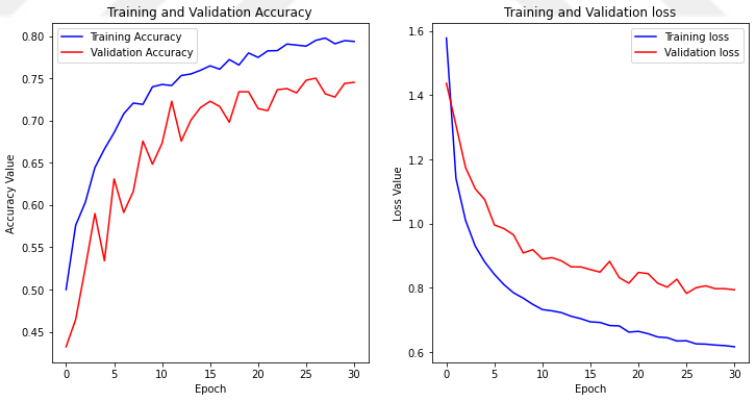

**Şekil 4-6.** Senaryo3 Verisi ile Sınıflandırma Modeli

Öte yandan, türetilmiş verilerin modele olan etkisi Şekil 4-3'da incelendiğinde, modeli oldukça etkiledikleri gözlemlenmektedir. Denemeler sayesinde, senaryo3 ve senaryo1 veri setleri arasındaki tek fark olan suya batma verisinin gemi sınıflandırılması üzerinde büyük bir etkisi olduğu gözlemlenmiştir. Suya batma verisi, başlangıç veri setinde en fazla eksik olan veridir. Dolayısıyla bu verinin modele olan

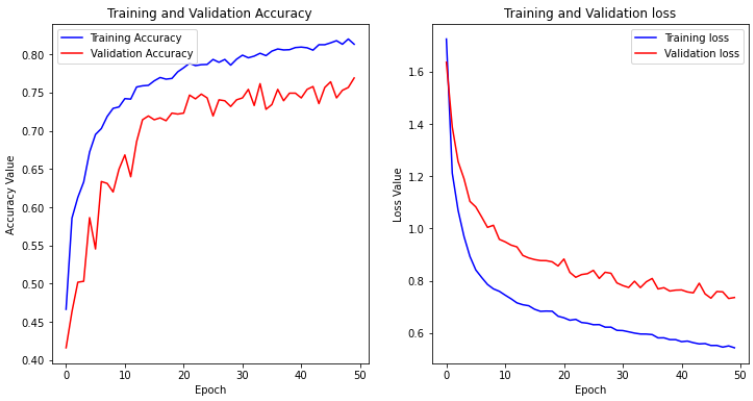
etkisi (Şekil 4-2 ve Şekil 4-3) göz önünde bulundurulduğunda, verinin tamamlanmasının önem arz ettiğini gösteriyor.

### 4.4 Veri Yükleme Sonrası Gemi Sınıflandırma Modeli Performans Analizi

Bu çalışmada, gemi sınıflandırma modelinde suya batma verisinin kullanılmasının önemli olduğu anlaşıldıktan sonra bu veri, kurulan K-NNR modeli ile tamamlanmıştır. Veri yükleme işlemi sonrası, Şekil 4-7'de görüldüğü gibi model performansı senaryo2 için %79 iken senaryo3 için Şekil 4-8'te görüldüğü %80 çıkmıştır. Veri tamamlamadan önceki doğruluk değerleri ile kıyaslandığında model doğruluklarında artış gözlemlenmiştir. Eksik verilerin tamamlanması, veri setindeki dengesizlikleri daha da vurgulayabilir. Eksik verilerin tamamlanması işlemi, eksik değerlerin çoğunluk sınıfına daha yakın değerlerle doldurulmasıyla gerçekleşebilir. Bu durumda, tamamlanan veri setinde çoğunluk sınıfına ait örneklerin sayısı artabilir ve azınlık sınıfına ait örneklerin sayısı azalabilir. Dolayısıyla, modelin doğruluğu düşebilir.




**Şekil 4-7.** Veri Yükleme Sonrası Senaryo2 Verisi ile Model Doğruluğu




**Şekil 4-8.** Veri Yükleme Sonrası Senaryo3 Verisi ile Model Doğruluğu

Doğruluk değerinin artması, aslında modelin gemi sınıflarını daha iyi tanımayı öğrenmesi anlamına gelebilir. Model, dengesiz sınıf dağılımını daha iyi dengelemeye çalışırken azınlık sınıfına daha fazla dikkat edebilir ve bu nedenle çoğunluk sınıfında daha fazla hatalı tahmin yapabilir. Bu nedenle, dengesiz sınıf dağılımına sahip bir veri setinde eksik verilerin tamamlanması sonucunda doğruluk değerinde artış görülmesi, modelin azınlık sınıfını daha iyi tanımayı öğrendiğini ve daha dengeli bir sınıflandırma performansı sergilediğini gösterebilir.

Tablo 3-4'te gözlemlenebildiği gibi suya batma verisinin tamamlanmasının ardından senaryo2 ve senaryo3 verilerinde yolcu, balıkçı ve çekme gemilerine ait veri sayısında artış meydana gelmesi, genel veri setine göre azınlık sayılabilecek bu sınıflardaki veri artışı sayesinde modelin bu türleri daha iyi öğrenmesi sağlanmıştır.

Dengesiz sınıf dağılımlarına sahip bir veri setinde, sınıflar arasında büyük bir dengesizlik olduğunda doğruluk değeri yanıltıcı olabilir. Veri yükleme öncesi ve sonrası olacak şekilde modelin performansını değerlendirmeye yardımcı olacak metrikler, Tablo 4-2'da gösterilmiştir. Tabloda senaryo1'e ait bir veri olmamasının nedeni, senaryo 1'in suya batma verilerinin girdi olarak değerlendirilmediği bir durumu yansıtmasıdır.

**Tablo 4-2.** Veri Yükleme Öncesi ve Sonrası Performans Metrikleri

| Metrik | Eğitim/ Test | Senaryo1 | | Senaryo2 | | Senaryo3 | |
|---|---|---|---|---|---|---|---|
| | | Yükleme Öncesi | Yükleme Sonrası | Yükleme Öncesi | Yükleme Sonrası | Yükleme Öncesi | Yükleme Sonrası |
| Hassasiyet (precision) | Eğitim | 0.7169 | - | 0.7877 | 0.8074 | 0.7730 | 0.7994 |
| | Test | 0.7172 | - | 0.7782 | 0.7994 | 0.7659 | 0.7994 |
| Geri Çağırma (recall) | Eğitim | 0.7351 | - | 0.7972 | 0.8099 | 0.7842 | 0.7975 |
| | Test | 0.7304 | - | 0.7814 | 0.7975 | 0.7816 | 0.7975 |
| F1 skoru | Eğitim | 0.7121 | - | 0.7820 | 0.8067 | 0.7712 | 0.8067 |
| | Test | 0.7073 | - | 0.7686 | 0.7948 | 0.7682 | 0.7948 |

Tablo 4-2'da görüldüğü üzere eksik verilerin tamamlanmasından sonra, model performansı artmıştır. Bunun nedeni eksik verilerin tamamlanması öncesi ve sonrası gemi tipine göre veri dağılımlarının farklı olması ve eksik verilerin tamamlanması sonrası azınlıkta olan gemi türlerinin azlığına vurgu yapılmamış olmasıdır. Bu sayede model bütün veri türlerini daha iyi öğrenmiştir. Ek olarak, aynı veri sayısı ve dağılıma sahip olan suya batma verisi içermeyen senaryo1 veri seti ile model doğruluğu %73 iken, eksik verilerin tamamlanması sonrası suya batma verisi içeren senaryo2'nin model doğruluğu %80 ve senaryo3'ün model doğruluğu %79 olduğu görülmüştür. Suya batma verisi yüklemesi ardından modelin performansının senaryo1'e kıyasla hem senaryo2'de hem senaryo3'te artması sayesinde eksik verilerin tamamlanmasının modele performansında iyileştirme yapıldığı ispatlanmıştır.

Öte yandan, senaryo2 ile senaryo3 arasındaki fark olan türetilmiş verilerin aslında sınıflandırma modeline girdi olarak eklenmesinin de modelin performansını negatif yönde etkilediği görülmüştür. Bu durumun oluşma nedeni, türetilmiş verilerin hesaplanmasında statik veriler için önemli bir öznitelik olduğunu ispatladığımız suya batma verisinin yer almamasından kaynaklandığı düşünülebilir. Bu durumda, türetilmiş verilerin modelin öğrenmesini etkilediği göz önünde bulundurarak, suya batma verisinin de hesaplamalara dâhil edildiği yeni özniteliklerin oluşturulması gelecekte yapılabilecek bir çalışmadır.

# 5. TARTIŞMA

Geçmişte yapılan gemi sınıflandırması ile ilgili çalışmalar incelenmiştir. Eksik verilerin veri setinden atıldığı ve suya batma verisinin de genelde çok fazla eksik olması nedeniyle girdi olarak alınmadığı görüldü. Yapılan çalışmalarda kimi zaman SAR görüntü verileri ile birlikte çalışabilmek adına, kimi zaman ise uzunluk ve genişliğin yetersiz görülmesi nedeniyle gemi karakteristiğini daha iyi açıklayabileceği düşünülen yeni verilerin türetildiği görüldü. Bu veri türetişinde de suya batma verisinin dâhil edilmediği görüldü.

Özellikle dengesiz dağılıma sahip veri setleri ile sınıflandırma yapılırken, doğru sınıflandırma yapabilmek için azınlık sayılan türden verilerin de öğrenilmesi gerekmektedir. Eksik verilerin tamamlanmasının dengesiz veri setlerine denge getirebileceği gibi azınlık verilerin azlığına daha çok vurgu yapabilir.

Özetle çalışma, liman ve deniz güvenliğine fayda sağlayabilecek gemi sınıflandırılması probleminde kullanılan veri setinde önemli özniteliklerin bulunmasını ve bu özniteliklerdeki eksik verilerin tamamlanması sayesinde veri setindeki dengesizliğin azalması ile daha doğru ve güvenilir bir sınıflandırma modeli oluşturulmasına olanak sağlar.

# 6. SONUÇ

Bu çalışmanın amacı, AIS verisi kullanılarak yapılan gemi sınıflandırmasında kullanılacak statik verilerindeki eksikliğin giderilmesi sayesinde sınıflandırma modelinin iyileştirilmesidir. Bu çalışmanın yapılabilmesi için üç ayrı veri seti üzerinden yapılan denemeler sonucu, suya batma verisinin gemi sınıflandırılmasında üzerinde oldukça etkili olduğu görülmüştür. Eksik verilerin atılması veya sıfır ile doldurulmasına karşı çıkan bu çalışmada, K-NNR modeli ile AIS statik verilerinden özellik önemi en yüksek olsan suya batma mesafesindeki eksik verilerin tamamlanması sayesinde gemi sınıflandırma modelinde doğruluğun ve performansın arttığı görülmüştür.

Başlangıç veri setinde oldukça eksik olan suya batma mesafesi verisinin veri setindeki eksikliğinin giderilmesinin veri dengesizliğinde azınlık sayılabilecek balıkçı, yolcu ve çekme gemilerindeki veri sayısında artış sağlaması ile gemi sınıfına göre veri setinin dağılımında bu gemi türleri için iyileşme gözlemlenmiştir. Ayrıca aynı veri sayısı ve dağılıma sahip olan suya batma verisi içermeyen senaryo1 veri seti ile model doğruluğu ile eksik verilerin tamamlanması sonrası suya batma verisi içeren senaryo2 ve senaryo3'ün model doğrulukları karşılaştırıldığında, eksik verilerin tamamlanmasının modele performansında iyileştirme yapıldığı ispatlanmıştır.

Bu çalışmada, eksik verilerin veri setinden atılmasının azınlık sınıflardan sayılabilecek balıkçı, yolcu ve çekme gemilerinin sınıf azlığına daha çok vurgu yaptığı görülmüştür. Bu nedenle senaryo2 ve senaryo3 veri setlerinde eksik verilerin tamamlanması öncesi veri dağılımı nedeniyle model doğruluğu ve performans oranlarının eksik verilerin tamamlanması sonrasına göre düşük olduğu gözlemlendi.

Daha önce birçok çalışmada gemi karakteristiğini ortaya koyduğu belirlendiği için kullanılan türetilmiş verilerden gemi boyutu (f3), geminin yüzey alanı, genişlik ve

uzunluk oranları (f5, f6) ve geminin şekli ve asimetrisi (f9) özelliklerinin gemi sınıflandırmasında geminin uzunluğu ve genişliğinin tek başına kullanılmasından daha önemli olduğu görülmüştür. Ancak kullanılan türetilmiş verilerin suya batma verisine göre özellik önemlerinin fazla olmasına rağmen modelin doğruluğunu ve performansını daha az etkiledikleri görülmüştür.

# 7. SINIRLAMALAR VE GELECEK ÇALIŞMA

Mevcut çalışmanın bazı sınırlamaları bulunmaktadır. Bu sınırlamalardan biri, veri setinin dengeli bir şekilde gemi türlerini barındırmamasıdır. Her gemi türünden homojen bir dağılıma sahip veri setiyle yapılacak sınıflandırma çalışmalarının doğruluğu daha yüksek olacaktır. Bu çalışmada eksik verilerin tamamlanmasının gemi sınıflandırma modeline olan etkisi araştırılırken eksik verilerin tamamlanması ile birkaç gemi türü için veri seti daha dengeli bir hale gelmiştir. Daha önce yapılan çalışmalarda da veri seti dengesizliğinin giderilmesi yerine belirli gemi türlerini kapsayacak çalışmalar yapıldığı görülmüştür. Ancak gelecek çalışmalarda birkaç farklı AIS verisi birleştirilerek gemi türü bakımından dengeli bir veri seti oluşturulması hedeflenerek tüm gemi türlerini sınıflandırabilen modellerin kurulması hedeflenebilir.

Bu çalışmada gemi sınıflandırması probleminde özellik önemi oldukça yüksek bulunan suya batma verisi kullanılarak yeni türetilmiş veriler ile veri seti zenginleştirilebilir. Bu çalışmada, uzunluk ve genişlik kullanılarak türetilen yeni özniteliklerin özellik önemlerinin, uzunluk ve genişlik verilerinin özellik öneminden daha büyük olduğu gözlemlenmiştir. Suya batma verisi ile de türetilecek yeni özniteliklerin özellik önemleri suya batma verisinden daha büyük olabilir.

## 8. BİLGİLENDİRME

Bu çalışma İstanbul Kültür Üniversitesi Bilimsel Araştırma Projeleri Koordinasyon Birimi Tarafından Desteklenmiştir. Proje numarası: İKÜ-BAP2012.

Veri seti talep edilmesi durumunda araştırma amaçlı olarak paylaşılacaktır.

# 9. KAYNAKLAR


[1] Singh, S. K., & Heymann, F. (2020, April). Machine learning-assisted anomaly detection in maritime navigation using AIS data. In 2020 IEEE/ION Position, Location and Navigation Symposium (PLANS) (pp. 832-838). IEEE.

[2] Huang, Yu, et al. "GPU-accelerated compression and visualization of large-scale vessel trajectories in maritime IoT industries." *IEEE Internet of Things Journal* 7.11 (2020): 10794-10812.

[3] Lv, Shenmin. "Construction of marine ship automatic identification system data mining platform based on big data." *Journal of Intelligent & Fuzzy Systems* 38.2 (2020): 1249-1255.

[4] Dobrkovic, A., Iacob, M. E., & van Hillegersberg, J. (2018). Maritime pattern extraction and route reconstruction from incomplete AIS data. *International journal of Data science and Analytics*, *5*, 111-136.

[5] Zhang, Daheng, Yingjun Zhang, and Chuang Zhang. "Data mining approach for automatic ship-route design for coastal seas using AIS trajectory clustering analysis." *Ocean Engineering* 236 (2021): 109535.

[6] Liu, C., & Chen, X. (2013). Inference of single vessel behaviour with incomplete satellite-based AIS data. The Journal of navigation, 66(6), 813-823.

[7] Zhong, H., Song, X., & Yang, L. (2019, July). Vessel classification from space-based ais data using random forest. In *2019 5th International Conference on Big Data and Information Analytics (BigDIA)* (pp. 9-12). IEEE.

[8] Gutierrez-Torre, A., Berral, J. L., Buchaca, D., Guevara, M., Soret, A., & Carrera, D. (2020). Improving maritime traffic emission estimations on missing data with CRBMs. *Engineering Applications of Artificial Intelligence*, *94*, 103793.

[9] Kim, Y., Steen, S., & Muri, H. (2022). A novel method for estimating missing values in ship principal data. *Ocean Engineering*, *251*, 110979.

[10] Iphar, Clément, Aldo Napoli, and Cyril Ray. "Detection of false AIS messages for the improvement of maritime situational awareness." *Oceans 2015-mts/ieee washington*. IEEE, 2015.

[11] Zhou, Y., Daamen, W., Vellinga, T., & Hoogendoorn, S. P. (2019). Ship classification based on ship behavior clustering from AIS data. *Ocean Engineering*, *175*, 176-187.

[12] T. Zhang et al., "HOG-ShipCLSNet: A Novel Deep Learning Network With HOG Feature Fusion for SAR Ship Classification," in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-22, 2022, Art no. 5210322, doi: 10.1109/TGRS.2021.3082759.

[13] Lang, H., Wu, S., & Xu, Y. (2018). Ship classification in SAR images improved by AIS knowledge transfer. IEEE Geoscience and Remote Sensing Letters, 15(3), 439-443.

[14] Yan, Zhenguo, Xin Song, and Lei Yang. "Research on AIS Data Aided Ship Classification in Spaceborne SAR Images." *Proceedings of the 2022 11th International Conference on Computing and Pattern Recognition*. 2022.

[15] Wang, Y., Yang, L., Song, X., Chen, Q., & Yan, Z. (2021). A Multi-Feature Ensemble Learning Classification Method for Ship Classification with Space-Based AIS Data. *Applied Sciences*, *11*(21), 10336.

[16] Adnan, F. A., Jamaludin, K. R., Wan Muhamad, W. Z. A., & Miskon, S. (2022). A review of the current publication trends on missing data imputation over three decades: direction and future research. *Neural Computing and Applications*, *34*(21), 18325-18340.

[17] Zhen, Rong, et al. "Maritime anomaly detection within coastal waters based on vessel trajectory clustering and Naïve Bayes Classifier." *The Journal of Navigation* 70.3 (2017): 648-670.

[18] Liang, Maohan, Yang Zhan, and Ryan Wen Liu. "MVFFNet: Multi-view feature fusion network for imbalanced ship classification." *Pattern Recognition Letters* 151 (2021): 26-32.

[19] Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

[20] Adnan, F. A., Zakaria, M. H., & Ibrahim, S. (2020). 60-year research history of missing data: a bibliometric review on Scopus database (1960–2019). *Appl Math Comput Intell*, *9*, 75-86.

[21] Resmî Gazete, 11 Eylül 2007 tarih ve 26640 sayılı, Otomatik tanımlama sistemi (AIS) Klas-B Cs cihazının, gemilere donatılmasına ve özelliklerine dair tebliğ.

[22] https://tr.wikipedia.org/wiki/D%C3%BCnyan%C4%B1n_en_uzun_gemileri_listesi

# 10. EKLER

Bulunmamaktadır.