T.C.
EGE ÜNİVERSİTESİ
Fen Bilimleri Enstitüsü

MAKİNE ÖĞRENMESİ TEKNİKLERİ KULLANILARAK TOKSİK ALG VE BİYOTOKSİN VARLIĞININ TAHMİNİ

Yüksek Lisans Tezi

Alper MEMİŞ

Bilgisayar Mühendisliği

İzmir
2019

T.C.
EGE ÜNİVERSİTESİ
Fen Bilimleri Enstitüsü

MAKİNE ÖĞRENMESİ TEKNİKLERİ KULLANILARAK TOKSİK ALG VE BİYOTOKSİN VARLIĞININ TAHMİNİ

Alper MEMİŞ

Danışman : Dr. Öğr. Üyesi Özgür GÜMÜŞ

Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Yüksek Lisans Programı

İzmir
2019

Alper MEMİŞ tarafından **Yüksek Lisans** tezi olarak sunulan "**Makine Öğrenmesi Teknikleri Kullanılarak Toksik Alg ve Biyotoksin Varlığının Tahmini**" başlıklı bu çalışma EÜ Lisansüstü Eğitim ve Öğretim Yönetmeliği ile EÜ Fen Bilimleri Enstitüsü Eğitim ve Öğretim Yönergesi'nin ilgili hükümleri uyarınca tarafımızdan değerlendirilerek savunmaya değer bulunmuş ve .....tarihinde yapılan tez savunma sınavında aday oybirliği/oyçokluğu ile başarılı bulunmuştur.

Jüri Üyeleri:		**İmza**
Jüri Başkanı	**: Dr. Öğr. Üyesi Özgür GÜMÜŞ**	**.....**
Raportör Üye	**: Doç. Dr. Şükrü YILDIRIM**	**.....**
Üye	**: Dr. Öğr. Üyesi Korhan KARABULUT**	**.....**

EGE ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ

ETİK KURALLARA UYGUNLUK BEYANI

EÜ Lisansüstü Eğitim ve Öğretim Yönetmeliğinin ilgili hükümleri uyarınca Yüksek Lisans Tezi olarak sunduğum "Makine Öğrenmesi Teknikleri Kullanılarak Toksik Alg ve Biyotoksin Varlığının Tahmini" başlıklı bu tezin kendi çalışmam olduğunu, sunduğum tüm sonuç, doküman, bilgi ve belgeleri bizzat ve bu tez çalışması kapsamında elde ettiğimi, bu tez çalışmasıyla elde edilmeyen bütün bilgi ve yorumlara atıf yaptığımı ve bunları kaynaklar listesinde usulüne uygun olarak verdiğimi, tez çalışması ve yazımı sırasında patent ve telif haklarını ihlal edici bir davranışımın olmadığını, bu tezin herhangi bir bölümünü bu üniversite veya diğer bir üniversitede başka bir tez çalışması içinde sunmadığımı, bu tezin planlanmasından yazımına kadar bütün safhalarda bilimsel etik kurallarına uygun olarak davrandığımı ve aksinin ortaya çıkması durumunda her türlü yasal sonucu kabul edeceğimi beyan ederim.

.... / .... / 2019

İmzası

Alper MEMİŞ

ÖZET

MAKİNE ÖĞRENMESİ TEKNİKLERİ KULLANILARAK TOKSİK ALG VE BİYOTOKSİN VARLIĞININ TAHMİNİ

MEMİŞ, Alper

Yüksek Lisans Tezi, Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı: Dr. Öğr. Üyesi Özgür GÜMÜŞ

Kasım 2019, 55 sayfa

Bu tezde deniz suyunda yapılan analizlerdeki toksik alg türlerinin, midye gibi kabuklu deniz ürünlerinde tespit edilen biyotoksin ile arasında var olduğu düşünülen korelasyonun ortaya çıkarılması amaçlanmış olup, bunun yanında su ve midye üzerinde yapılan analizlerin bu duruma etkisi incelenmiştir. Midyelerde biriken biyotoksin, midyeyi tüketen insanlar için ciddi bir tehlikedir. Biyotoksin ve toksik fitoplankton varlığının tahmini de bu çalışmanın temelindedir. Üretim şartlarının sağlanması için zamana, deniz suyu ve midye analizleri için ise yüksek harcamalara ihtiyaç duyulmaktadır. Bu çalışmada ise biyotoksin ve toksik fitoplankton varlığının tahmini ile bu harcamaların önüne geçilebilineceği düşünülmektedir. Burada yapılan tahminler makine öğrenmesi teknikleri ile gerçekleştirilmiştir.

Bu çalışma için yaklaşık 3 yıllık bir veri ele alınmıştır. Veri içeriğinde midyeden ve sudan alınan örneklerin fiziksel, kimyasal ve biyolojik analiz sonuçları bulunmaktadır. Mevcut veriye bakıldığında bir dengesizlik gözlenmektedir. Yani biyotoksin varlığına çok az rastlanmıştır. Bu durum literatürde “dengesiz veri” (imbalanced data) olarak adlandırılmaktadır. Dolayısıyla veri dengeli bir hale getirildikten sonra üzerinde makine öğrenmesi çalışmaları yapılmıştır. Yapılan çalışmalar neticesinde en başarılı tahmin, Rastgele Orman metodu ile elde edilmiştir.

Anahtar sözcükler: Makine Öğrenmesi, Rastgele Orman, Dengesiz Veri, Tahmin, Kabuklu Deniz Ürünleri, Biyotoksin.

ABSTRACT

PREDICTION OF TOXIC ALGAE AND BIOTOXIN PRESENCE USING MACHINE LEARNING TECHNIQUES

MEMİŞ, Alper

MSc in Computer Eng.

Supervisor: Asst. Prof. Özgür GÜMÜŞ

November 2019, 55 pages

In this thesis, it is aimed to reveal the correlation between the toxic algae species and the biotoxin detected in shellfish such as mussels in seawater analysis. In addition, the effect of water and mussel analysis on this situation was examined. The accumulation of biotoxins in mussels is a serious danger for people who consume mussels. Estimation of the presence of biotoxin and toxic phytoplankton is also the basis of this study. Time is needed to meet production conditions and high expenditure is required for sea water and mussel analyzes. In this study, it is thought that these expenditures can be prevented by estimating the presence of biotoxin and toxic phytoplankton. Estimates made here are made by machine learning techniques.

For this study, a data of approximately 3 years has been discussed. Physical, chemical and biological analysis results of samples from mussel and water is available in the data content. An imbalance is observed in the current data. In other words, the presence of biotoxin is very rare. This is called imbalanced data in the literature. Therefore, after the data was stabilized, machine learning studies were carried out. As a result of the studies, the most successful estimation was obtained by the Random Forest method.

Keywords: Machine Learning, Random Forest, Imbalanced Data, Prediction, Shellfish, Biotoxin

ÖNSÖZ

Bu çalışmada makine öğrenmesi yöntemlerinin özel bir konuda nasıl sonuçlar verdiği incelenmiştir.

Herkesin bildiği gibi makine öğrenmeleri uygulamalarında bir veriye ihtiyaç vardır. Bu çalışmada kullanılan veri ve çalışmanın amacı, tezi özel kılan taraftır.

Konunun belirlenmesindeki öncülüğü ile İzmir Tarım ve Orman İl Müdürlüğü, Balıkçılık ve Su Ürünleri Şube Müdürü Sayın Levent ÖZKAN' a teşekkür ederim. Tezin hazırlanma sürecinin her aşamasında değerli bilgilerini ve zamanını benden esirgemeyerek her fırsatta çalışmamla yakından ilgilenen, eleştirileriyle yol gösteren danışman hocam Dr. Öğr. Üyesi Özgür GÜMÜŞ' e teşekkür ederim.

İZMİR

…/…/2019

Alper MEMİŞ

İÇİNDEKİLER

Sayfa

İÇ KAPAK ..... …ii

KABUL ONAY SAYFASI ..... …vi

ETİK KURALLARA UYGUNLUK BEYANI ..... …iv

ÖZET ..... …vii

ABSTRACT viii

ÖNSÖZ x

İÇİNDEKİLER DİZİNİ xii

ŞEKİLLER DİZİNİ xv

TABLOLAR DİZİNİ xvi

SİMGELER VE KISALTMALAR DİZİNİ xvii

1.GİRİŞ 1

2.GENEL BİLGİLER 3

3.GEREÇ VE YÖNTEM 7

3.1 Ortam ve Dil 7

3.2 Veri Seti 7

3.3 Veri Ön İşleme 8

İÇİNDEKİLER (devam)

Sayfa

3.3 Metotlar.....10

3.4 Değerlendirme Ölçütleri.....12

4. BULGULAR.....14

4.1 Biyotoksin Varlığı Tahmini için SMOTE Öncesi Bulgular.....14

4.2 Biyotoksin Varlığı Tahmini için SMOTE Sonrası Bulgular.....16

4.3 Biyotoksin ile Diğer Parametreler Arasındaki Korelasyon Bulguları.....17

4.4 Toksik Fitoplankton Varlığı Tahmini için SMOTE Öncesi Bulgular.....20

4.5 Toksik Fitoplankton Varlığı Tahmini için SMOTE Sonrası Bulgular.....22

4.6 Toksik Fitoplankton ile Deniz Suyu Parametreleri Arasındaki Korelasyon Bulguları.....23

4.6.1 SMOTE Öncesi Korelasyon Bulguları.....23

4.6.2 SMOTE Sonrası Korelasyon Bulguları.....26

5.SONUÇ VE TARTIŞMA.....29

KAYNAKLAR DİZİNİ.....31

TEŞEKKÜR.....34

ÖZGEÇMİŞ.....35

ŞEKİLLER DİZİNİ

Şekil Sayfa

3.1 Hata Matrisi (Confusion Matrix).....12

4.1 Biyotoksin varlığı tahmini için SMOTE uygulanmamış veri seti.....14

4.2 Biyotoksin varlığı tahmini için SMOTE uygulanmış veri seti.....16

4.3 Tüm özniteliklerin biyotoksin ile korelasyonu.....17

4.4 Tüm parametreler arasındaki korelasyon matrisi 18

4.5 Biyotoksin varlığı tahmininde Random forest ile "feature importance" sonuçları.....19

4.6 Toksik fitoplankton varlığı tahmini için SMOTE uygulanmamış veri seti….....20

4.7 Toksik fitoplankton varlığı tahmini için SMOTE uygulanmış veri seti. 22

4.8 SMOTE öncesi deniz suyu özniteliklerinin toksik fitoplankton ile korelasyonu.....23

4.9 SMOTE öncesi öznitelikler arasındaki korelasyon matrisi 24

4.10 Toksik fitoplankton varlğı tahmininde SMOTE öncesi Random forest ile "feature importance" sonuçları.....25

4.11 SMOTE sonrası deniz suyu özniteliklerinin toksik fitoplankton ile korelasyonu.....26

4.12 SMOTE sonrası deniz suyu parametrelerinin arasındaki korelasyon matrisi…. 27

4.13 Toksik fitoplankton varlğı tahmininde SMOTE sonrası Random forest ile "feature importance" sonuçları.....28

TABLOLAR DİZİNİ

Tablo Sayfa

2.1 Dünya su ürünleri üretim miktarı 3

3.1 Veri seti detayı 8

3.2 Veri ön işlemesi öncesi deniz suyu analizi parametreleri veri kesiti 9

3.3 Veri ön işlemesi öncesi midye analizi parametreleri veri kesiti (Ecoli sütunu ve Biyotoksin türleri) 9

3.4 Veri ön işlemesi sonrası deniz suyu ve midye analizi parametreleri birleştirilmiş veri kesiti 10

4.1 Biyotoksin varlığı tahmini için makine öğrenmesi tekniklerinin SMOTE uygulanmayan veri üzerinde denenmesi sonucu elde edilen değerlendirme ölçütleri.. 14

4.2 Biyotoksin varlığı tahmini için makine öğrenmesi tekniklerinin SMOTE uygulanmayan veri üzerinde çapraz geçerleme yöntemi uygulanarak denenmesi sonucu elde edilen değerlendirme ölçütleri.. 15

4.3 Biyotoksin varlığı tahmini için SMOTE ve Çapraz Geçerleme (Cross Validation) sonucu elde edilen değerlendirme ölçütleri.. 16

4.4 Toksik fitoplankton varlığı tahmini için makine öğrenmesi tekniklerinin SMOTE uygulanmayan veri üzerinde denenmesi sonucu elde edilen değerlendirme ölçütleri 20

4.5 Toksik fitoplankton varlığı tahmini için makine öğrenmesi tekniklerinin SMOTE uygulanmayan veri üzerinde çapraz geçerleme yöntemi uygulanarak denenmesi sonucu elde edilen değerlendirme ölçütleri.. 21

4.6 Toksik fitoplankton varlığı tahmini için SMOTE ve Çapraz Geçerleme (Cross Validation) sonucu elde edilen değerlendirme ölçütleri.. 22

SİMGELER VE KISALTMALAR DİZİNİ

Simgeler Açıklama

Kısaltmalar

SMOTE Synthetic Minority Oversampling Technique (Sentetik Azınlık Örnekleme Tekniği)

QDA Quadratic Discriminant Analysis (Kuadratik Diskriminant Analizi)

SVM Support Vector Machine (Destek Vektör Makinesi)

1. GİRİŞ

Ülkemizin üç yanı denizlerle çevrilidir. Buna rağmen su ürünleri tüketimi, diğer (kırmızı et, tavuk vb.) ürünlere karşı daha azdır. Hâlbuki su ürünleri beslenme için faydalı ve gereklidir. Midye eti de; diğer su ürünleri gibi iyi bir besin kaynağıdır (Mol, 2006).

Faydalı özelliklerinin yanı sıra, midyelerin tüketiminde bazı durumlarda zehirlenmelere ve hastalıklara sebebiyet verebilmektedir. Bunları özetleyecek olursak; midyelerin tazeliğinden kaynaklı problemler, taşınması ve işlem görmesi sırasındaki mikrobiyolojik faktörler ve midyelerde bulunabilecek zararlı madde ve mikroorganizmalar. Midyeler, yaşadıkları doğal ortamlarında bölgenin etkisi ile virüs, ağır metal, biyotoksinler ve çeşitli bakterileri içerebilmektedirler. Bunun sebebi ise midyelerin bulundukları bölgedeki suyu filtre etmeleridir. Sudaki zararlı maddeler bu sayede midyede birikmiş olur.

Bu çalışmada, sudaki toksik algler ile midye içerisinde biriken biyotoksin varlığı arasındaki ilişkinin gösterilmesi amaçlanmıştır. Burada daha kapsamlı tahmin sonuçları elde etmek adına toksik alglerin yanı sıra, su parametreleri ve midye üzerinde yapılan analiz parametrelerinden de faydalanılmıştır. Makine öğrenmesi yöntemleri, bu ilişkileri çıkarmakta yardımcı olabilecektir. Bu tezde biyotoksin ve toksik fitoplankton varlığının tahmini ve diğer parametrelerle olan korelasyonunun incelenmesi için birden çok makine öğrenmesi yöntemi denenmiştir. Bunların tahmin sonuçlarını değerlendirmek için standart performans ölçütleri kullanılmış olup, kendi içlerinde karşılaştırma yapılmıştır.

Bu tezde, biyotoksin ve toksik fitoplankton varlığının tahmini ve etki eden parametrelerin belirlenmesine katkıda bulunmak amaçlanmıştır. Biyotoksin ve toksik fitoplankton varlığının tahmin yöntemi geliştirildiğinde birçok harcamadan ve zamandan tasarruf edilmiş olunacaktır. Ayrıca bildiğimiz kadarıyla bu çalışma, biyotoksin ve toksik fitoplankton varlığının tahmini konusunda Türkiye'de yapılan ilk çalışmadır.

Tezin bölüm içeriklerinde, Bölüm 2'de literatür ve genel bilgiler, Bölüm 3'te veri seti, makine öğrenmesi metotları ve değerlendirme

ölçütleri, Bölüm 4’te çalışma sonunda elde edilen bulgular, Bölüm 5’te bulgular üzerindeki yorumlamalar ve sonuçlar anlatılmıştır.

2. GENEL BİLGİLER

Dünya genelinde kaliteli ve sağlıklı besin talebi giderek artmaktadır. Su ürünleri yüksek ve nitelikli protein içeriği ile vücudumuzdaki birçok sistemin güçlenmesinde rol almaktadır. İçerdiği birçok vitamin ve mineral ile gelişim ve beslenme açısından büyük bir önem taşımaktadır. Dünya genelinde artan kaliteli ve sağlıklı besin ihtiyacını karşılamak adına su ürünleri yetiştiriciliği giderek gelişmiş ve hızlı bir şekilde artış göstermiştir.

Dünyada su ürünleri yetiştiriciliği başlangıcı iki bin yıl öncesine kadar dayanmaktadır. Dünyada en hızlı büyüyen gıda üretim sektörü, FAO tarafından, su ürünleri yetiştiriciliği olarak belirlenmiştir. Ayrıca günümüzde çok popüler bir konu olmak üzere giderek yaygınlaşmaktadır (Subasinghe et al., 2009). Bu konuya ilişkin dünya su ürünleri üretim miktarı tablosu Tablo 2.1 de gösterilmiştir.

Yıllar	İç sular (milyon ton)	Deniz (milyon ton)	Toplam (milyon ton)
2011	38.6	23.2	61.8
2012	42.0	24.4	66.4
2013	44.8	25.4	70.2
2014	46.9	26.8	73.7
2015	48.6	27.5	76.1
2016	51.4	28.7	80.0

Tablo 2.1. Dünya su ürünleri üretim miktarı (FAO, 2017).

Su ürünleri yetiştiriciliğinin bir kısmını da midyeler oluşturmaktadır. Gıda bileşimi açısından midyeler; doymamış yağ asitlerinin yüksek olduğu, doymuş yağ asitlerini düşük oranda içermekte olan, iyi birer omega-3 kaynaklarıdır. Midye etinde ortalama 9 g/100 g (100 gram ette yaklaşık 9 gram) esansiyel aminoasitler (vücut tarafından üretilmeyen, besinlerle alınması zorunlu olan) bulunmaktadır (Kinsella, 1988; Ackman, 1988).

Midyelerin beslenmesi diğer deniz ürünleri yetiştiriciliğinden farklı olarak, deniz suyundaki inorganik ve organik maddeleri süzerek gerçekleşmekte ve ilave bir yemleme yapılmamaktadır. Özellikle su içerisinde var olan fitoplankton, midyelerin beslenmesinde çok önemli bir yer tutmaktadır. Bu nedenle kabuklu üretiminin kontrolünün ve izlenmesinin sağlanması amacıyla balık yetiştiriciliğine ek olarak bir takım yasal sınırlamalar ve düzenlemeler getirilmiştir.

Midyelerin biyolojik yapıları gereği yaşadığı ve yetiştiği alanlar kıyıya yakın alanlar olduğundan dolayı kimyasal, biyolojik ve fiziksel çevresel faktörler (insan kaynaklı kirleticiler, endüstriyel atıklar vb.) ile kontaminasyona (kirlenmeye) maruz kalabilmektedir. Sudaki kimyasallar (civa, kadmiyum ve kurşun gibi ağır metaller) ve fitoplankton artışı ile ortaya çıkan toksik maddeler (PSP, ASP, DSP, NSP) dolayısı ile midyelerin insanlar tarafından tüketilmesi sonucunda zehirlenmeler, bazı hastalıklar ve ölümlere varan sonuçlar ortaya çıkabilmektedir.

Midyeler bünyelerine hem sudaki asılı partikülleri hem de suda çözünen metali alabilmektedirler. Yapılan bazı çalışmalarda suyun kimyasal yapısına karşı hassas oldukları belirlenmiştir. Kadmiyum ve kurşun bulunan ortamların incelenmesi sonucunda midye larvalarında anormallik belirlenmiş ve kadmiyumun ergin bireyleri kurşundan daha fazla etkilediği belirtilmiştir (Varlık,1991).

Tüm bu riskli durumlara bakıldığında, midye üretimi yapılacak alanlarda fiziksel, kimyasal ve biyolojik verilerin takip edilmesi gerekmektedir. Avrupa Birliği Uyum Yasaları çerçevesinde; "Avcılık Veya Yetiştiricilik Yoluyla Elde Edilen Çift Kabuklu Yumuşakça Üretim Alanlarının Belirlenmesi, Sınıflandırılması, Ürün Alımına Açılıp Kapatılması Ve Numune Alımına İlişkin Uygulama Talimatı" kapsamında midye yetiştiriciliği yapmak için üretim alanlarının sınıflandırılması ve izleme programı uygulanması gerekmektedir. Bu çalışmalar T.C. Tarım ve Orman Bakanlığı ve İl Müdürlükleri gözetiminde ve denetiminde yürütülmektedir.

İzleme programı kapsamında üründe düzenli olarak mikrobiyolojik, biyotoksin, ağır metal analizi ile birlikte deniz suyunda ise su kalitesi ve

toksik alg analizleri yapılmaktadır. Biyotoksin ve fitoplankton analizleri haftada bir, mikrobiyolojik analizler ise iki haftada bir olmak üzere periyodik olarak gerçekleştirilmektedir.

Gelişen teknoloji ve bununla birlikte endüstrinin ilerlemesiyle birçok ağır metal doğaya atık olarak bırakılır. Ağır metaller, deniz ortamında oldukça düşük konsantrasyonlarda bulunmaktadır ve sucul ortamlara girişleri, insan kaynaklı veya doğal olabilmektedir. En önemli metal girdileri ise endüstriyel kaynaklıdır. Deniz ortamında yaşayan bitki ve hayvanlar bu metalleri bünyelerinde biriktirirler (Bilgin and Uluturhan-Suzer, 2015). Deniz suyundaki ağır metaller sudaki organizmalarda toksik etki göstermektedir. Bu tür organizmalara toksik algler adı verilmektedir. Denizde yaşayan çift kabuklu yumuşakçalarda ise biyotoksin birikmesine sebebiyet verdiği düşünülmektedir.

Çift kabuklu yumuşakçalar, en başta Avrupa ve Pasifik ülkeleri olmak üzere ciddi bir tüketim miktarına sahiptir. Bu canlı türlerinin en yaygın ve popüler olanı midyelerdir. Avcılık yolu ile de elde edilen midyelerin ticari üretimleri sağlık açısından tüketim riski taşıması sebebiyle genellikle kültür yoluyla yapılmaktadır. Çift kabuklu yumuşakça ticaret hacminin büyümesi için hijyen koşullarının yerine getirilmesi gerekmektedir (Çolakoğlu ve Çolakoğlu, 2013) .

Üretim şartlarının sağlanması için zamana, deniz suyu ve midye analizi için ise yüksek harcamalara ihtiyaç duyulmaktadır. Bu çalışmada, biyotoksin ve toksik alg varlığının tahmini ve bunların kendi arasındaki korelasyonun bulunması amaçlanmaktadır. Ayrıca bunlara etki eden en önemli parametreler incelenmiştir. Bu çalışmanın sonucunda elde edilecek olası yüksek tahmin performansı ile büyük bir ekonomik kazanç ve zaman tasarrufu sağlanacağı düşünülmektedir.

Bu ve buna benzer çalışmalarda kontamine kabuklu deniz hayvanlarının tüketimi, insanlarda ciddi hastalıklara hatta ölümlere neden olabileceğinden sıkça bahsedilmektedir. Bu kontaminasyon sonucunda kabuklu deniz ürünü çiftliklerinin kapatılması gerekebilmektedir (D'este et al., 2012). Hastalıkları ve ölümleri önlemek adına yapılan çalışmalarda kabuklu deniz ürünü çiftliklerinin kapatılmasının nedenini saptamak ve açılması / kapatılmasını öngörmek için yeni makine öğrenme algoritmaları

geliştirilmiş ve test edilmiştir (Rahman and D'este, 2015).

Bu çalışmada esas nokta, insan sağlığıdır. İnsan sağlığını tehdit eden çift kabuklu deniz ürünlerinin önceden tespiti yapılarak, yetiştiricilik yapılan bölgelerin kontrol ve denetim altına alınmasını sağlamaktır.

3. GEREÇ VE YÖNTEM

3.1 Ortam ve Dil

Bu tezin uygulama kısmında, son zamanlarda popüler olan ve gittikçe kullanımı artan Python dilinden faydalanılmıştır. Python'ı öne çıkaran asıl etkenler ise makine öğrenmesi alanında geniş kütüphanelere, yerli-yabancı kaynak açısından bolluğa ve geniş bir yelpazeye sahip olmasıdır (Python, 2019).

Platform olarak ise bu dilin kullanımını kolaylaştıran özellikle makine öğrenmesi alanında yardımcı olan Jupyter Notebook kullanılmıştır. Anaconda Navigator uygulaması içerisinde kullanılan Jupyter Notebook, açık kaynaklı bir web uygulamasıdır (Anaconda, 2019). Jupyter Notebook, veri temizleme ve dönüştürme, sayısal simülasyon, istatistiksel modelleme, veri görselleştirme, makine öğrenmesi ve çok daha fazla kullanım alanını içermektedir. İçerisinde barındırdığı pratik kütüphane kullanımları, çıktıların okunurluğu ve düzeni birçok ihtiyacı karşılamaktadır (Jupyter, 2019).

3.2 Veri Seti

Bu tezde kullanılan veriler, İzmir İl Tarım ve Orman Müdürlüğü ve Balıkesir İl Tarım ve Orman Müdürlüğü, Balıkçılık ve Su Ürünleri Şube Müdürlükleri tarafından yaklaşık 3 yıl boyunca düzenli olarak; deniz suyu ve midyeden alınan örneklerin analizi sonucu elde edilmiştir. Ayrıca günlük ortalama deniz suyu sıcaklığı verileri Meteoroloji Genel Müdürlüğünden temin edilmiştir. Sıcaklık verileri, örneklemelerin yapıldığı midye çiftliklerine en yakın Meteoroloji Genel Müdürlüğü istasyonlarından alınmıştır. Daha sonra diğer veriler ile tarih bazında eşleştirilmiştir. Deniz suyu örneği 2 haftada bir alınırken, midye örneği haftada bir kez alınmaktadır. Bu sebeple deniz suyu örneklerinin alındığı gün esas alınmıştır. Mücbir sebeplerden (hava koşulları, cihaz arızası vb.) ötürü veri alınamayan haftalar da mevcuttur. Sonuç olarak bu veriler düzenlendiğinde, eksik veri içeren satır ve sütunlar temizlendiğinde 285 örnek (satır) ve bu örneklere ilişkin 11 parametre (sütun) içeren veri kümesi elde kalmıştır. Bu veri setinin boyutu makine öğrenmesi algoritmaları için yetersizdir ve bu veri kümesinde odaklandığımız biyotoksin varlığı yalnız 7 satırda (örnek),

toksik fitoplankton varlığı ise 57 satırda mevcuttur. 11 parametre (sütun) şunlardır; E.Coli, Civa (Hg), Kadmiyum (Cd), Kurşun (Pb), Nitrat, Tuzluluk, PH, Silis, Sıcaklık (Deniz suyu sıcaklığı günlük ortalaması), Toksik Fitoplankton ve Biyotoksin. Tablo 3.1 ‘ de veri seti detaylı olarak gösterilmektedir.

Deniz Suyu Analizi Parametreleri	Midye Analizi Parametreleri
Nitrat	E. Coli
Tuzluluk	Civa
Ph	Kadmiyum
Silis	Kurşun
Sıcaklık	Biyotoksin
Toksik Fitoplankton	

Tablo 3.1: Veri seti detayı

3.2.1 Veri Ön İşleme

Bu tezde, yaklaşık 3-4 yıl boyunca ve 2 haftada bir düzenli olarak alınan veriler kullanılmıştır. Bu verinin çok olması beklenir. Fakat resmi tatil günleri, hava muhalefetleri ve teknik problemlerden dolayı veri alınamamasından kaynaklı veri miktarı beklenenden düşüktür. Ayrıca çeşitli sütunlardan oluşan bu veri grubunda, herhangi bir sütundaki eksik veri sebebiyle veri grubundaki eksik veri içeren satır tamamen çıkartılmıştır. Bu çıkarılan satırlar veri boyutuna oranla bakıldığında çok ufak kalmaktadır. İzmir İl Tarım ve Orman Müdürlüğü ve Balıkesir İl Tarım ve Orman Müdürlüğü verisi birleştirilirken eşleşmeyen parametreler (sütunlar) de çıkarılmıştır.

Makine öğrenmesi yöntemlerini kullanmak için verilerin tamamen sayısallaştırılması gerekmektedir. Elde kalan verilerde Toksik Fitoplankton, Civa, Kadmiyum, Kurşun ve Nitrat parametreleri için veri grubunun çok büyük çoğunluğu “Tespit Edilemedi” şeklinde görülmektedir. Bu ifadenin belli bir eşik değerine göre mi yazıldığı tespit edilememiştir. Ayrıca “Tespit Edilemedi” dışında kalan veriler ise çok azdır. Bu sebepten “Tespit Edilemedi” olarak nitelendirilen veriler “0” olarak, geriye kalan az miktardaki veriler de “1” olarak değiştirilmiştir. Detaylandırılacak olursa, örneğin civa sütununda “1” var ise midyede civa görülmüş anlamına

gelmektedir. Eğer "0" ise midyede civaya rastlanılmadığı ya da eşik değerin altında kaldığı anlamı taşımaktadır. Biyotoksin sonuçları ham veride, biyotoksin türleri ayrı ayrı olarak kaydedilmiştir. Bu çalışma için Biyotoksin türleri gözetilmeksizin, eğer herhangi bir Biyotoksin türü gözlenmiş ise o bölgede var anlamına gelen "1" olarak adlandırılmıştır. Yani biyotoksinin herhangi bir türü gözlenmiş ise tek sütunda birleştirilmiştir. Eğer hiçbir biyotoksin türüne rastlanmamış ise "0" olarak adlandırılmıştır. Tablo 3.3 'de midye analizi verilerinden bir kesit gösterilmiştir. Bu tablo Ecoli ve Biyotoksin türlerini içermektedir. Tabloda gösterilen Biyotoksin türlerinden herhangi birinin tespit edilmesi, veri ön işlemesi sonucu Biyotoksin sütununda "1" ile ifade edilmiştir. Tablo 3.2 ve Tablo 3.3 'de verilerin ön işleme öncesi hali ve Tablo 3.4 ' de verilerin ön işleme sonrası deniz suyu analizi verileri ile midye analizi verileri birleştirilmiş şekli gösterilmiştir.

Toksik Fitoplankton	Amonyum İyonu	Fosfat İyonları	Nitrat İyonları	Tuzluluk	PH	Nitrit İyonları	Silis
	Tespit Edilmedi	Tespit Edilmedi	Tespit Edilmedi	25,18	8,50	Tespit Edilmedi	0,076
Tespit Edildi	Tespit Edilmedi	Tespit Edilmedi	Tespit Edilmedi	26,8	9,13	Tespit Edilmedi	0,056
	Tespit Edilmedi	Tespit Edilmedi	Tespit Edilmedi	27,05	8,84	Tespit Edilmedi	0,175
	Tespit Edilmedi	Tespit Edilmedi	Tespit Edilmedi	24,99	8,75	Tespit Edilmedi	0,064
	Tespit Edilmedi	Tespit Edilmedi	0,676	24,85	8,49	0,084	0,61
	Tespit Edilmedi	Tespit Edilmedi	Tespit Edilmedi	27,49	8,66	Tespit Edilmedi	0,14
	Tespit Edilmedi	Tespit Edilmedi	Tespit Edilmedi	29,93	8,77	Tespit Edilmedi	0,016

Tablo 3.2 Veri ön işlemesi öncesi deniz suyu analizi parametreleri veri kesiti

ECOLİ	PSP	ASP	DSP	OKADAİK ASİT	YESSO TOKSİNLER	SAXOT OKSİN	AZASPİR
3500 EMS/100g	ÜRÜN İÇ SICAKLIĞI NEDENİYLE NUMUNE RED						
330 EMS/100g	Bulunamadı	Tespit Edilemedi	Bulunamadı	Analiz Yapılamadı	Analiz Yapılamadı	Saptanamadı	Analiz Yapılamadı
140 EMS/100g	Bulunamadı	Tespit Edilemedi	Bulunamadı	Analiz Yapılamadı	Analiz Yapılamadı	Saptanamadı	Analiz Yapılamadı

Tablo 3.3 Veri ön işlemesi öncesi midye analizi parametreleri veri kesiti (Ecoli sütunu ve Biyotoksin türleri)

ECOLİ	CİVA	KADMIYUM	KURSUN	NITRAT	TUZLULUK	PH	SILIS	SICAKLIK	TOKSIK FITOPLANKTON	BIYOTOKSIN
3500	0	0	0	0	25,18	8,50	0,076	19,2	0	0
330	0	0	0	0	26,8	9,13	0,056	18,3	1	0
140	0	0	0	0	27,05	8,84	0,175	16,4	0	0
40	0	0	0	0	24,99	8,75	0,064	17	0	0
90	0	0	0	1	24,85	8,49	0,61	17,1	0	0
20	0	0	0	0	27,49	8,66	0,14	16	0	0
50	0	0	0	0	29,93	8,77	0,016	15,9	0	0

Tablo 3.4 Veri ön işlemesi sonrası deniz suyu ve midye analizi parametreleri birleştirilmiş veri kesiti

3.3 Metotlar

Önceki bölümde bahsedilen veriler, çeşitli makine öğrenmesi algoritmalarına tabi tutulmuştur. Bu denemelerin sonucunda başta doğruluk (accuracy) değerleri %90'ları bulmuş olsa da yanıltıcı bir sonuç olduğu düşük f1-skor (f1-score), duyarlılık (recall) ve kesinlik (precision) değerleri ile tespit edilmiştir. Duyarlılık değerleri %0'larda çıkmıştır. Veri sayısının da yetersiz olduğu düşünüldüğünde, bu problemi çözmek için başta literatürde çapraz geçerleme (cross validation) olarak geçen yöntem denenmiştir (Nar, 2018). Çapraz geçerleme, bir makine öğrenmesi metodunda yapılan testin hatasını daha iyi tahmin edebilmek için model seçiminde kullanılan bir tekniktir. Çapraz geçerlemenin temelinde, eğitim verileri setinden doğrulama kümeleri olarak bilinen örnek gözlem bölümlerini oluşturmak yatmaktadır. Bu tezde kullanılan yöntem çapraz geçerleme yöntemlerinden birisi olan "Stratified-K-Fold" çapraz geçerleme yöntemidir (StratifiedKFold, 2019). "Stratified-K-Fold" çapraz geçerleme yönteminde veri kümesi k tane alt kümeye bölünmekte ve tahmin modeli oluşturma işlemi k kez tekrarlanmaktadır. Her defasında, k alt kümelerinden biri test kümesi olarak kullanılırken diğer k-1 alt kümesi bir eğitim kümesi oluşturmak üzere bir araya getirilir. Ardından, tüm k denemelerindeki ortalama hatası hesaplanır. Stratified-K-Fold' un en belirgin özelliği, az olan pozitif değerleri her test ve eğitim kümesine yüzdeleri koruyarak dağıtmasıdır. Yani biyotoksin görülen satırlar yüzdeleri korunarak dağıtılmıştır. Fakat "Stratified-K-Fold" çapraz geçerleme yöntemi de duyarlılık değerlerini pek istenilen yönde değiştirmemiştir. Bunun

sonrasında bu problemin asıl sebebinin literatürde “dengesiz veri” (imbalanced data) olarak geçen dengesiz veri dağılımı olduğu anlaşılmıştır (Kartal ve Özen, 2017). Yani veri kümesinde yukarıda bahsettiğimiz gibi 285 satır verinin sadece 7 satırında biyotoksin, 57 satırında ise toksik fitoplankton varlığı görülmüştür. Makine öğrenmesi algoritmaları biyotoksin ve toksik fitoplankton varlığı olan satırları tahmin etmek yerine, veri setinde büyük oranda mevcut olan biyotoksin ve toksik fitoplankton varlığının olmadığı satırları tahmin etmiş ve yüksek oranda doğruluk sonuçları vermiştir. Bu, makine öğrenmesi algoritması için eğitim seti olarak ayrılan verinin büyük bir çoğunluğunun, doğal olarak biyotoksin ve toksik fitoplankton varlığı içermeyen satırlardan oluşmasından kaynaklanmaktadır (He and Garcia, 2008). Bu istenilmeyen bir durumdur. Bu tezde amaçlanan ise biyotoksin ve toksik fitoplankton varlığının tahmin edilebilmesi ve bunun diğer parametrelerle korelasyonunun tespit edilmesidir.

Dengesiz veri problemini çözmek adına geliştirilen birçok yöntem mevcuttur. Bunlara genel itibariyle bakılacak olursa, “resampling” başlığı altında ikiye ayrılmaktadır. “Undersampling”, veri setindeki çok olan veriyi azınlık sınıfla eşleştirmek için bazı gözlemleri rasgele veya belli şartlarda silme işlemidir. “Oversampling”, veri setindeki az olan veriyi çoğunluk sınıfla eşleştirmek için bazı gözlemleri rasgele veya belli şartlarda çoğaltma işlemidir (Altini, 2015). Bu çalışmada veri setinin kısıtlı olması sebebiyle “oversampling” tercih edilmiştir. “Oversampling” yöntemlerinden rasgele veri arttırma yerine, Sentetik Azınlık Örnekleme Tekniği “Synthetic Minority Over-sampling Technique” (SMOTE) tercih edilmiştir. SMOTE yönteminde her azınlık sınıfı örneği alınır ve bu örneğin k komşusunun herhangi birine ya da tümüne bakılarak sentetik örnekler oluşturulur. Böylece azınlık sınıfı aşırı örneklenmiş olur (Chawla, 2002). Diğer örnekleme yöntemlerinden en temel farkı, azınlık sınıfındaki örneklerin kopyalanarak çoğaltılması yerine yakın komşularına bakılarak sentetik örneklerin üretilmesidir. SMOTE tekniği ile azınlık veriler çoğaltıldıktan sonra makine öğrenmesi için yeterli veri seti oluşturulmuştur.

Bu çalışmada 11 farklı tekli ve topluluk (ensemble) makine öğrenmesi yöntemleri kullanılmıştır. Kullanılan makine öğrenmesi yöntemleri şunlardır: Logistic Regression, Nearest Neighbors, Linear SVM, Gradient Boosting Classifier, Decision Tree, Random Forest, Neural Net,

Naive Bayes, AdaBoost, QDA, Gaussian Process. Ayrıca iyi sonuç veren makine öğrenmesi teknikleri üzerinde, “grid search” yöntemi ile modeller için en iyi parametreler belirlenmiştir (Lerman, 1980). “Grid search” yöntemi, modeller için uygun olan çeşitli parametreleri, makine öğrenmesi fonksiyonları üzerinde deneyerek en iyi doğruluk oranlarını veren parametreleri bulan bir yöntemdir (Joseph, 2018). Makine öğrenmesi yöntemleri sonuçlarını değerlendirirken öncelikle doğruluk ölçütü ile başlanılmıştır. Fakat doğruluk çoğu zaman tek başına yeterli değildir. Bu sebeple bu çalışmada birçok model performans değerlendirme ölçütlerinden faydalanılmıştır (Alpaydın, 2009).

Korelasyon bulgularını göstermek için ise korelasyon matrisi ve korelasyon grafiği kullanılmıştır (Shaikh, 2018). Ayrıca makine öğrenmesi teknikleri arasında en iyi performansı gösteren random forest yöntemi ile sonuca etki eden önemli öznitelikleri bulmak için “feature importance” metodundan yararlanılmıştır (Lewinson, 2019).

3.4 Değerlendirme Ölçütleri

Hata matrisi (confusion matrix), makine öğrenmesinde hedef verisi önceden belli olan veri setlerinden elde edilen modellerin performansını değerlendirmek üzere en sık kullanılan yöntemdir. Modelin, pozitif ve negatif örnekleri içerisinde barındıran test veri setini ne ölçüde sınıflandırdığını gösteren matris’e hata matrisi (Şekil 3.2) denir (Arat, 2018).

		Gerçek	
		Pozitif(1)	Negatif(0)
Tahmin Edilen	Pozitif(1)	DP	YP
	Negatif(0)	YN	DN

Şekil 3.1: Hata matrisi

DP (Doğru Pozitif): Test verisindeki değer ile modelin ürettiği sınıf aynıdır (her ikisi de pozitiftir). Yani doğru sınıflandırma yapılmıştır.
YN (Yanlış Negatif): Test verisindeki değer ile modelin ürettiği sınıf farklıdır, gerçekte pozitif iken negatif sınıflandırılmıştır. Yani hatalı sınıflandırma yapılmıştır.
YP (Yanlış Pozitif) : Gerçek değer negatif iken pozitif olarak sınıflandırılmıştır. Yani hatalı sınıflandırma yapılmıştır.
DN (Doğru Negatif): Gerçek değer negatif iken negatif olarak sınıflandırılmıştır. Yani doğru sınıflandırma yapılmıştır.

Doğruluk, modelin girdilerden doğru şekilde tahmin edilen gözlemlerin, toplam gözleme oranıdır (Shang, 2015).

$$Doğruluk = \frac{Doğru\ Pozitif + Doğru\ Negatif}{DP + YP + YN + DN}$$

Kesinlik, modelin girdilerden pozitif sınıf etiketini tahmin etmekteki etkililiğini gösterir. Doğru sınıflandırılan pozitif örneklerin toplam pozitif örnek sayısına oranıdır (Shang, 2015).

$$Kesinlik = \frac{Doğru\ Pozitif}{Doğru\ Pozitif + Yanlış\ Pozitif}$$

Duyarlılık, doğru sınıflandırılan pozitif örneklerin toplam pozitif tahmin edilen örneklere oranıdır (Shang, 2015).

$$Duyarlılık = \frac{Doğru\ Pozitif}{Doğru\ Pozitif + Yanlış\ Negatif}$$

F1-Skor, kesinlik ve duyarlılık performans değerlendirme ölçütlerinin harmonik ortalamasıdır. Her iki ölçütü birlikte değerlendirme imkânı verir. (Powers, 2011)

$$F1_{Skor} = 2 \times \frac{Kesinlik \times Duyarlılık}{Kesinlik + Duyarlılık}$$

4. BULGULAR

4.1 Biyotoksin Varlığı Tahmini için SMOTE Öncesi Bulgular

Bu tezde birçok bulgu mevcuttur. Gereç ve yöntem bölümünde bahsedilen makine öğrenmesi yöntemleri, öncelikle SMOTE uygulanmayan veriler üzerinde denenmiştir. SMOTE uygulanmamış verilerin 278 'i biyotoksin gözlenmeyen, 7 'si ise biyotoksin gözlenen satırlardan oluşmaktadır. Şekil 4.1 ' de veri setinin matrisi gösterilmektedir.

```
0    278
1      7
Name: Biyotoksin, dtype: int64
```

Şekil 4.1 Biyotoksin varlığı tahmini için SMOTE uygulanmamış veri seti

Biyotoksin için yapılan çalışmalarda kullanılan veri seti bütün öznitelikleri (EColi, Civa, Kadmiyum, Kurşun, Nitrat, Tuzluluk, Ph, Silis, Sıcaklık, Toksik Fitoplankton) içermektedir. Veri seti üzerinde uygulanan makine öğrenmesi teknikleri doğruluk, f1- skor, kesinlik ve duyarlılık bazında değerlendirildiğinde Logistic Regression 0,97 doğruluk, 0 f1-skor, 0,69 kesinlik ve 0 duyarlılık ile en iyi sonucu vermiştir. Sırasıyla diğer makine öğrenmesi teknikleri ile elde edilen sonuçlar Tablo 4.1'de görülmektedir.

	Eğitim Doğruluk	Test Doğruluk	Duyarlılık	Kesinlik	f1_skor
Logistic Regression	0,976526	0,972222	0	0,694953	0
Nearest Neighbors	0,976526	0,972222	0	0,028578	0
Linear SVM	0,976526	0,972222	0	0,092559	0
Random Forest	1	0,972222	0	0,702835	0
Neural Net	0,976526	0,972222	0	1.656.716	0
AdaBoost	1	0,972222	0	0,058614	0
Gradient Boosting Classifier	1	0,958333	0	0,391935	0
QDA	1	0,958333	0	0,870573	0
Gaussian Process	0,99061	0,958333	0	0,746556	0
Decision Tree	1	0,930556	0	0,001222	0
Naive Bayes	0,549296	0,472222	1	0,001099	1

Tablo 4.1 Biyotoksin varlığı tahmini için makine öğrenmesi tekniklerinin SMOTE uygulanmayan veri üzerinde denenmesi sonucu elde edilen değerlendirme ölçütleri

Fakat tablo 4.1'de görülen doğruluk değerlerinin yüksek olması yanıltıcıdır. Veri sayısının az olması buradaki yanıltıcı faktör olarak düşünülebilir. Bunun için gereç ve yöntem bölümünde anlatılan, çapraz geçerleme metodu ile verinin farklı bölgelerinden kesitler alarak, yani verinin farklı bölgelerinden test ve eğitim kümeleri oluşturularak, tekrar değerlendirmeye tabi tutulmuştur. Bunun sonucunda ise Logistic Regression ve diğer makine öğrenmesi teknikleri duyarlılık, yani kısaca biyotoksin varlığını tahmin etme konusunda tekrar başarısız sonuçlar vermişlerdir. Veri seti üzerinde 4 katlı (fold) çapraz geçerleme sonrası çıkan sonuçların ortalama değerleri Tablo 4.2'de görülmektedir.

	Eğitim Doğruluk	Test Doğruluk	Duyarlılık	Kesinlik	f1_skor
AdaBoost	1	0,971975	0,125	0,25	0,166667
Naive Bayes	0,588315	0,606869	0,875	0,050354	0,094678
Logistic Regression	0,975445	0,975497	0	0	0
Nearest Neighbors	0,975445	0,975497	0	0	0
Linear SVM	0,975445	0,975497	0	0	0
Gradient Boosting Classifier	1	0,971926	0	0	0
Decision Tree	1	0,957988	0	0	0
Random Forest	1	0,975497	0	0	0
Neural Net	0,975445	0,975497	0	0	0
QDA	0,997674	0,972025	0	0	0
Gaussian Process	0,992969	0,954418	0	0	0

Tablo 4.2 Biyotoksin varlığı tahmini için makine öğrenmesi tekniklerinin SMOTE uygulanmayan veri üzerinde çapraz geçerleme yöntemi uygulanarak denenmesi sonucu elde edilen değerlendirme ölçütleri

Tablo 4.2' de de görüldüğü gibi test doğruluk sonuçları yüksek doğruluk oranına sahiptir. Fakat çapraz geçerleme yöntemi uygulanması sonucunda da duyarlılık ve f1-skor sonuçlarında istenen başarı elde edilememiştir.

4.2 Biyotoksin Varlığı Tahmini için SMOTE Sonrası Bulgular

Bir sonraki aşamada SMOTE, veri setine uygulanmıştır. Bu sayede veri setinde azınlık olarak bulunan biyotoksin varlığı olan satırlar ile biyotoksin varlığı olmayan satırların sayıları eşitlenmiştir. Şekil 4.2 ‘ de veri setinin matrisi gösterilmektedir.

```
1.0    278
0.0    278
Name: Biyotoksin, dtype: int64
```

Şekil 4.2 Biyotoksin varlığı tahmini için SMOTE uygulanmış veri seti

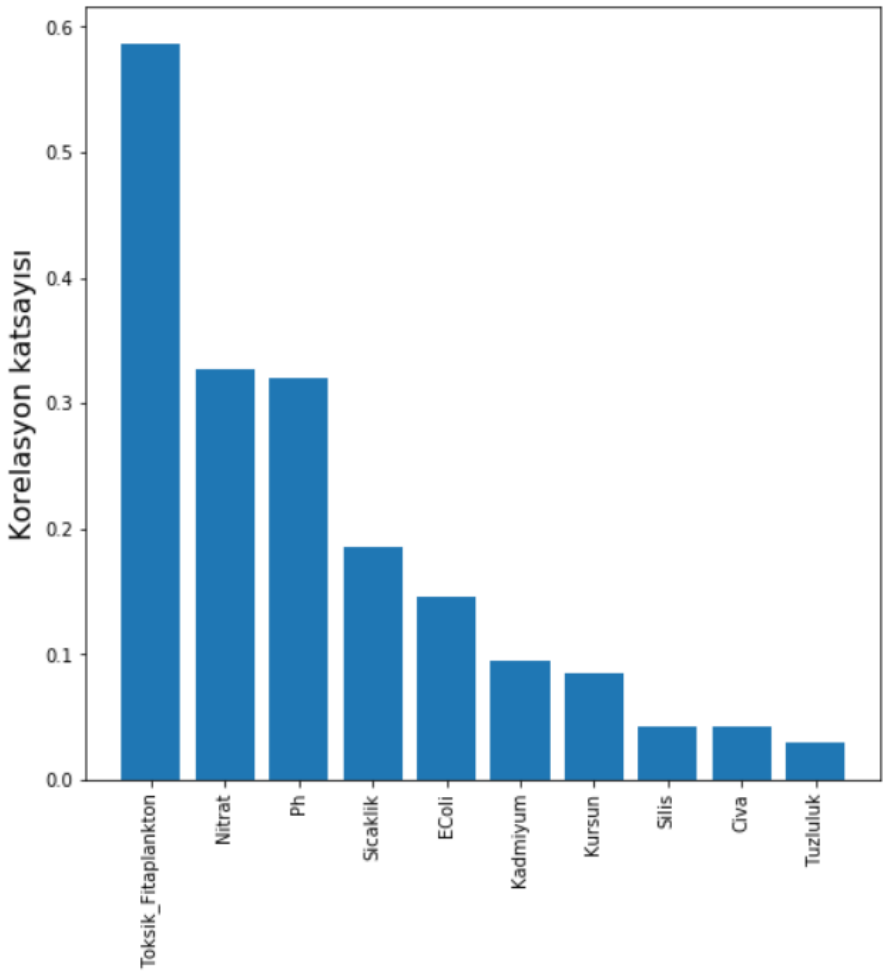
Yeni oluşan veri setine makine öğrenmesi yöntemlerini ve 4 katlı çapraz geçerleme uyguladığımızda, en iyi performansı Random Forest göstermiştir. Tablo 4.3‘ de sonuçlar görülmektedir.

	Eğitim Doğruluk	Test Doğruluk	Duyarlılık	Kesinlik	f1_skor
Random Forest	1	0,982013	0,978364	0,98566	0,981933
Gradient Boosting Classifier	1	0,980254	0,985611	0,975295	0,980392
AdaBoost	1	0,978468	0,978416	0,978718	0,978516
Linear SVM	0,979015	0,964053	0,996377	0,936172	0,965243
Decision Tree	1	0,962293	0,985559	0,942815	0,963339
Gaussian Process	0,982011	0,955072	1.000.000	0,91773	0,957055
Nearest Neighbors	0,948444	0,931703	1.000.000	0,880113	0,936158
Neural Net	0,899918	0,891977	0,949741	0,853368	0,898096
Naive Bayes	0,807534	0,805849	0,996377	0,722382	0,837222
Logistic Regression	0,832743	0,830952	0,866822	0,810107	0,836709
QDA	0,543217	0,528701	0,057402	0,5	0,101685

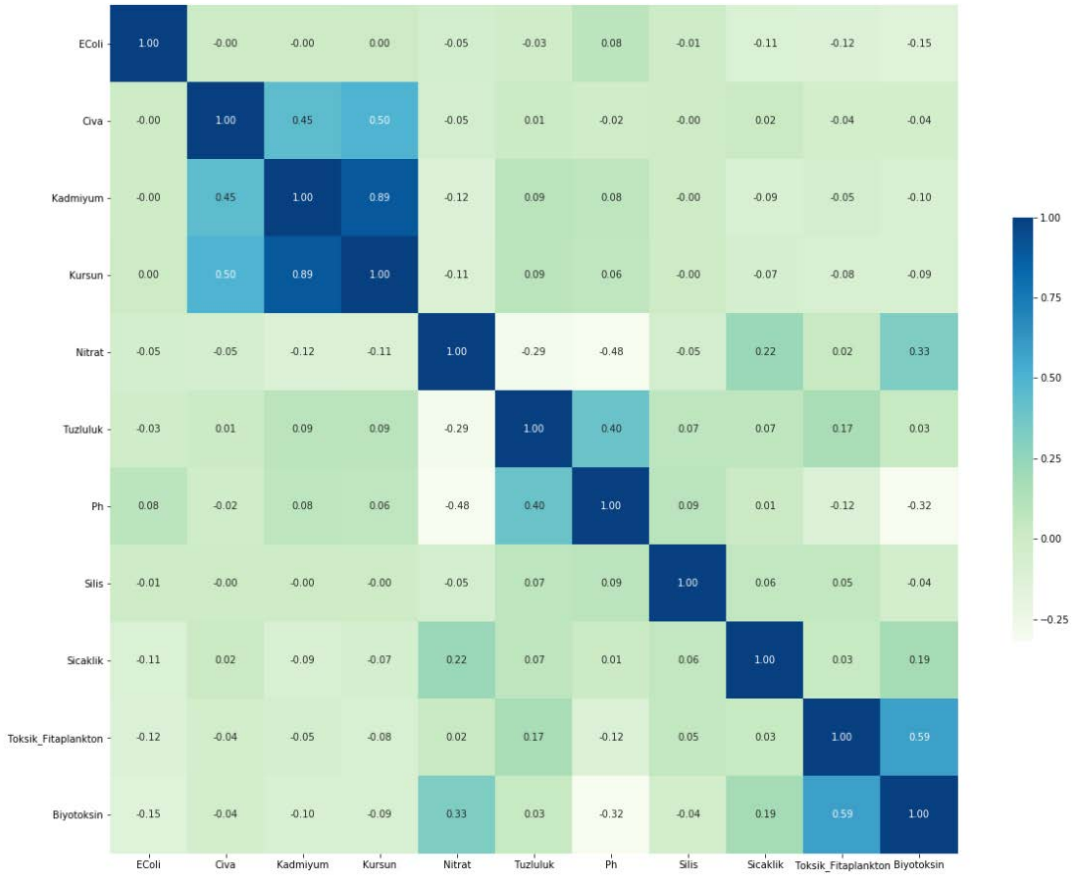
Tablo 4.3 Biyotoksin varlığı tahmini için SMOTE ve Çapraz Geçerleme (Cross Validation) sonucu elde edilen değerlendirme ölçütleri

4.3 Biyotoksin ile Diğer Parametreler Arasındaki Korelasyon Bulguları

SMOTE uygulandıktan sonra elde edilen veri kümesinde, tezin de temel amaçlarından olan biyotoksinin, toksik fitoplankton (toksik algler) ile arasında var ise korelasyonun tespit edilmesi için korelasyon grafiği ve korelasyon matrisi oluşturulmuştur. Bunun sonucunda toksik fitoplankton, beklenildiği gibi biyotoksin ile korelasyonu en yüksek öznitelik olarak ortaya çıkmıştır. Şekil 4.3'de tüm özniteliklerin biyotoksin ile korelasyon grafiği, Şekil 4.4'de ise tüm özniteliklerin birbirleriyle ilişkisini ifade eden korelasyon matrisi gösterilmiştir.

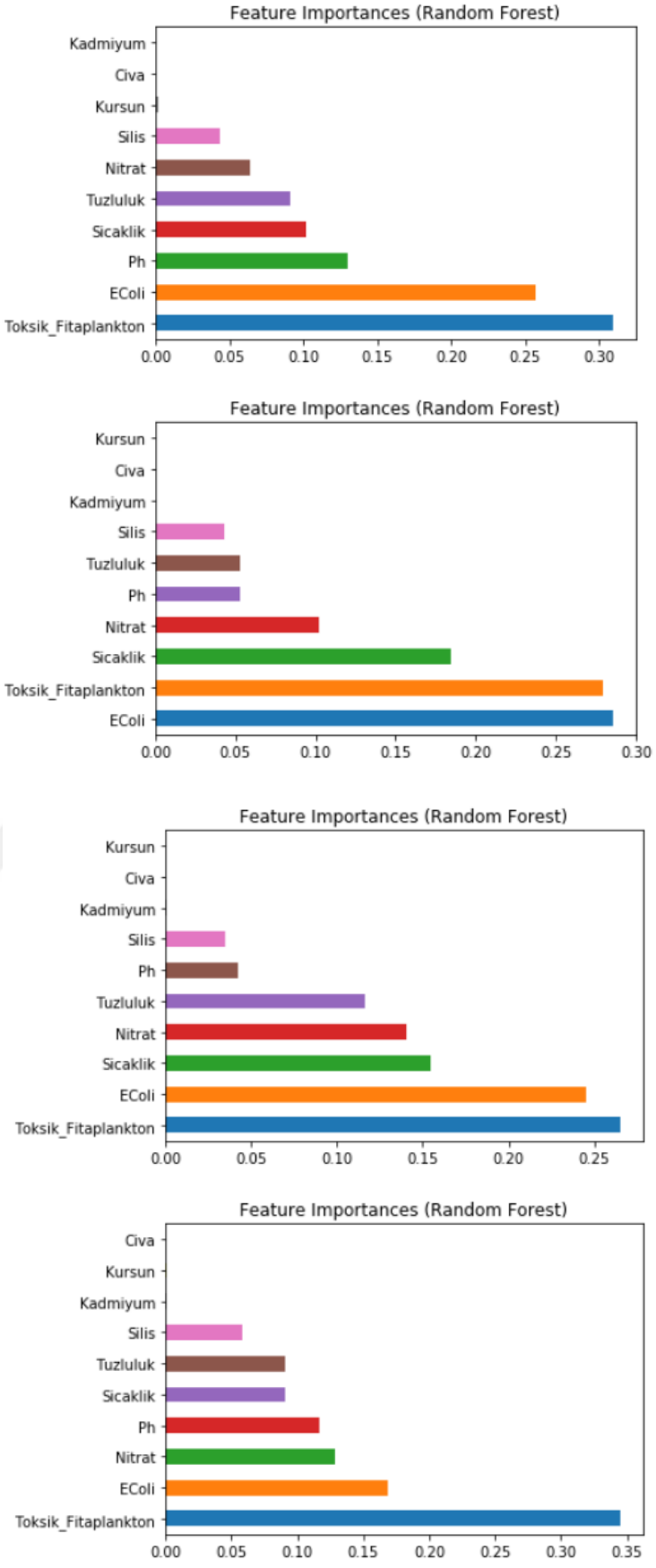


Şekil 4.3 Tüm özniteliklerin biyotoksin ile korelasyonu



Şekil 4.4 Tüm parametreler arasındaki korelasyon matrisi

Son olarak “feature importance” olarak bilinen, tahmin sonucuna etki eden önemli özniteliklerin belirlenmesine çalışılmıştır. Bunun için random forest makine öğrenmesi ve 4 katlı çapraz geçerleme yönteminden yararlanılmıştır. Bunun sonucunda toksik fitoplankton, biyotoksin varlığının tahmininde veri setindeki öznitelikler arasında en önemli öznitelik olarak bulunmuştur. Şekil 4.5’ deki grafikte, 4 katlı çapraz geçerlemedeki her bir kat, tüm özniteliklerin önem oranları ile gösterilmiştir.



Şekil 4.5 Biyotoksin varlığı tahmininde Random forest ile “feature importance” sonuçları

4.4 Toksik Fitoplankton Varlığı Tahmini için SMOTE Öncesi Bulgular

Bu tezde birçok bulgu mevcuttur. Gereç ve yöntem bölümünde bahsedilen makine öğrenmesi yöntemleri, öncelikle SMOTE uygulanmayan veriler üzerinde denenmiştir. SMOTE uygulanmamış verilerin 228'i toksik fitoplankton gözlenmeyen, 57'si ise toksik fitoplankton gözlenen satırlardan oluşmaktadır. Şekil 4.6 ' da veri setinin matrisi gösterilmektedir.

```
0    228
1     57
Name: Toksik_Fitaplankton, dtype: int64
```

Şekil 4.6 Toksik fitoplankton varlığı tahmini için SMOTE uygulanmamış veri seti

Toksik Fitoplankton varlığı tahmini için yapılan çalışmalarda kullanılan veri seti, sadece deniz suyu analizi parametrelerini (Nitrat, Tuzluluk, Ph, Silis, Sıcaklık, Toksik Fitoplankton) içermektedir. Veri seti üzerinde uygulanan makine öğrenmesi teknikleri doğruluk, f1- skor, kesinlik ve duyarlılık bazında değerlendirildiğinde Nearest Neighbors 0,81 doğruluk, 0,28 f1-skor, 0,002 kesinlik ve 0,28 duyarlılık ile en iyi sonucu vermiştir. Sırasıyla diğer makine öğrenmesi teknikleri ile elde edilen sonuçlar Tablo 4.4'de görülmektedir.

	Eğitim Doğruluk	Test Doğruluk	Duyarlılık	Kesinlik	f1_skor
Nearest Neighbors	0,807512	0,819444	0,285714	0,002331	0,285714
Gaussian Process	0,920188	0,75	0,214286	0,01975	0,214286
Decision Tree	1	0,625	0,214286	0,001737	0,214286
AdaBoost	0,910798	0,763889	0,071429	0,050024	0,071429
Logistic Regression	0,802817	0,805556	0	0,014628	0
Neural Net	0,802817	0,805556	0	0,105625	0
Naive Bayes	0,802817	0,805556	0	0,001053	0
QDA	0,802817	0,805556	0	0,00229	0
Linear SVM	0,826291	0,791667	0	0,002621	0
Random Forest	1	0,791667	0	0,759789	0
Gradient Boosting Classifier	1	0,652778	0	0,426798	0

Tablo 4.4 Toksik fitoplankton varlığı tahmini için makine öğrenmesi tekniklerinin SMOTE uygulanmayan veri üzerinde denenmesi sonucu elde edilen değerlendirme ölçütleri

Fakat tablo 4.4'de görülen doğruluk değerlerinin yüksek olması yanıltıcıdır. Veri sayısının az olması buradaki yanıltıcı faktör olarak düşünülebilir. Bunun için gereç ve yöntem bölümünde anlatılan, çapraz geçerleme metodu ile verinin farklı bölgelerinden kesitler alarak, yani verinin farklı bölgelerinden test ve eğitim kümeleri oluşturularak, tekrar değerlendirmeye tabi tutulmuştur. Veri seti üzerinde 4 katlı (fold) çapraz geçerleme sonrası çıkan sonuçların ortalama değerleri Tablo 4.5'de görülmektedir. Buna göre Nearest Neighbors ve diğer makine öğrenmesi teknikleri duyarlılık açısından toksik fitoplankton varlığını tahmin etme konusunda biyotoksin varlığını tahmin etmeye nazaran daha iyi sonuçlar vermiştir. Fakat yine de bu sonuçlar 0,33 den küçük olduğu için başarısız olarak kabul edilebilir.

	Eğitim Doğruluk	Test Doğruluk	Duyarlılık	Kesinlik	f1_skor
Decision Tree	1	0,71562	0,332143	0,310268	0,320115
Gaussian Process	0,920468	0,726379	0,247619	0,286029	0,261671
Gradient Boosting Classifier	0,98948	0,757874	0,194048	0,320076	0,238291
AdaBoost	0,91229	0,743887	0,195238	0,270563	0,220714
Nearest Neighbors	0,836257	0,743936	0,158333	0,275198	0,200216
Random Forest	1	0,768339	0,122619	0,361111	0,179926
Linear SVM	0,842092	0,782473	0,071429	0,279762	0,108281
QDA	0,781202	0,782619	0,116667	0,097222	0,106061
Naive Bayes	0,760086	0,75132	0,116667	0,067308	0,085366
Logistic Regression	0,803508	0,800029	0	0	0
Neural Net	0,803508	0,800029	0	0	0

Tablo 4.5 Toksik fitoplankton varlığı tahmini için makine öğrenmesi tekniklerinin SMOTE uygulanmayan veri üzerinde çapraz geçerleme yöntemi uygulanarak denenmesi sonucu elde edilen değerlendirme ölçütleri

Tablo 4.5'de de görüldüğü gibi test doğruluk sonuçları yüksek doğruluk oranına sahiptir. Fakat çapraz geçerleme yöntemi uygulanması sonucunda da duyarlılık ve f1-skor sonuçlarında istenen başarı elde edilememiştir.

4.5 Toksik Fitoplankton Varlığı Tahmini için SMOTE Sonrası Bulgular

Bir sonraki aşamada SMOTE, veri setine uygulanmıştır. Bu sayede veri setinde azınlık olarak bulunan toksik fitoplankton varlığı olan satırlar ile toksik fitoplankton varlığı olmayan satırların sayıları eşitlenmiştir. Şekil 4.7 ‘ de veri setinin matrisi gösterilmektedir.

```
1.0    228
0.0    228
Name: Toksik_Fitaplankton, dtype: int64
```

Şekil 4.7 Toksik fitoplankton için SMOTE uygulanmış veri seti

Yeni oluşan veri setine makine öğrenmesi yöntemlerini ve 4 katlı çapraz geçerleme uyguladığımızda, en iyi performansı Random Forest göstermiştir. Tablo 4.6‘ da sonuçlar görülmektedir.

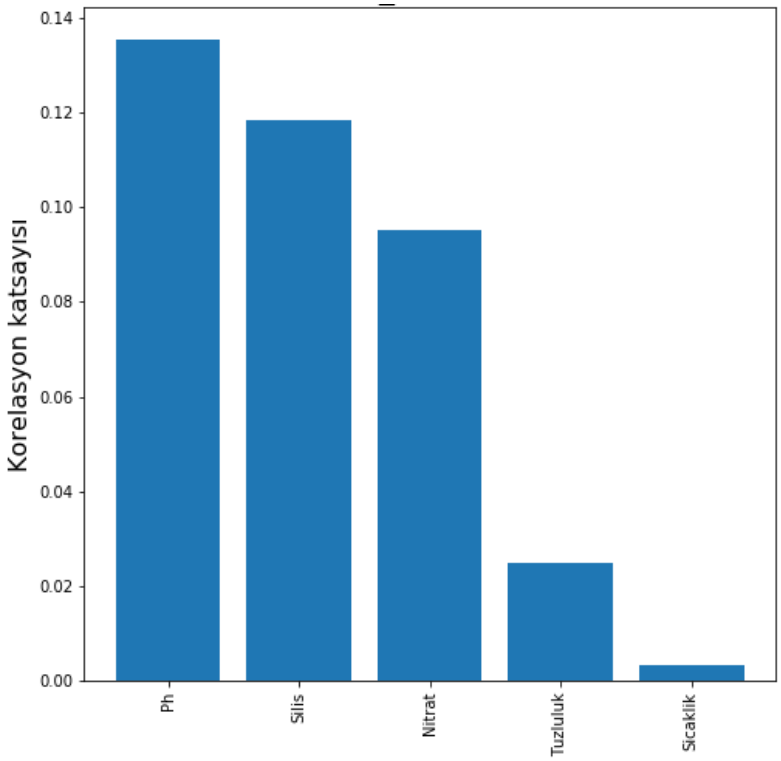
	Eğitim Doğruluk	Test Doğruluk	Duyarlılık	Kesinlik	f1_skor
Random Forest	1	0,822368	0,828947	0,817039	0,822228
Decision Tree	1	0,809211	0,837719	0,793037	0,814473
Gradient Boosting Classifier	0,975877	0,809211	0,824561	0,798678	0,810921
Gaussian Process	0,910819	0,769737	0,850877	0,732995	0,786868
Linear SVM	0,876462	0,765351	0,776316	0,762092	0,767854
Nearest Neighbors	0,833333	0,739035	0,842105	0,698707	0,762969
AdaBoost	0,859649	0,730263	0,732456	0,728464	0,730249
Neural Net	0,679094	0,638158	0,675439	0,628901	0,650662
Logistic Regression	0,572368	0,561404	0,644737	0,549685	0,592566
QDA	0,553363	0,541667	0,197368	0,742424	0,196306
Naive Bayes	0,540205	0,515351	0,188596	0,555556	0,180326

Tablo 4.6 Toksik fitoplankton varlığı tahmini için SMOTE ve Çapraz Geçerleme (Cross Validation) sonucu elde edilen değerlendirme ölçütleri

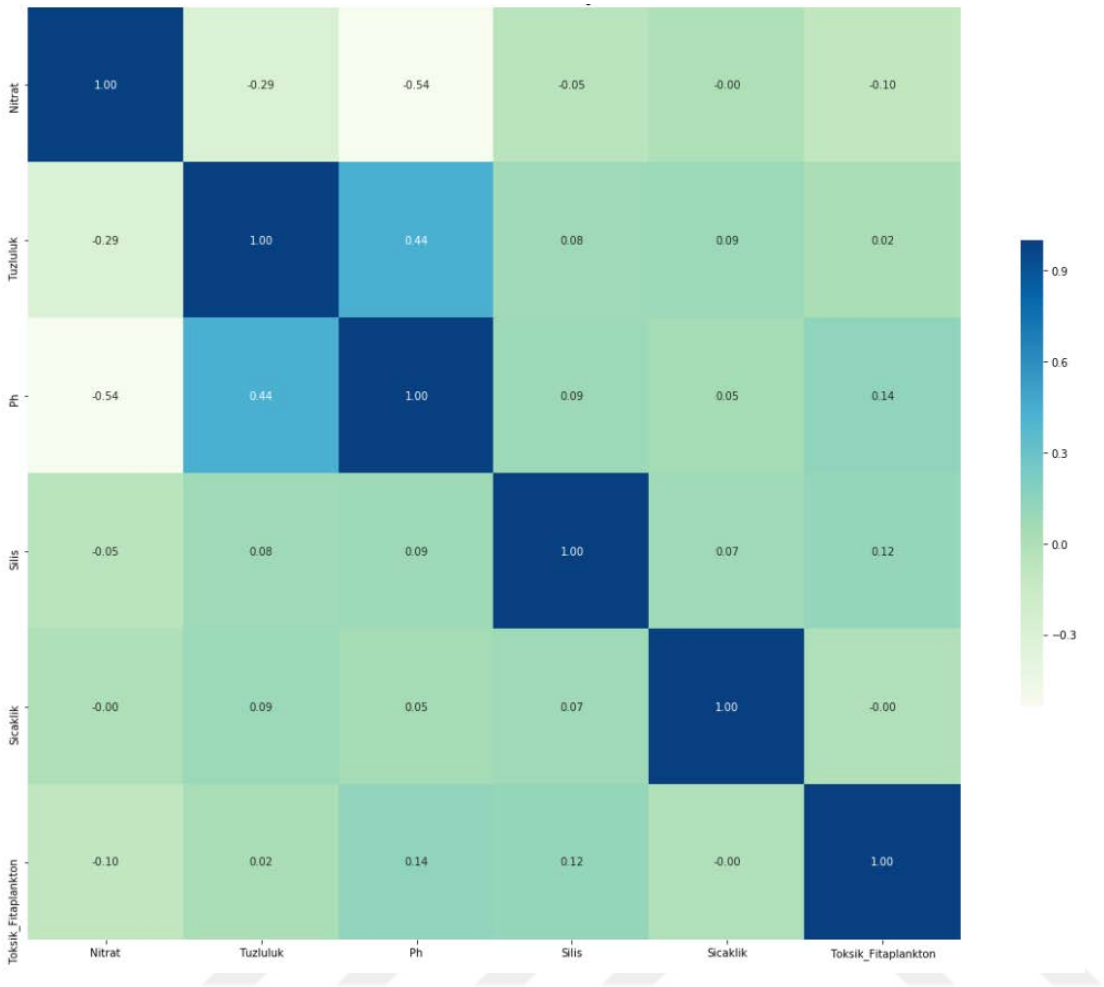
4.6 Toksik Fitoplankton ile Deniz Suyu Parametreleri Arasındaki Korelasyon Bulguları

4.6.1 SMOTE Öncesi Korelasyon Bulguları

SMOTE uygulanmadan önce elde olan veri kümesinde, deniz suyu parametreleri ve toksik fitoplankton (toksik algler) korelasyonun tespit edilmesi için korelasyon grafiği ve korelasyon matrisi oluşturulmuştur. Bunun sonucunda Ph parametersi, toksik fitoplankton ile korelasyonu en yüksek öznitelik olarak ortaya çıkmıştır. Şekil 4.8'de deniz suyu özniteliklerinin toksik fitoplankton ile korelasyon grafiği, Şekil 4.9'da ise tüm deniz suyu parametrelerinin birbirleriyle ilişkisini ifade eden korelasyon matrisi gösterilmiştir.

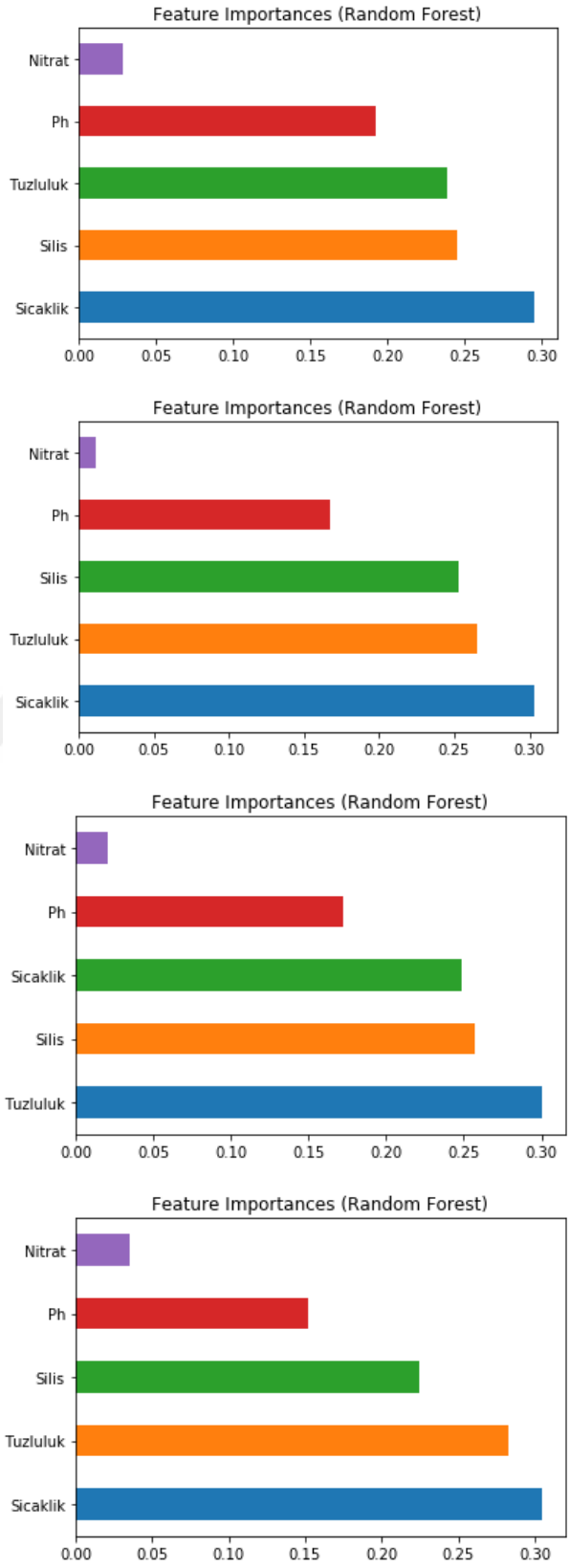


Şekil 4.8 SMOTE öncesi deniz suyu özniteliklerinin toksik fitoplankton ile korelasyonu



Şekil 4.9 SMOTE öncesi tüm deniz suyu parametrelerinin korelasyon matrisi

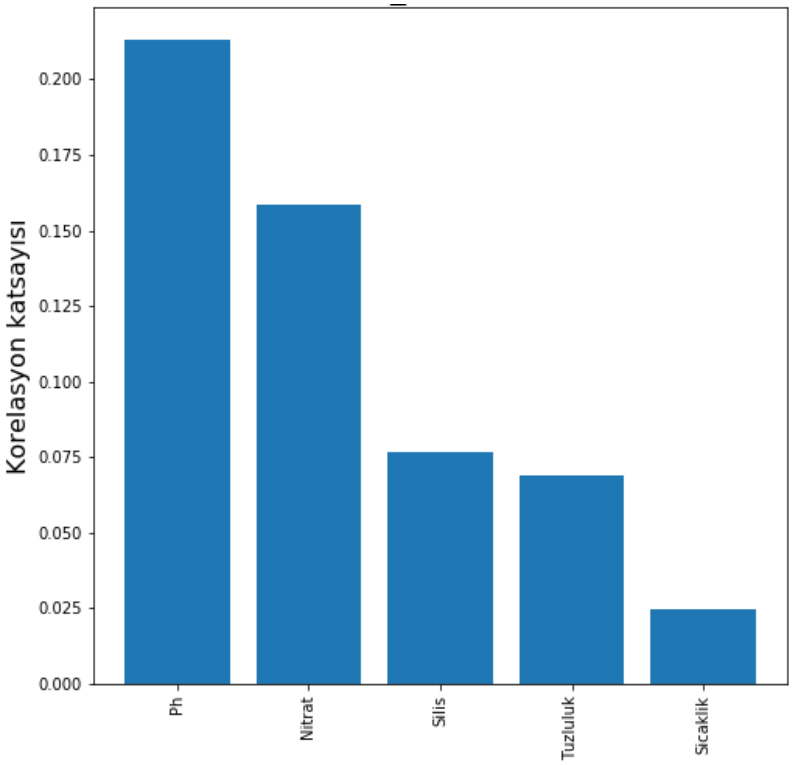
"Feature importance" metodu ile tahmin sonucuna etki eden önemli özniteliklerin belirlenmesine çalışılmıştır. Bunun için random forest makine öğrenmesi ve 4 katlı çapraz geçerleme yönteminden yararlanılmıştır. Bunun sonucunda sıcaklık, toksik fitoplankton varlığının tahmininde veri setindeki öznitelikler arasında en önemli öznitelik olarak bulunmuştur. Şekil 4.10' daki grafikte, 4 katlı çapraz geçerlemedeki her bir kat, tüm özniteliklerin önem oranları ile gösterilmiştir .



Şekil 4.10 Toksik fitoplankton varlığı tahmininde SMOTE öncesi Random forest ile “feature importance” sonuçları

4.6.2 SMOTE Sonrası Korelasyon Bulguları

SMOTE uygulandıktan sonra elde edilen veri kümesinde, deniz suyu parametreleri ve toksik fitoplankton (toksik algler) korelasyonun tespit edilmesi için korelasyon grafiği ve korelasyon matrisi oluşturulmuştur. Bunun sonucunda Ph parametersi, toksik fitoplankton ile korelasyonu en yüksek öznitelik olarak ortaya çıkmıştır. Şekil 4.11'de özniteliklerin toksik fitoplankton ile korelasyon grafiği, Şekil 4.12'da ise tüm özniteliklerin birbirleriyle ilişkisini ifade eden korelasyon matrisi gösterilmiştir.

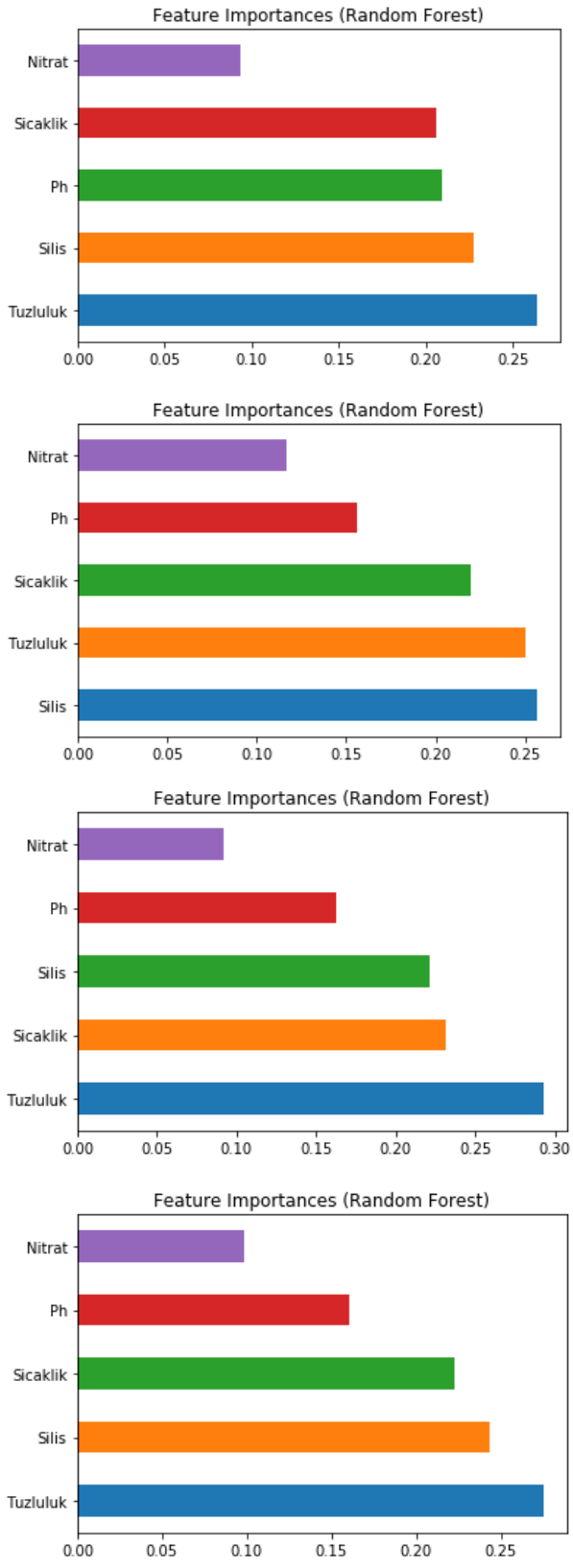


Şekil 4.11 SMOTE sonrası deniz suyu özniteliklerinin toksik fitoplankton ile korelasyonu



Şekil 4.12 SMOTE sonrası deniz suyu parametrelerinin arasındaki korelasyon matrisi

"Feature importance" ile sonuca etki eden önemli özniteliklerin belirlenmesine çalışılmıştır. Bunun için random forest makine öğrenmesi ve 4 katlı çapraz geçerleme yönteminden yararlanılmıştır. Sonucunda her katta farklı sonuçlar hakim olmak üzere, veri setindeki öznitelikler arasında tuzluluk, sıcaklık ve silis önemli öznitelikler olarak bulunmuştur. Şekil 4.13'daki grafikte, 4 katlı çapraz geçerlemedeki her bir kat, tüm özniteliklerin önem oranları ile gösterilmiştir.



Şekil 4.13 Toksik fitoplankton varlığı tahmininde SMOTE sonrası Random forest ile “feature importance” sonuçları

5. SONUÇ VE TARTIŞMA

Bu tezde gerçek, kısıtlı ve hatta dengesiz veri kümesi ile çalışmanın zorlukları ve bunlara ait çözüm yolları anlatılmıştır. Kısıtlı veriyi çoğaltmak için dengesiz veri setinde sayı olarak az olan tarafı SMOTE yöntemi ile arttırmak iki problemi de bir nebze gidermiştir. Dengesizliği tam olarak ortadan kaldırsa da verinin kısıtlı olma durumu tam anlamıyla değişmemiştir. Veri kümesi dengeli bir şekle geldikten sonra tezin temel amaçlarından biri olan biyotoksin varlığının toksik fitoplanktonlarla var olduğu düşünülen ilişkisi, korelasyon tablosu ve "feature importance" ile ortaya konulmuştur. Su parametreleri (Sıcaklık, Nitrat İyonları, Tuzluluk, Ph, Silis, Toksik Fitoplankton) ve midye üzerinden alınan örneğin analizi sonucunda oluşan parametreler (E. Coli, Civa, Kadmiyum, Kurşun, Biyotoksin varlığı) ile biyotoksin varlığı arasında olan korelasyonda en yüksek oranla toksik fitoplankton, ikinci sırada ise Nitrat parametresi gözlenmekte olup "feature importance" metodunda ikinci sırada E.Coli parametresi gözlenmektedir. Toksik Fitoplankton için SMOTE öncesi korelasyon bulguları ayrıca incelenmiştir. SMOTE öncesi korelasyon sonucunda Ph ilk sırada gözükmektedir. Fakat makine öğrenmesi yöntemlerinden random forest ile "feature importance" metodu uygulandığında, toksik fitoplankton varlığı tahmininde en etken faktör sıcaklık olarak çıkmıştır. SMOTE sonrası bulunan korelasyon sonuçlarında Ph birinci sırada gelirken ikinci sırada nitrat parametresi vardır. "Feature importance" metodunda ise birinci sırada toksik fitoplankton ile ilişkili olarak tuzluluk parametresi gözükmektedir. Tuzluluğun arkasından ise sıcaklık ve silis parametreleri gelmektedir. Bu çalışma bu konuda yapılan ilk çalışma olması sebebiyle bulgular kıyaslanamamıştır.

Veri kümesi kısıtlı olduğu durumlarda, kullanılan makine öğrenmesi yöntemlerinin doğruluğunu, tahminin gerçekliğini arttırmak için başvurulan bir diğer çözüm yöntemi çapraz geçerlemeden faydalanılmıştır. Veri kümesi elverişli hale geldiğinde Biyotoksin ve toksik fitoplankton varlığının tahmini üzerinde yapılan çalışmalar daha anlamlı olmuştur. Bu çalışmada 11 farklı makine öğrenmesi yöntemi kullanılmış, en iyi sonucu Random Forest vermiştir. "Grid search" yöntemi uygulandığında, Random Forest için verilen varsayılan parametreler ile hemen hemen aynı doğruluk oranında sonuçlar çıkmıştır. Biyotoksin varlığı tahmini için doğruluk 0,98, kesinlik 0,98, duyarlılık 0,97 ve f1-skor 0,98 olarak çıkmıştır. Toksik fitoplankton varlığı tahmini için ise doğruluk 0,82, kesinlik 0,81, duyarlılık 0,82 ve f1-skor 0,82 olarak çıkmıştır. Fakat bu sonuçların yüksek doğruluk oranı taşıması, tek bir test/eğitim veri setiyle sağlanması nedeniyle doğruluk ve gerçekçilik anlamında soru işaretleri oluşturmaktadır. Bunu gidermek adına, çapraz geçerleme denenmiş, sonuçlar çapraz geçerleme uygulaması sonrasında da değişmemiştir.

Bu çalışmanın sonucunda biyotoksin ve toksik fitoplankton varlığının önceden tahmin modeli daha da geliştirildiğinde ve saha çalışmalarıyla karşılıklı çıktılar test edildiğinde, zaman ve harcamalardan tasarruf edileceği anlaşılmaktadır. Bu tür çalışmalar yaygınlaştığı ve uygulamaya konulmaya başlandığında midye yetiştiricilik çiftliklerindeki önceden tespit edilen, oluşacağı düşünülen biyotoksin ve toksik fitoplankton varlığına erken müdahale ile kapatma risklerine karşı önlemler alınabilinecektir.

Ülkemizin üç tarafının denizlerle çevrili olması ve su ürünlerinde giderek artan yetiştiricilik sektörünün gelişimi göz önünde bulundurulduğunda bu çalışmanın önemi daha da artmaktadır. Büyüyen midye üreticiliği sektörüne, bu çalışma sonucu öngörülen tasarruflar büyük katkılar sağlayacaktır.

Korelasyon ve “feature importance” sonucunda toksik fitoplankton, ph ve E.coli, biyotoksin üzerindeki en önemli etkenler olarak gözükmektedir. Korelasyon ve “feature importance” sonuçlarına göre biyotoksin için etken gözüken parametreler, bu konudaki çalışmalara yön verebilir. Bu parametreler üzerinde yoğunlaşmak daha etkin bir sonuca götürecektir. Aynı şekilde toksik fitoplankton üzerindeki önemli gözüken parametreler üzerinde çalışmalar yapmak, toksik fitoplankton varlığını tahmin etme çalışmalarına zemin oluşturabilir. Ayrıca deniz suyunda yapılan analizlerin çeşitliliği arttırılıp ve uzun yıllar düzenli bir veri ile makine öğrenmesi metotları çalışıldığında toksik fitoplankton varlığı önceden tahmin sistemi, erken alarm sistemi geliştirilebilir.

Çalışma sonucunda Random Forest yöntemi biyotoksin ve toksik fitoplankton varlığı tahmini üzerinde birçok değerlendirme ölçütü sonucunda yüksek doğruluk oranı vermiştir. Mevcut olan veriler kısıtlı veriler olduğu için tahmin modeli geliştirmede kullanılmamış başka bir veri kümesi ile denenme şansı olmamıştır. Bu sebeple gelecek çalışmalarda yeni biriken, kayıt altına alınan veriler üzerinde denemeler yapılması düşünülmektedir.

KAYNAKLAR DİZİNİ

Ackman, R.G., 1988, Concerns for utilisation of marine lipids and oils, *Food Technology*, 151-160 s..

Altini M., 2015 "Dealing with Imbalanced Data: Undersampling, Oversampling and Proper Cross-Validation" https://www.marcoaltini.com/blog/dealing-with-imbalanced-data-undersampling-oversampling-and-proper-cross-validation/ (Erişim tarihi 19 Temmuz 2019)

Alpaydın, E., 2009, Introduction to machine learning. *MIT press*.

Anaconda https://www.anaconda.com/ (Erişim tarihi 19 Temmuz 2019)

Arat, B. 2018 "Model Performans Ölçütleri-1" https://berkarat.com/performans-degerlendirme/ (Erişim tarihi 19 Temmuz 2019)

Bilgin, M. and Uluturhan-Suzer E. 2015, Assessment of heavy metal accumulation in Mytilus galloprovincialis and Tapes decussatus (Bivalvia) distributed in the Homa Lagoon (Izmir Bay). *Ege Journal of Fisheries and Aquatic Sciences* 32.1: 1-8.

Chawla, Nitesh V., 2002, et al. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321-357.

Çolakoğlu, S. ve Çolakoğlu, F. A., 2013, Çift kabuklu yumuşakçaların ekonomik açıdan durumu. *Gıda Dergisi* 73

D'este, C., Rahman A. and Turnbull A., 2012, Predicting shellfish farm closures with class balancing methods. In: Australasian Joint Conference on Artificial Intelligence. *Springer, Berlin, Heidelberg*, 2012. p. 39-48.

FAO, Statistics, http://www.fao.org/statistics/en/, (Erişim tarihi: 24.10.2019)

He, H. and Garcia, E. A., 2008, Learning from imbalanced data. *IEEE Transactions on Knowledge & Data Engineering*, 9: 1263-1284.

Joseph, R. 2018 "Grid Search for model tuning" https://towardsdatascience.com/grid-search-for-model-tuning-3319b259367e/ (Erişim tarihi 19 Temmuz 2019)

Jupyter https://jupyter.org/ (Erişim tarihi 19 Temmuz 2019)

Kartal, E. ve Özen Z., 2017, Edition: 1, Chapter: 8, Publisher: Sakarya Üniversitesi Kütüphanesi Yayınevi, *Mühendislikte Yapay Zeka ve Uygulamaları*, pp.109-131

KAYNAKLAR DİZİNİ (devam)

Kinsella, J.E., 1988, Fish and seafoods: nutritional ımplications and quality ıssues, *Food Technology*, 42(5):146-150 p.

Lerman, P. M., 1980, Fitting segmented regression models by grid search. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(1), 77-84

Lewinson E. 2019 "Explaining Feature Importance by example of a Random Forest" https://towardsdatascience.com/explaining-feature-importance-by-example-of-a-random-forest-d9166011959e/ (Erişim tarihi 19 Temmuz 2019)

Mol, S. 2006 "Fayda ve Riskleriyle Midye" http://www.dunyagida.com.tr/haber/fayda-ve-riskleriyle-midye/1973/ (Erişim tarihi 19 Temmuz 2019)

Nar, K. 2018 "Cross Validation (Çapraz Doğrulama) Nedir?" https://kadirnar.com/post/cross-validationcapraz-dogrulama-nedir-veri-on-isleme/49/ (Erişim tarihi 19 Temmuz 2019)

Powers, D. M., 2011, Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

Python http://www.python.org/ (Erişim tarihi 19 Temmuz 2019)

Rahman, Ashfaqur; D'este, Claire., 2015, Shellfish farm closure prediction and cause identification using machine learning methods. *Computers and Electronics in Agriculture*, 110: 241-248.

Shaikh, R. 2018 "Feature Selection Techniques in Machine Learning with Python" https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e/ (Erişim tarihi 19 Temmuz 2019)

Shang, K. P. 2015. "Accuracy, Precision, Recall", https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9/ (Erişim tarihi 19 Temmuz 2019)

StratifiedKFold https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html/ (Erişim tarihi 19 Temmuz 2019)

Subasinghe, R., Soto, D. and Jia, J. 2009, Global aquaculture an ditsrole in sustainable development, Aquaculture 1: 2-9pp.

KAYNAKLAR DİZİNİ (devam)

Varlık, B., 1991, *Investigation of effects of someheavy metal (Cd-Pb) to the different development stages of Mytilus galloprovincialis*, Dokuz Eylül Üniversitesi Fen Bilimleri Enstitüsü, İzmir.

TEŞEKKÜR

Bu çalışma süresince gerekli verilerin sağlanmasında kolaylık gösteren İzmir Tarım ve Orman İl Müdürlüğü çalışanlarına, özellikle kıymetli görüşlerinden yararlandığım ve yakın ilgisini esirgemeyen Sayın Levent ÖZKAN'a, tezin biçimlendirmesinde katkıda bulunan Arş. Gör. Erkan YAŞAR 'a, desteklerini hiç esirgemeyen Koordinatörüm Yük. Müh. Erdinç VESKE 'ye ve her zaman yanımda olan aileme teşekkürü bir borç bilirim.

.... / .... / 2019

İmzası

Alper MEMİŞ

ÖZGEÇMİŞ

Alper MEMİŞ

İLETİŞİM BİLGİLERİ

- **Tarımsal Araştırmalar ve Politikalar Genel Müdürlüğü, Hayvancılık ve Su Ürünleri Araştırmaları Dairesi Başkanlığı,**
- **Üniversiteler Mah. Dumlupınar Bulvarı, Eskişehir Yolu 10. Km, Çankaya/ANKARA 06800.**
- **Telefon: +90 (506) 2177290**
- **E-posta: alper.memis@tarimorman.gov.tr**

EĞİTİM

- **Lise, Mehmetçik Anadolu Lisesi**
- **Lisans, Kırıkkale Üniveristesi, Bilgisayar Mühendisliği Bölümü, 2014.**
- **Yüksek Lisans, Ege Üniversitesi, Bilgisayar Mühendisliği Bölümü, 2016-devam ediyor.**

DENEYİM

- **TUSAŞ(TAI) - Türk Havacılık ve Uzay Sanayii A.Ş. – Stajyer**
- **ASELSAN - Stajyer**
- **İzmir Tarım ve Orman İl Müdürlüğü**
- **Tarımsal Araştırmalar ve Politikalar Genel Müdürlüğü**