SELÇUK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

SELÇUK
ÜNİVERSİTESİ

T.C.

SELÇUK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

DENGESİZ BAL PETEĞİ VERİ SETİNDE SINIFLANDIRMA PERFORMANSININ ANALİZİ

Serkan ÖZGÜN

YÜKSEK LİSANS TEZİ

Bilgisayar Mühendisliği Anabilim Dalı

Eylül-2022
KONYA

TEZ KABUL VE ONAYI

Serkan ÖZGÜN tarafından hazırlanan "DENGESİZ BAL PETEĞİ VERİ SETİNDE SINIFLANDIRMA PERFORMANSININ ANALİZİ" adlı tez çalışması 09/09/2022 tarihinde aşağıdaki jüri tarafından oy birliği / ~~oy çokluğu~~ ile Selçuk Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Jüri Üyeleri	**İmza**
Başkan Doç.Dr. Sait Ali UYMAZ	……..
Danışman Doç.Dr. Mehmet Akif ŞAHMAN	……..
Üye Dr.Öğr.Üyesi Selahattin ALAN	……..

Yukarıdaki sonucu onaylarım.

Prof. Dr. Sait GEZGİN
FBE Müdürü

TEZ BİLDİRİMİ

Bu tezdeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edildiğini ve tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

DECLARATION PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

İmza

Serkan ÖZGÜN

Tarih: 09/09/2022

ÖZET

YÜKSEK LİSANS TEZİ

DENGESİZ BAL PETEĞİ VERİ SETİNDE SINIFLANDIRMA PERFORMANSININ ANALİZİ

Serkan ÖZGÜN

Selçuk Üniversitesi Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Doç. Dr. Mehmet Akif ŞAHMAN

2022, 56 Sayfa

Jüri
Doç.Dr. Mehmet Akif ŞAHMAN
Doç.Dr. Sait Ali UYMAZ
Dr.Öğr.Üyesi Selahattin ALAN

Arıcılık faaliyetleri Türkiye ve Dünya için önemli bir tarımsal faaliyettir. Arıcılık, Türkiye' deki kırsal kesimlerin kalkınmasına katkısı nedeniyle sosyo-ekonomik anlamda önem arz etmektedir. Ayrıca arıcılık faaliyetleri sonucunda üretilen ürünler insanlar için önemli besin kaynaklarıdır. Bu nedenle arıcılık faaliyetlerinde doğru yöntemlerin kullanılması arıcılık faaliyetlerinin sürdürülebilirliği için önemlidir. Üreticiler tarafından bilinçsiz ve gerekli teknikler kullanılmadan gerçekleştirilen arıcılık faaliyetleri, elde edilecek ürünlerin kalitesini ve verimini negatif yönde etkilemektedir. Bal, arıcılık faaliyetleri sonucunda elde edilen en önemli çıktılardan birisidir. Bal üretim sürecinde birçok aşama yer almaktadır. Bu aşamalardan biri de bal hasadı aşamasıdır. Bal hasadı aşamasında doğru yöntem ve tekniklerden faydalanılması üretilen bal miktarını ve kalitesini arttırmaktadır. Ayrıca bilinçli arıcılık faaliyetleri yersiz yavru arı kayıplarından kaçınılarak, arı kolonisi varlığının korunmasında da etkilidir. Bu tez çalışmasında, bal hasadındaki yavru arı kayıplarını azaltmak için bal peteği üzerinde 'kapalı larva hücrelerinin' tespiti bir sınıflandırma problemi olarak ele alınmıştır. Çalışmada 38 adet bal peteği görüntüsünden faydalanılarak veriseti oluşturulmuştur. Verisetinde kapalı larva hücreleri ve diğerleri olmak üzere iki sınıf için etiketle yapılmıştır. Veri setindeki etiketlenmiş iki sınıfa ait veri oranının yaklaşık 1/5 olduğu görülmüştür. Sınıflar arasındaki dengesizliğin giderilerek sınıflandırma başarısını arttırmak istenmiştir. Bunun için literatürde iyi bilinen ve güncel beş farklı veri düzeyinde aşırı örnekleme (SMOTE, Borderline-SMOTE1, Borderline-SMOTE2, Safe-Level-SMOTE ve DEBOHID) yaklaşımdan faydalanılmıştır. Dengelenmiş veriler üzerindeki sınıflandırma başarısını göstermek için üç farklı sınıflandırıcıdan (K- En Yakın Komşu (kNN), Karar Ağacı(KA) ve Destek Vektör Makineleri (DVM)) faydalanmıştır. Sınıflandırma sonuçları F1-Skor, G-Ortalama ve AUC metrikleri ile değerlendirilmiştir. Sınıflandırma işlemleri sonucunda sentetik veri üretme yöntemleri ile dengeli hale getirilen veri setlerinde sınıflandırma başarısının arttığı görülmüştür.

Anahtar kelimeler: Bal peteği sınıflandırma, sentetik veri üretimi, kapalı larva tespiti, sınıflandırma

ABSTRACT

MS THESIS

ANALYSIS OF CLASSIFICATION PERFORMANCE ON IMBALANCED HONEYCOMB DATASET

Serkan ÖZGÜN

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCE OF SELÇUK UNIVERSITY
THE DEGREE OF MASTER OF SCIENCE IN COMPUTER ENGINEERING

Advisor: Assoc. Prof. Dr. Mehmet Akif ŞAHMAN

2022, 56 Pages

Jury
Assoc.Prof.Dr. Mehmet Akif ŞAHMAN
Assoc.Prof.Dr. Sait Ali UYMAZ
Asst.Prof.Dr. Selahattin ALAN

Beekeeping activities are an important agricultural activity for Turkey and the World. Beekeeping is important in socio-economic terms due to its contribution to the development of rural areas in Turkey. In addition, the products produced as a result of beekeeping activities are important food sources for humans. For this reason, using the right methods in beekeeping activities is important for the sustainability of beekeeping activities. Beekeeping activities carried out by the producers unconsciously and without using the necessary techniques negatively affect the quality and yield of the products to be obtained. Honey is one of the most important outputs obtained as a result of beekeeping activities. There are many stages in the honey production process. One of these stages is the honey harvest stage. Utilizing the right methods and techniques during honey harvesting increases the amount and quality of honey produced. In addition, conscious beekeeping activities are also effective in preserving the existence of the bee colony by avoiding unnecessary baby bee losses. In this thesis, the detection of 'closed larval cells' on the honeycomb is considered as a classification problem in order to reduce the loss of baby bees in honey harvest. In the study, a dataset was created by using 38 honeycomb images. The dataset was constructed with labels for two classes, closed larval cells and others. It was seen that the data ratio of the two labeled classes in the data set was 1/5. It was aimed to increase the classification success by eliminating the imbalance between the classes. For this, five different data-level oversampling approaches (SMOTE, Borderline-SMOTE1, Borderline-SMOTE2, Safe-Level-SMOTE and DEBOHID) that are well-known and current in the literature were used. Three different classifiers (K-Nearest Neighbor (kNN), Decision Tree and Support Vector Machines (DVM)) were used to show the classification success on balanced data. Classification results were evaluated with F1-Score, G-Mean and AUC metrics. As a result of the classification processes, it was observed that the classification success increased in the data sets that were balanced with synthetic data generation methods.

Keywords: Honeycomb classification, synthetic data production, closed larva detection, classification

ÖNSÖZ

Tez çalışmamın konusunun belirlenmesinde ve çalışmamın her aşamasında yardımlarını esirgemeyen, pozitif eleştirileri ile çalışmama yön veren, uzun süren tez çalışmam boyunca büyük sabır gösteren değerli danışman hocam Doç.Dr. Mehmet Akif ŞAHMAN' a teşekkür eder ve saygılarımı sunarım.

Tez çalışmam boyunca gösterdikleri sabır ve destekleri ile motivasyon kaynağı olan eşim Rabia ÖZGÜN, kızlarım Ezgi Dilda ÖZGÜN ve Eylül Aze ÖZGÜN' e, bu noktaya gelmemi sağlayan annem Aliye ÖZGÜN ve babam Ahmet ÖZGÜN' e teşekkür ederim.

Serkan ÖZGÜN
KONYA-2022

İÇİNDEKİLER

KISALTMALAR

AUC	**:**	Eğri Altında Kalan Alan (Area Under The Curve)
CUST	**:**	Küme Alt Örnekleme Tekniği (Cluster Undersampling Technique)
DE	**:**	Diferansiyel evrim
DEBOHID	**:**	Yüksek Düzeyde Dengesiz Veri Kümeleri İçin Bir Diferansiyel Evrim Tabanlı Aşırı Örnekleme Yaklaşımı (A Differantial Evolution Based Oversampling Appoach For Highly Imbalanced Datasets)
KA	**:**	Karar Ağacı (Decision Tree)
DVM	**:**	Destek Vektör Makineleri (Support Vector Machine)
FN	**:**	Yanlış Negatif (False Negative)
FP	**:**	Yanlış Pozitif (False Positive)
G-Mean	**:**	G-Ortalama
kNN	**:**	K-En Yakın Komşu (K-Nearest Neighbors)
LEM2	**:**	Örnek Modülden Öğrenme (Learning From Exampling Module)
OSS	**:**	Tek Taraflı Seçim (One Side Selection)
PSO	**:**	Parçacık Sürü Optimizasyonu (Particle Swarm Optimisation)
RBF	**:**	Radyal Temel Fonksiyon (Radial Basis Function)
ROS	**:**	Rastgele Aşırı Örnekleme (Random Over-Sampling)
RUS	**:**	Rastgele Az Örnekleme (Random Under-Sampling)
SMOTE	**:**	Sentetik Azınlık Aşırı Örnekleme Tekniği (Synthetic Minority Over-sampling Technique)
TN	**:**	Doğru Negatif (True Negative)
TP	**:**	Doğru Pozitif (True Positive)
TPR	**:**	Doğru Pozitif Oranı (True Positive Rate)
YSA	**:**	Yapay Sinir Ağları

1. GİRİŞ

Arıcılık tüm dünya için önemli bir tarımsal faaliyet ve gelir kaynağı olduğu gibi ülkemizde de önemli bir yere sahiptir (Semerci, 2017). Türkiye zengin florası, geniş coğrafyası ve koloni varlığı ile dünya arıcılığında önde gelen ülkelerden birisidir (Karlıdağ ve Köseman, 2015). Türkiye'nin arıcılık faaliyetlerinde iyi performans sergilemesini gerektiren bu avantajlara ve son zamanlarda göstermiş olduğu olumlu gelişmelere rağmen bal üretiminde ve ticaretinde beklenilen gelişmeyi göstermediği görülmektedir (Şeker ve ark., 2017). Türkiye' de arıcılık sektörü yıllar içinde sürekli gelişim göstermektedir. Yıllara bağlı olarak kovan sayısı sürekli artış gösterse de bu artış üretime tam olarak yansımamıştır. Bu durumun ortaya çıkmasında iklim değişikliği gibi küresel faktörler ve bal üretimi sürecinde yanlış yöntemlerin kullanılması gibi yerel sorunlar etkilidir. Bu sorunlar kovan başına verimde düşüşe sebep olmaktadır (Burucu ve Gülse Bal, 2017). Uygun yöntemlerin uygulanması sonucunda arıcılığın Türkiye'nin kırsal bölgeler için önemli bir ekonomik kaynak olacağı ve sürekli bir üretim faaliyeti olacağı düşünülmektedir (Karlıdağ ve Köseman, 2015).

Gıda ve Tarım Örgütü' nün (FAO) 2016 yılı verilerine göre Türkiye, %5,6'lık oranla dünya bal üretiminde %28,1'lik orana sahip Çin' den sonra ikinci sıradadır. Türkiye'yi %4,5'lik oranla İran takip etmektedir. Bal üretiminde önemli bir noktada olan Türkiye kovan varlığında ise 3. sırada yer almaktadır. (Burucu ve Gülse Bal, 2018). Türkiye' de bu sayısal verilere rağmen koloni başına bal verimi 14,3 kg olarak gerçekleşmiştir. Bu değer dünya ortalamasından %32 düşüktür. Çin' de bu koloni başına bal verimi 50 kg' nin üzerindedir (Semerci, 2017). Türkiye'nin coğrafi konumu, nektar kaynağı bakımından sahip olduğu çeşitlilik, bölgeler arasındaki topografik farklılıklardan kaynaklanan bölgeden bölgeye değişen yılın farklı dönemlerindeki çiçeklenme, arı gen kaynakları bakımından sahip olduğu zenginlik gibi avantajları göz önünde bulundurulduğunda arıcılıkta yüksek bir performans göstermesi beklenir. Ancak tüm bu avantajlara rağmen sayısal veriler değerlendirildiğinde beklenen performansın oluşmadığı görülmektedir (Kekeçoğlu ve ark., 2007).

Bal üretimi sürecinde karşılaşılan problemlerden biri yavru arı kayıpları veya yavru arı bulundurması sebebiyle hasat edilmeyen petekler dolayısıyla ortaya çıkan verim kaybıdır. Bal petekleri üzerinde yavru arılara ait hücrelerde bulunmaktadır. Mevcut yöntemlerle yapılan hasat işlemlerinde balın süzme evresinde yavru arılar telef olmaktadır bu nedenle yavru arı bulunduran petekler hasat edilememektedir. Arıcıların

üretilen bal miktarını arttırmak amacıyla yavrulardan vazgeçmesi ve içinde yavru hücrelerin bulunduğu petekleri süzmesi uzun vadede verim kaybına neden olmaktadır. Bununla birlikte yavru arılara ait hücreler bal süzme esnasında bala karışmaktadır. Dolayısıyla balın homojenliğini kaybolmaktadır. Arıcıların yavru arıları korumak amacıyla petekleri hasat etmemesi de verim kaybına sebep olmaktadır. Arıcıların yavrulardan vazgeçmesi ise koloninin zayıflamasına sebep olmaktadır. Zayıflayan koloni kendisini dış etkenlerden (yağmalama gibi) koruyamamaktadır. Bu sebeplerden dolayı hasat sırasında doğru petek seçimi ve doğru hasat yaklaşımı büyük önem taşımaktadır.

Makine öğrenmesi ve veri madenciliği kavramlarının ortaya çıkması ile sınıflandırma problemlerinde karşılaşılan sorunlardan biri de dengesiz sınıf dağılımına sahip veri setlerinde ortaya çıkan sınıflandırma başarısı problemidir. Eşit sınıf dağılımlarına sahip olmayan veri setleri üzerinde yapılan çalışmalarda çoğunluk sınıfı örneklerinin azınlık sınıf örneklerini bastırdığı görülmektedir. Bu durum makine öğrenmesinde yeni bir araştırma alanının gelişmesinin önünü açtı; dengesiz veri setlerinde öğrenme.

Dengesiz veri setleri, birçok gerçek dünya sınıflandırma probleminde ortaya çıkan bir sorundur. Sınıflandırma başarısı veri setlerindeki sınıfların dengeli temsil edilmesi ile doğrudan ilişkilidir. Sınıfları temsil eden verilerin örneklem sayısı, veri seti içindeki sınıf örnek sayıları aynı oranda temsil edilirse sınıflandırma başarısının yüksek olması beklenir (Kaya ve ark., 2021). Bu sorunun ortaya çıkmasında farklı etkenler bulunmaktadır ve probleme göre farklı nedenlerden kaynaklanan dengesiz veri sorunları görülmektedir. Veri toplamadaki maliyet sorunları, veri toplama sürecindeki mahremiyet, örneklem sayısının az olması dengesiz veri seti problemlerinin görülmesinde başlıca sebeplerden bazılarıdır (Sun ve ark., 2009). Tıbbi teşhisler (Mazurowski ve ark., 2008), kredi kartı dolandırıcılığı tespiti (Zareapoor ve Yang, 2017), metin sınıflandırması problemleri (Li ve ark., 2010), yazılım hatası tahmini (Wang ve Yao, 2013), üretim tesislerindeki kusurlu ürünlerin tespiti (Cateni ve ark., 2014), bilgisayar ağlarına izinsiz girişimlerin tespiti (Cieslak ve ark., 2006), okyanus yüzeyinde petrol sızıntılarının tespiti (Kubat ve ark., 1998) gibi alanlarda dengesiz veri seti problemleri görülmüştür.

Tez çalışmasında, bir bal peteğinde kapalı larva gözlerinin tespiti bir sınıflandırma problemi olarak ele alınmıştır. Bir bal peteğinde farklı hücreler bulunabilmektedir. Bu hücreler; içi boş hücreler, içi bal dolu açık hücreler, içi bal dolu kapalı(sırlı) hücreler, içinde polen olan açık hücreler, içinde larva olan açık hücreler, içinde larva olan kapalı hücreler ve diğerleri (arı, bal peteği çıtası vb.) şeklinde tanımlanabilir (Farahmand, 2022).

Bu tez çalışmasında kapalı larva hücreleri tespit edilmesi istenildiğinden, iki sınıf için etiketleme yapılmıştır. Bu etiketler kapalı larva hücreleri ve diğerleri olarak etiketlenmiştir. Veri seti oluşturulurken, kapalı larva hücrelerinin sayısının, diğer hücrelerin sayısına göre daha az örneklem sayısına sahip olduğu görülmektedir. Bu örneklem sayısındaki belirgin oransal dengesizlikten dolayı oluşturulan veri seti, dengesiz olarak ifade edilebilmektedir.

Tez çalışmasında kullanılan bal peteği görüntüleri 5x5 piksel olacak şekilde parçalanmış ve uygun şekilde etiketlenerek veri seti oluşturulmuştur. Oluşturulan veri seti önce temel sınıflandırıcılar kullanılarak sınıflandırılmıştır. Daha sonrasında ise literatürde iyi bilinen aşırı örnekleme yaklaşımları ile veri setindeki sınıf örnek sayıları dengelenmiştir. Dengelenen veri setleri aynı sınıflandırma yaklaşımları kullanılarak sınıflandırılmıştır. Orijinal (dengesiz) veri seti ile dengeli veri setinin sınıflandırma başarısı karşılaştırılmıştır. Yapılan detaylı analizler, sentetik veriler üretilerek dengelenmiş veri setindeki sınıflandırma başarısının arttırıldığını göstermektedir.

2. KAYNAK ARAŞTIRMASI

Bu bölümde bal petekleri ve dengesiz veri seti problemi ile ilgili literatürdeki çalışmalar incelenmiştir. Literatürde petek görüntüleri üzerinde yapılan çalışmalarla sık karşılaşılmasa da dengesiz veri setleri üzerinde yapılan birçok çalışma vardır.

Güngörmüş (Güngörmüş, 2020), arı sütü üretim sürecinde en meşakkatli işlemlerden biri olan larva transferinin hızlandırılması ve arı sütü üretimin arttırılabilmesi için bir çalışma gerçekleştirmiştir. Arı sütü üretimi için ideal boyutta olan larvaları tespit etmek gerekmektedir. Yapılan çalışmada görüntü işleme teknikleri kullanılarak larvaların konumu ve özellikleri tespit edilmiştir. Bunun için oluşturdukları düzenek ile 60 farklı petek fotoğrafı çekilmiştir. Fotoğrafların 40 tanesi eğitim için, 20 tanesi test için kullanılmıştır. Fotoğraflardaki larvalar etiketlenerek evrişimsel sinir ağı yöntemlerinden biri olan Faster R-CNN ile eğitilmiştir. Çalışma sonucunda larvaların konumu ve özellikleri başarıyla tespit edilmiştir.

Sparavigna (Sparavigna, 2016), işçi arıların petekteki hücrelerini işçi arıların hücrelerinden ayırt edebilmek için görüntü bölütleme yöntemi ile bir çalışma gerçekleştirmiştir. Önerilen bölütlemede siyah beyaz görüntüye dönüştürülen bal peteğinin orijinal görüntüsünün eşiklenmesine dayanmaktadır. Bu görüntüler her biri petek hücresi içeren "süper piksel" olarak bilinen birden çok piksel kümesine bölünmüştür. Süper pikseller etiketlenir ve her etiket hücrenin boyutuna karşılık gelir. Böylece her bir hücrenin boyutu kolayca ölçülmüş ve işçi arılara ait hücreler erkek arıların hücrelerinden ayırt edilmiştir.

Alves ve arkadaşları(Alves ve ark., 2020), yaptıkları çalışma ile Derin öğrenme yaklaşımlarını kullanarak bir uygulama geliştirmişlerdir. On üç farklı konvolüsyonel sinir ağı (CNN) mimarisi ile eğittikleri ve sınıflandırdıkları petek görüntülerini yumurta, larva, kapalı kuluçka, polen, nektar, bal ve diğerleri olarak sınıflandırmışlardır.

Farahmand (Farahmand, 2022), yedi ayrı sınıftan oluşan 103.451 eğitim ve 25.863 test görüntüsü içeren bal peteği veri seti üzerinde derin öğrenme yöntemleri kullanarak sonuçları karşılaştırmıştır. Çalışmada AlexNet, Vgg16, Vgg19 ve ResNet50 +Xception, Inception-v3 ve SSCNN derin öğrenme algoritmaları kullanılmıştır. F1-skor, kesinlik, duyarlılık ve Roc eğrisi değerlendirme metrikleri ile performansını ölçtüğü karşılaştırma sonucunda AlexNet (eğitim başarısı %95 ve doğrulama başarısı %94) diğer algoritmalara göre daha iyi sonuçlar vermiştir.

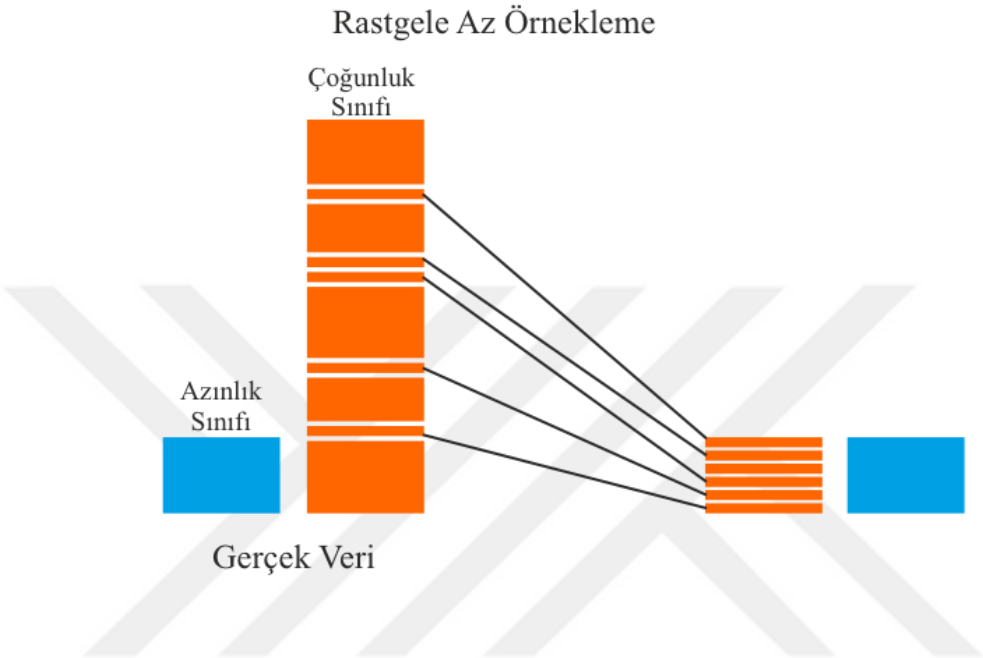
Dengesiz veri setleri gerçek dünya problemlerinde yaygın olarak karşılaşılan bir sorundur. Dengesiz veri seti problemi bir sınıfın diğer sınıf veya sınıflara göre örneklem sayısının çok olduğu durumlarda ortaya çıkar (Chawla ve ark., 2004). Dengesiz veri seti araştırma alanının temelleri öncelikle 2000' li yılların başlarında gerçekleştirilen "Amerikan Yapay Zeka Derneği Konferansı' da" (Japkowicz, 2000) atıldı. 2003 yılında ICML-KDD konferansıyla artık bu konu tek başına üzerine çalışmalar yapılan bir konu haline geldi (Fernández ve ark., 2018).

Makine öğrenmesi ve veri madenciliği alanlarının genişlemesi ve büyük veri kavramlarının hayatımıza girmesi ile yeni zorluklarla karşılaşırken dengesiz öğrenme hakkında yeni gelişmeler sağlanmıştır. Veri setindeki sınıfların dengesizlik problemi genellikle doğru sınıflandırmayı zorlaştırır. Bu sorunun etkilerini azaltmak için çeşitli yöntemler önerilmiştir. Chawla ve ark. (Chawla ve ark., 2004), sınıf dengesizliği problemini veri düzeyi ve algoritmik düzey olmak üzere iki farklı kategoriye ayırmışlardır. Veri düzeyi tekniklerinin de rastgele az örnekleme, rastgele aşırı örnekleme, yönlendirilmiş alt örnekleme ve yönlendirilmiş aşırı örnekleme olarak dört alt başlık içerdiğini belirtmişlerdir. Algoritmik düzeyde çözümlerin ise sınıfların maliyetini ayarlama, karar ağacı ile çalışırken ağaç yaprağındaki olasılık tahmini, karar eşiğini ayarlamak ve iki sınıf yerine bir sınıftan öğrenmeye dayalı yöntemler olarak ayrıldıklarını belirtmişlerdir (Japkowicz, 2000; Chawla ve ark., 2003). García ve Herrera (García ve Herrera, 2009), sınıf dengesizliği problemini çözmek için önerilen yaklaşımları çalışma biçimlerine göre algoritmik düzey, veri düzeyi ve bu iki düzeyi birleştiren üçüncü bir düzey olmak üzere üçe ayırıyor (Kaya ve ark., 2021). Chen ve ark.(Chen ve ark., 2021), bu yöntemleri veri düzeyi, algoritma düzeyi ve melez (hibrit) yöntemler olmak üzere üç başlıkta toplamıştır. Buna göre sınıf dengesizliği literatürde veri düzeyi, algoritmik düzey, maliyete duyarlı ve topluluklar (melez) olmak üzere dört başlıkta ele alınmıştır (Kaya ve ark., 2021).

2.1. Veri Düzeyi

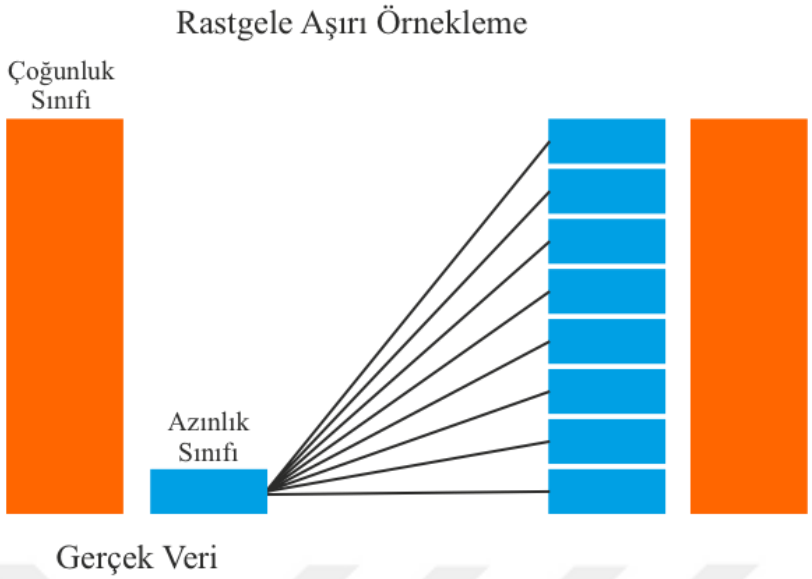
Genellikle standart sınıflandırma algoritmaları, veri setlerindeki sınıfların dengeli olduğunu varsayar ve bu kabul ile önerilir veya geliştirilirler (He ve Garcia, 2009). Veri düzeyinde veriseti dengeleme yöntemleri, eğitim veri setini standart bir öğrenme algoritmasına uygun hale getirmek için eğitim veri setindeki sınıf örneklerini aşırı (over-sampling) veya az (under-sampling) örnekleme yaklaşımları ile dengelemeye odaklanır (Krawczyk, 2016). Literatürde en çok kullanılan veri düzeyi yöntemleri 'Rastgele Az

Örnekleme' (Random Under-Sampling - RUS), 'Rastgele Aşırı Örnekleme' (Random Over-Sampling - ROS) ve 'Sentetik Veri Üretme' yöntemidir (Gümüştaş, 2019). Sentetik veri üretme, veri düzeyi yaklaşımının en yaygın kullanılan yöntemlerinden biridir (Pristyanto ve ark., 2018).



Şekil 2.1: Rastgele az örnekleme.

Rastgele az örnekleme, çoğunluk sınıftan rastgele seçilen bazı verilerin silinmesi ilkesine dayanır (Şekil 2.1). Bu şekilde iki veri arasında sınıflar arası denge sağlanmaya çalışılmaktadır. Bu yöntemin en büyük dezavantajı veri kaybıdır (Sağlam ve ark., 2021). Rastgele seçilen veriler arasında yararlı verilerin silinmesi istenmeyen bir sonuçtur (Tahir ve ark., 2012). Rastgele aşırı örneklemede ise azınlık sınıfı örnek sayısı çoğunluk sınıfı örnek sayısına eşitlenene kadar rastgele seçilen azınlık örneklerinin tekrar azınlık sınıfına eklenerek (Şekil 2.2) örnek sayısının artmasını sağlayan yöntemdir (Aydin Hakli, 2018). Bu durumda aşırı öğrenmeye sebep olabilmektedir ve bu durum modelin dezavantajı olarak görülebilir. (Çürükoğlu, 2019).



Şekil 2.2.: Rastgele aşırı örnekleme.

Chawla ve arkadaşları (Chawla ve ark., 2002), Sentetik Azınlık Aşırı Örnekleme Tekniği (SMOTE: Synthetic Minority Over-sampling Technique) olarak adlandırdıkları bir tekniği önermişlerdir. Bu yaklaşım azınlık sınıfı örneklemlerini sentetik olarak arttırma ve dengesiz veri setlerine çözüm bulma yaklaşımıdır (Aydilek, 2018). Bu yöntem ile rastgele örnekleme yönteminden farklı olarak mevcut veriler analiz edilerek bu verilere benzer yeni veriler oluşturulur (Pir, 2022).

SMOTE yönteminin önerilmesinden sonra bu yöntemi geliştirmeye yönelik yeni yaklaşımlar önerilmiştir. Bu çalışmalardan biri Han ve arkadaşları (Han ve ark., 2005) tarafından önerilmiştir. Bu çalışmada Borderline-SMOTE1 ve Borderline-SMOTE2 adında iki yöntem önerilmiştir. Mevcut SMOTE tekniğinden farklı olarak tüm azınlık sınıfı örneklerini göz önünde bulundurmak yerine, sadece sınır çizgisi adı verilen, sınıfların sınır bölgelerindeki verileri temsil eden veriler SMOTE tekniği ile çoğaltılmıştır (Bunkhumpornpat ve ark., 2009). Borderline-SMOTE1 yönteminde sınır çizgisindeki azınlık sınıfı örnekleri göz önünde bulundurulurken Borderline-SMOTE2 yönteminde farklı olarak azınlık sınıfı örnekleri ile beraber çoğunluk sınıfı örnekleri de göz önünde bulundurulmuştur (Kaya ve ark., 2021).

Bunkhumpornpat ve ark. (Bunkhumpornpat ve ark., 2009), Safe-Level-SMOTE adından yine SMOTE yöntemine dayanan yeni bir yöntem önermiştlerdir. kNN(K-En

Yakın Komşu - K-Nearest Neighbors) ile güvenli seviye olarak belirlenmiş bölgedeki aynı ağırlık değerine sahip azınlık verileri örneklenmiştir.

He ve arkadaşları (He ve ark., 2008), dengesiz veri seti problemi için ADASYN(Adaptive Synthetic Sampling Approach for Imbalanced Learning) adında yeni bir örnekleme yaklaşımı sunmuşlardır. ADASYN' ın temel fikri, azınlık sınıfı örneklerini kolay ve zor olarak etiketlemektir. Öğrenmesi daha kolay azınlık örneklerine kıyasla öğrenmesi daha zor olan azınlık sınıfı örnekleri için daha fazla sentetik verinin üretilmesini sağlar. Üretilen yeni veriler yine SMOTE yöntemi gibi üretilir.

Van Hulse ve arkadaşları (Van Hulse ve ark., 2007), iki sınıfa sahip veya sahip olacak şekilde düzenlenen 35 gerçek yaşam verisi,11 öğrenme algoritması ve 7 örnekleme tekniği kullanarak bir çalışma gerçekleştirmişlerdir. Çalışmalarında sınıflandırma sonuçlarını ölçmek için Eğri Altında Kalan Alan (Area Under The Curve - AUC), Kolmogorov-Smirnov (Hand, 2005) istatistiği, geometrik ortalama, F-Ölçütü, doğruluğu ve doğru pozitif oranı (True Positive Rate - TPR) ölçütlerini kullanmışlardır.

He ve arkadaşları (He ve Garcia, 2009) "Learning From Imbalanced Data" adlı çalışmada geniş kapsamlı bir araştırma yapmıştır. Konu ile ilgili daha önce yapılan çalışmalar ve dengesiz verilerin hangi alanlarda ortaya çıktığı incelenmiş sorunun giderilmesi için performans değerlendirme kriterleri hakkında bilgi vermişlerdir.

Yavaş ve arkadaşları (Yavaş ve ark., 2020), Covid-19 vakalarından toplanan laboratuvar test sonuçlarına göre test sonucu pozitif veya negatif sınıfa ait hastaları SMOTE ve YSA modeli kullanarak daha yüksek bir başarıya sahip sınıflandırma başarısı elde etme için çalışma gerçekleştirmişlerdir. Çalışma sonucunda SMOTE ile dengelenen verilere ait sınıflandırma başarısının daha yüksek olduğunu tespit etmişlerdir.

Topal ve Amasyalı (Topal ve Amasyalı, 2021), 33 veri kümesi üzerinde SMOTE, Borderline-SMOTE ve Rastgele Veri Üretim algoritmaları ile bir çalışma gerçekleştirmiştir. 33 veri kümesi dengesizlik oranına göre dengeli, kısmen dengeli-dengesiz ve dengesiz olmak üzere 3 gruba ayrışmıştır. Çalışma sonucunda temelde dengesiz veri setleri için geliştirilmiş olan bu algoritmalar dengeli veri kümelerinde de başarıya olumlu yönde katkı sağlamıştır. Sınırda değerler üreten Borderline-SMOTE algoritmasının dengeli veri kümelerinde, SMOTE algoritmasının ise kısmen dengeli-dengesiz veri kümelerinde başarılı olduğunu tespit etmişlerdir.

2.2. Algoritma Düzeyi

Algoritma düzeyi yaklaşımı, azınlık sınıfı tanımak için algoritmayı ayarlayarak mevcut sınıflandırıcıyı güçlendirmeyi amaçlar (Sahare ve Gupta, 2012). Sınıf dengesizliği problemini algoritmik bakış açısı ile ele alan ve temel olarak mevcut algoritmaları ve yöntemleri dengesiz verilere uyarlayan farklı öneriler vardır. Bu öneriler arasında maliyete duyarlı öğrenme, tek sınıflı sınıflandırıcılar ve sınıflandırıcı toplulukları sayılabilir (Prati ve ark., 2009).

Herhangi bir makine öğrenimi sürecinde yanlış sınıflandırma maliyetleri, eğitim maliyetleri ve test maliyetleri gibi çeşitli maliyetler ortaya çıkar. Sınıflandırma görevlerinde amaç yanlış sınıflandırılan örneklerin sayısını azaltmak ve yanlış sınıflandırma maliyetleri olarak bilinen doğru sınıflandırılan örnek sayısını artırmaktır. Algoritmik düzeyde yöntemler ve maliyete duyarlı öğrenme, bir öğrenme sürecinde yanlış sınıflandırma maliyetlerini azaltmayı amaçlar (Alhakbani, 2019). Maliyete duyarlı öğrenme, yanlış sınıflandırma maliyetini hesaba katar. Maliyete duyarlı öğrenme algoritmaları, sınıflandırma hata oranını en aza indirmek yerine yanlış sınıflandırma maliyetlerini en aza indirmeyi amaçlar (Prati ve ark., 2009).

Zhang ve arkadaşları (Zhang ve ark., 2010), rastgele az örnekleme yaklaşımına alternatif olarak küme tabanlı az örnekleme yöntemini önermişlerdir. Küme tabanlı az örnekleme yaklaşımı ile çoğunluk sınıfının bilgi kaybı etkili bir şekilde önlenebiliyor. Bu yöntemde belirli oranda k kümeden temsili alt kümeler belirleniyor ve ardından azınlık ve çoğunluk sınıfları üzerinde doğruluk arttırmak için temsili alt kümeyi tüm azınlık sınıfı örneklerinin eğitim verileri olarak kullanmışlardır. Yapılan çalışma ile kümeleme tabanlı az örnekleme yaklaşımının rastgele az örnekleme yaklaşımına göre daha başarılı olduğunu göstermişlerdir.

Sowah ve arkadaşları (Sowah ve ark., 2016) kümelemeye dayalı başka bir az örnekleme yöntemi Cluster Undersampling Technique (CUST) yöntemini önermişlerdir. Bu yöntemde eğitim veri setindeki çoğunluk sınıfı örneklerinden gürültülü ve güvenilmez örnekler silinmiştir. Kalan çoğunluk sınıf örnekleri n alt kümeye bölünmüştür. Modelin performansı rastgele az örnekleme, rastgele aşırı örnekleme, SMOTE ve küme tabanlı az örnekleme performansı ile karşılaştırılmıştır. Çalışma sonucunda CUST' ın daha başarılı sonuçlar elde ettiği görülmüştür.

Grzymala-Busse ve arkadaşları (Grzymala-Busse ve ark., 2005), dengesiz veri kümeleriyle başa çıkmak için iki veri madenciliği yaklaşımını karşılaştırmıştır. İlk yaklaşım, 'Örnek Modülden Öğrenme' (LEM2) algoritması tarafından oluşturulan

orijinal veri setini kaydetmeye ve azınlık sınıfı için maliyeti değiştirmeye dayanır. İkinci yaklaşımda veri seti ikiye bölünür. Parçalardan biri LEM2 ile sınıflandırılırken diğer parça başka bir veri madenciliği algoritması olan EXPLORE ile sınıflandırılır.

Bagging (Bootstrap Aggregation), yönteminde ise veri seti aynı boyutta alt kümelere bölünür. Böylece her alt küme bir sınıflandırma problemi oluşturur. Belirli sınıflandırıcıların toplanması birleşik sınıflandırıcıya katkıda bulunur (Kaur ve ark., 2019).

2.3. Maliyete Duyarlı Yöntemler

Maliyete duyarlı yöntemler sınıf dengesizliği problemlerini çözmek için maliyet fonksiyonlarını kullanır. Yanlış sınıflandırma maliyetini araştırırlar. Diğer yöntemlere göre maliyete duyarlı yöntemler yanlış sınıflandırma maliyetlerinin verilerden belirlenememesi ve maliyetlerin hesaplanmasında zorlukların yaşanması sebebiyle daha az popülerdirler (Kaur ve ark., 2019).

Cao ve arkadaşları (Cao ve ark., 2013), maliyete duyarlı DVM için en iyi öznitelik setini ve sınıflandırma maliyetini seçmek için sarmalayıcı öznitelik seçim yöntemini entegre etmiştir. AUC ve G-Ortalama değerlendirme ölçütlerini kullanarak diğer yöntemler ile karşılaştırmalı sonuçları göstermişlerdir.

Dhar ve Cherkassky (Dhar ve Cherkassky, 2014), U-SVM (Universum-SVM) ile farklı yanlış sınıflandırma maliyetlerine sahip problemlere uyarlamış ve maliyete duyarlı U-SVM' yi önermişlerdir.

Qiu ve arkadaşları (Qiu ve ark., 2017), çok hedefli uyarlanabilir bir öznitelik seçim ölçüsü ve karar ağaçları oluşturmak ve test etmek için basit ama etkili bir yöntem önermişlerdir. Bu algoritma ağaçtaki her bir düğümde test edilecek uygun bir özniteliği bulmak için rastgele bir öznitelik seçim ölçüsü kullanır. Spesifik olarak, ağaç oluşturmadaki tüm nitelik uzayında rastgele bir arama yaptılar ve ortaya çıkan modele rastgele seçilmiş karar ağacı (RSDT) adını verdiler. Bu sayede RSDT, toplam test maliyetini önemli ölçüde azaltırken, aynı zamanda rakiplerine göre daha yüksek sınıflandırma doğruluğunu korumuştur.

2.4. Melez (Hibrit) Yöntemler

Melez yöntemler veri düzeyi ve algoritma düzeyinde yöntemlerin birleşimi ile oluşturulan yöntemlerdir (Kaur ve ark., 2019).

Zhang ve arkadaşları (Zhang ve ark., 2015), dengesiz sınıflandırma yöntemini ve topluluk öğrenme tekniğini birleştiren dengesiz duygu sınıflandırması için bir yöntem önermişlerdir. Topluluk öğrenimi çerçevesinde bu hibrit yöntem yetersiz örnekleme, önyüklemeli yeniden örnekleme (bootstrap re-sampling) ve veri kümesini işlemek için rastgele özellik seçimi yöntemlerini birleştirmiştir.

Cao ve Zhai (Cao ve Zhai, 2015), iki sınıflı dengesiz veri seti sınıflandırması için yeni bir hibrit yeniden örnekleme yaklaşımı önermişlerdir. İlk önce sentetik veriler üretmek için SMOTE tekniğini kullanmışlardır. Ardından çoğunluk sınıfına ait bazı örnekleri silmek için One Side Selection (OSS) yöntemini kullanmışlardır. Elde ettikleri yeni veri setini DVM ile sınıflandırmışlardır. Yaptıkları çalışmanın etkinliğini UCI üzerindeki 5 veri seti üzerinde uyguladıkları deneysel sonuçlar ile göstermişlerdir.

Tang ve arkadaşları (Tang ve ark., 2008), son derece dengesiz veri kümesi için özel olarak üretilmiş bir algoritma olan, tanecikli SVM-tekrarlı alt örnekleme anlamına gelen GSVM-RU'yu önerdi. Bunun arkasındaki ana fikir, iyi bir sınıflandırma doğruluğu elde etmek için gürültülü ve güvenilmez örneklerin silinmesi, faydalı ve tutarlı örneklerin çıkarılmasıdır. Önerilen teknik, karşılaştırma sırasında diğer yöntemlerden daha iyi performans göstermiştir.

Gao ve arkadaşları (Gao ve ark., 2011), SMOTE ve Parçacık Sürü Optimizasyonu (Particle Swarm Optimisation - PSO) destekli Radyal Temel Fonksiyon (Radial Basis Function - RBF) sınıflandırıcısını birleştirerek güçlü bir yöntem önermişlerdir. Dört farklı veri seti üzerinde üç yöntem ile karşılaştırdıkları sonuçlarda başarı elde etmişlerdir (Pir, 2022).

3. MATERYAL

Arıların yaşam alanı olan kovanlar arıların sağlıkları ve verimlilikleri ile doğru orantılıdır (Şeker ve ark., 2017). Farklı arı kovanı türleri bulunmaktadır. Arı kovanı seçimi; üretilecek arıcılık ürünlerinin niteliğine, devam eden araştırmalara ve hükümet politikalarına bağlıdır. Arıcılığın, tarımsal bir faaliyet olarak kabul edilmesi ve zaman içinde gelişmesiyle beraber üretimdeki verimin arttırılması amacıyla dünya üzerinde farklı kovan tipleri önerilmiştir. Kovan seçimindeki temel kriterler; kovanın ucuz, yüksek verimlilik sağlayacak, çevre dostu ve kullanımının kolay olmasıdır (Genç ve ark., 2020). Dünya genelinde en çok tercih edilen ve kabul görmüş modern kovan çeşitleri Langstroth ile Dadant tipi kovanlardır. İki kovan türünde sistem aynıdır ancak ölçüler farklılık göstermektedir. Langstroth kovan tipi Dünya arıcılırının %75 oranla kullandığı kovan tipidir (Genç ve ark., 2020). Langstroth kovan tipi, Langstroth tarafından arıların petekleri yaparken aralarında aynı mesafede boşluk bıraktığını keşfetmesiyle 1851 yılında geliştirilmiştir (Langstroth, 1857). Bu çalışmada kulllanılan kovanların ve bal peteklerinin standartı Langstroth kovan standartlarındadır.

Langstroth kovan tipi gezginci, kışların ılık geçtiği, sıcak ve kurak iklime sahip bölgelerde yapılan arıcılık faaliyetlerine uygundur. Şekil 3.1’ de görüleceği üzere Langstroth kovan tipi alttan üste doğru kovan dip tahtası ve uçuş tahtası, kuluçkalık, ballık, örtü tahtası ve kovan kapağı olmak üzere 5 kısımdan oluşmaktadır (Genç ve ark., 2020).

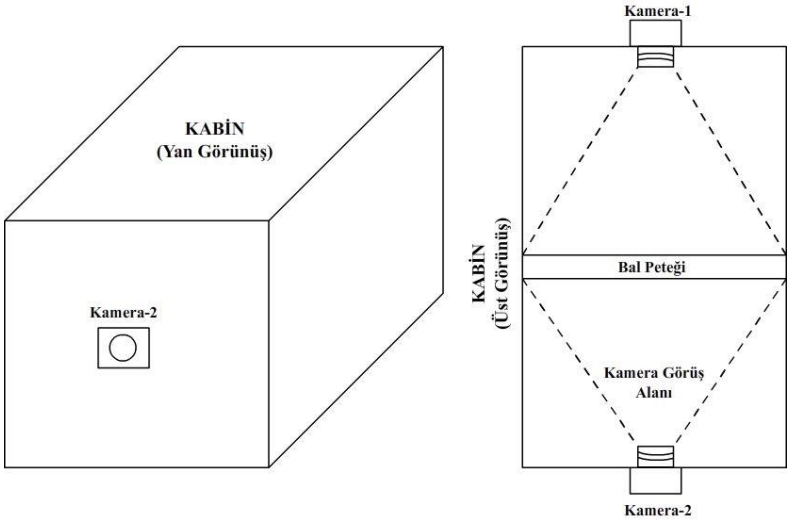
Şekil 3.1.: Langstroth kovan tipi yapısı

Kovanın en altında gerektiğinde çıkarılabilen kovan dip tahtası bulunmaktadır. Bu parça gerektiğinde çıkarılabilecek yapıdadır. Ancak ülkemizde yaygın olan gezginci arıcılık faaliyetleri sebebiyle kovan dip tahtası sabit olacak şekilde tasarlanmaktadır. Uçuş tahtası dip tahtası boyunca menteşeli ve kapanacak şekilde yapılmaktadır. Uçuş tahtası arılar için bir nöbet tutma yeri görevi görmektedir. Kanat çırparak kovana hava pompalayan arılar bu görevi uçuş tahtasında gerçekleştirir. Ayrıca arıların kovana giriş çıkışını da kolaylaştırır. Kuluçkalık arıların yavru yetiştirdiği kovanın ana parçasıdır. Aynı zamanda arıların kışladıkları ve gıda stokunu yaptığı yerdir. Ballık, hasat edilecek balın alındığı kısımdır. Ana arı kuluçkalıkta yer kalmadığında ballıkta da sürdürebilir. Örtü tahtası, kovan kapağı altına yerleştirilen iç kapak görevi görmektedir. Kovan kapağı tüm kovanın koruyucusudur. Kovanı yağmur ve kar suyundan korur.

Bu çalışmada kullanılacak görüntüler TÜBİTAK 1512 Tekno girişim sermayesi Desteği Programı 2170060 numaralı projeden alınmıştır. Bu görüntüler, toplamda 19 tane Langstroth standardında farklı özellikli bal peteğinden elde edilmiştir. Her bir peteğin iki yüzü olduğu için toplamda 38 adet bal peteği görüntüsü üzerinde çalışmalar yapılacaktır. Farklı özellikli peteklerle çalışmanın nedeni ise peteklerin farklı karakteristik özelliklere sahip olması ve çalışmanın kapsayıcılığının arttırılmasıdır. Bu karakteristik özellikleri, arının cinsi, bulunduğu bölge, bal için topladığı nektar (çam, fi, ay çiçeği, mısır, çiçek vb.), peteğe daha önce larva konulması gibi etkenler etki etmektedir. Bu etmenler peteklerin renk dokusunu değiştirmektedir.

Görüntüler elde edilirken BASLER acA2500-14uc alan tarama kamerası kullanılmıştır. Projede kullanılan BASLER acA2500-14uc alan tarama kamerası, maksimum 2590x1942, minimum 64x64 piksel ekran çözünürlüğünde görüntü sağlamaktadır. Ayrıca kazanım değeri 0-23,7 arasında değişebilmektedir. Proje çalışmasında görüntülerin elde edilirken dış ortamdaki gürültülerin engellenebileceği bir kabin hazırlanmış ve ışık kaynağı kullanılmıştır. Çerçeve görüntüleri 2590x1940 piksel boyutlarında çekilmiştir, fakat çalışma alanı sadece peteğin bulunduğu kısım ile sınırlandırılmıştır. Petek bölgesinin boyutları 1162x574 piksel olarak belirlenmiştir.

Görüntülerin elde edilmesinde kullanılan kabin düzeneğine ait, yan ve üst görünüş temsili olarak aşağıdaki Şekil 3.2.' de verilmiştir. Bal peteğindeki görüntüler anlık olarak her iki yüzeyi için alınacağından dolayı, aynı özelliklere sahip iki kamera kullanılmıştır. Kabinin tam ortasında ise bal peteği bulunmaktadır.

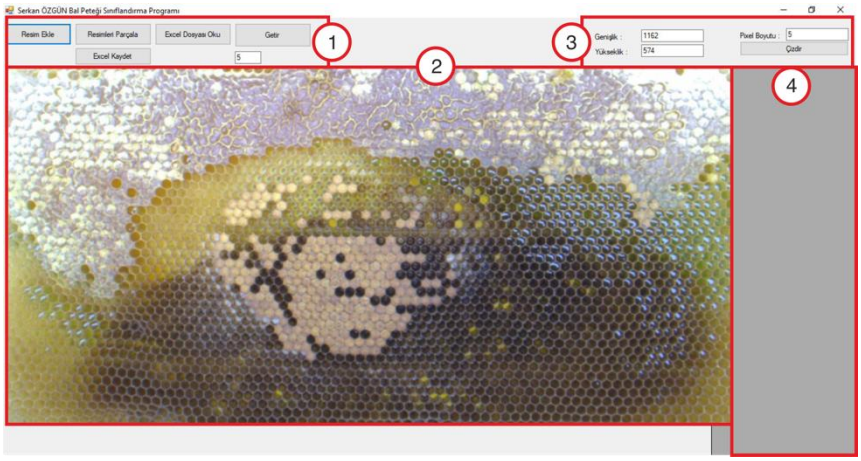


Şekil 3.2: Bal peteği görüntülerinin elde edilmesinde kullanılan kabin düzeneği

Bal peteğindeki görüntüler anlık olarak kaydedildikten sonra mevcut hali ile kullanılamamaktadır. Dolayısıyla bal peteğinin dışında kalan görüntüler silinmektedir. Kırpılmış görüntü ise veri setinin oluşturulmasında kullanılmıştır.

Şekil 3.3: Bal peteği görüntülerinin temizlenmesi (Farahmand, 2022).

Bal peteği hücreleri altıgen şekline sahiptir ancak sırlama (bal peteği hücresinin arılar tarafından kapatılması) işleminden sonra hücrelerin altıgen şeklinin tespit edilmesi zorlaşmaktadır. Dolayısıyla sınıfların belirlenmesinde 5x5 piksellik alanlardan görüntüler alınmış ve sınıfı etiketlenmiştir.



Şekil 3.4: Bal petekleri üzerinde kapalı larva hücrelerini etiketleme ve sınıflandırma programı

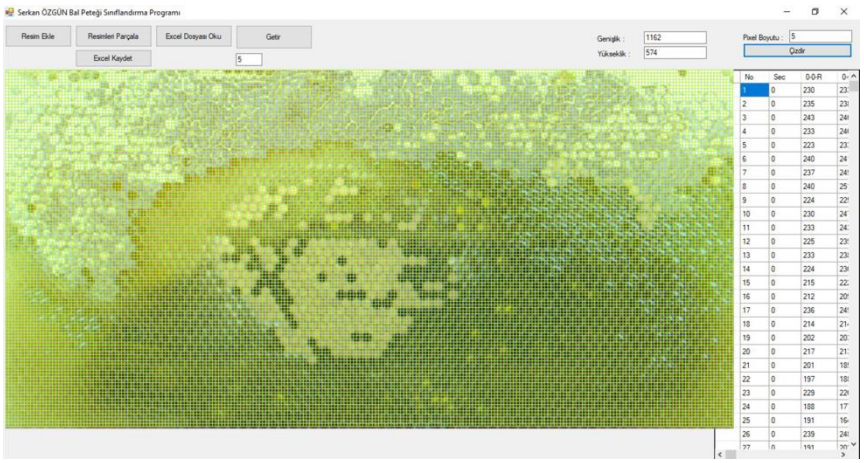
Kapalı larva bulunan alanların etiketlenmesi için Şekil 3.4' te görülen program geliştirilmiştir. Geliştirmiş olduğumuz program dört bölümden oluşmaktadır.

Birinci bölüm, program kullanımı için gerekli butonların bulunduğu bölümdür. Bu bölümde Resim Ekle butonu, üzerinde etiketleme yapılacak peteğe ait görüntüyü seçmek için kullanılmaktadır. Excel Kaydet butonu etiketlenmiş verileri excel formatında kaydetmek için kullanılmaktadır. Excel Dosyası Oku butonu daha önce excel dosyasına kaydedilmiş etiketleme verilerini geri yüklemek için kullanılmaktadır. Getir butonu, Excel Dosyasını Oku butonu ile yüklenmiş excel dosyasındaki etiketlenmiş alanları petek görüntüsü üzerinde göstermek için kullanılmaktadır (Şekil 3.6).

İkinci bölüm, programa yüklenen bal peteğine ait görüntünün gösterildiği kısımdır.

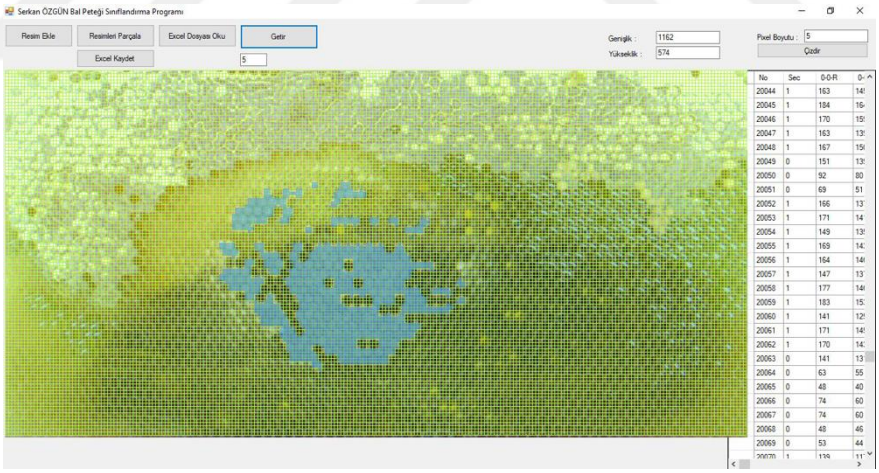
Üçüncü bölümde, seçilen petek görüntüsüne ait yükseklik ve genişlik değerleri görülmektedir. Çizdir butonu ile üstündeki kutucukta yazan değer boyutunda olacak şekilde görüntü üzerinde alanların sınırları çizilmektedir (Şekil 3.5). Bu çalışma için görüntüleri 5x5 piksel boyutunda parçalara ayırdığımız için 5 değeri görülmektedir.

Dördüncü bölümde, Çizdir butonu ile petek görüntüsü üzerinde oluşturulan ızgaralarla beraber sağ taraftaki kısımda, her 5x5 piksellik görüntü için bir satır olacak şekilde oluşan tablo görülecektir.



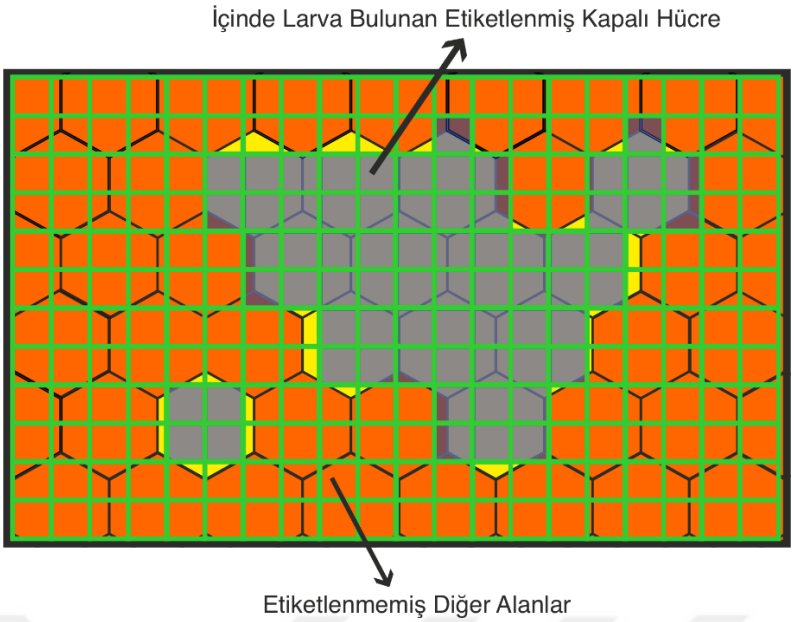
Şekil 3.5: Bal peteklerinden veri elde etme süreci

Şekil 3.5' de görüleceği üzere petek görüntüsü Çizdir butonu ile işaretlendikten sonra her bir 5x5 piksellik görüntü için sağ kısımdaki tabloda bir satır oluşmaktadır. Her bir satırda görüntü parçasının ait olduğu sınıf etiketi ve her bir piksel için sırasıyla R-G-B değerleri tutulmaktadır. Böylece 26.448 adet satır ve 76 sütundan oluşan bir tablo oluşmaktadır.



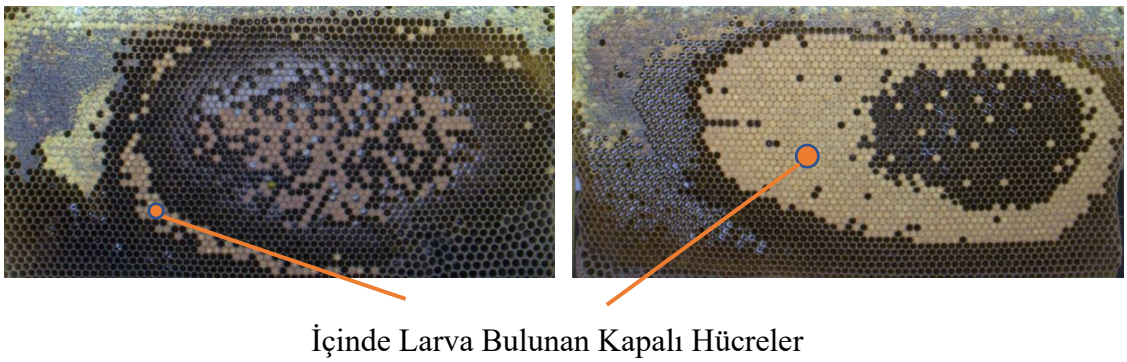
Şekil 3.6: Bal peteği üzerinde sınıfların etiketlenmesi

Kapalı larva hücreler, görüntü üzerinde işaretlendikten sonra etiketlenen kısımlar Şekil 3.6' da görüldüğü üzere faklı renk almaktadır ve sağ taraftaki tabloda sınıf etiketi "1" olmaktadır. Görüleceği üzere bal peteği görüntüleri 5x5 piksel olacak şekilde parçalanmış ve kapalı larva hücreler tek tek görsel üzerinde etiketlenmiştir.



Şekil 3.7: Bal peteği üzerinde içinde larva bulunan hücrelerin etiketlenmesinde karar süreci.

Etiketleme işlemine karar verilirken 5x5 piksellik resim parçasında, kapalı larva hücresinin kapladığı alan yaklaşık olarak %50' den fazla ise "1", tersi durumda ise "0" olacak şekilde etiketlenmiştir (Şekil 3.7). Bir kapalı larva hücre yaklaşık olarak 4 veya 5 adet 5x5 piksellik alan kaplamaktadır. Petek hücresinin komşu hücreleri yine 'içinde kapalı larva' bulunan bir hücre olduğunda birbirlerinin devamı şeklinde görüneceklerinden dolayı bir hücredeki 5x5 piksellik parça sayısı artabilmektedir. Ancak komşu hücresi kapalı larva içeren bir hücre olmadığında bu sayı daha az olmaktadır. Bu şekilde "larva bulunan kapalı hücreler" bir sınıf diğer alanlar bir sınıf olacak şekilde iki farklı sınıf elde edilmiştir.

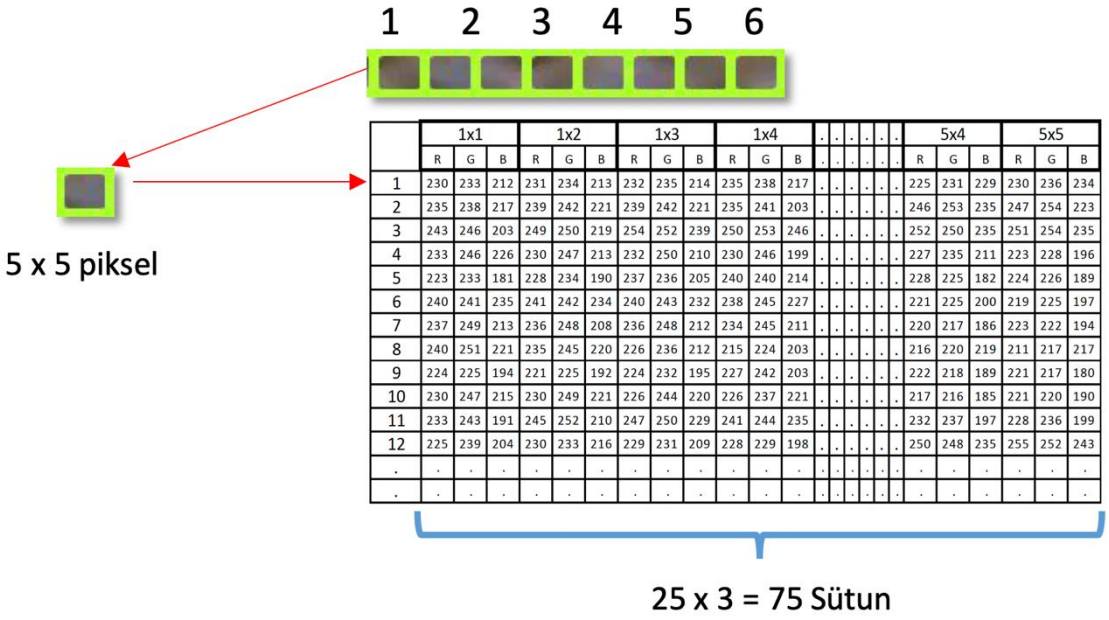


Şekil 3.8: Bal peteği üzerinde içinde larva bulunan hücreler.

Şekil 3.8' de görüleceği üzere petekler üzerinde renk dokuları farklılık göstermektedir. Ayrıca içinde larva bulunan alanlar seyrek ve dağınık olabileceği gibi yoğun ve bir arada da olabilmektedir. Bunlarında dışında üzerinde hiç larva bulunmayan peteklerimizde bulunmaktadır.

Bir petek görüntüsünde bakıldığında içinde yavru olan kapalı hücreler, içinde polen olan hücreler, bir miktar bal olan açık hücreler, içinde bal olan kapalı hücreler, içi boş hücreler ve içinde larva olan açık hücreler şeklinde farklı bölümler bulunabilir ve ayrı etiketlerler sınıflandırılabilir. Bu çalışmada verilerimiz içinde larva olan kapalı hücre ve diğerleri olmak üzere iki temel sınıfa ayrılmıştır. Arıcılık faaliyetlerinde içinde yavru bulunan kapalı hücrelerin tespiti hem arı neslinin devamı hem de yavruların bala karışıp homojenliği bozmaması açısından önemlidir. Çalışmadaki amacımız, içinde larva bulunan kapalı hücreleri tespit etmek olacağından diğer kısımlar tek bir sınıf olarak görülmüştür. Larva bulunan kapalı hücrelere ait görüntüler "1", diğer görüntüler "0" olarak etiketlenmiştir.

Her bir petek görüntüsü 5x5 olacak şekilde parçalanırken 232 sütun ve 114 satır elde edilmiş. Bu şekilde her petekten 26.448 resim parçası elde edilmiştir. 38 adet petek görüntüsünden elde ettiğimiz küçük resim sayısı 1.005.024 olmuştur. Bu görüntülerden 188.147 tanesi yavru olan kapalı hücrelere ait ve "1" olarak etiketlenmiştir. Geri kalan 816.877 adet veri "0" olarak etiketlenmiştir. İki sınıf arasında yaklaşık olarak 1/5 gibi bir oran olduğu görülmektedir.



	1x1			1x2			1x3			1x4			. . . . .	5x4			5x5		
	R	G	B	R	G	B	R	G	B	R	G	B		R	G	B	R	G	B
1	230	233	212	231	234	213	232	235	214	235	238	217	. . . . .	225	231	229	230	236	234
2	235	238	217	239	242	221	239	242	221	235	241	203	. . . . .	246	253	235	247	254	223
3	243	246	203	249	250	219	254	252	239	250	253	246	. . . . .	252	250	235	251	254	235
4	233	246	226	230	247	213	232	250	210	230	246	199	. . . . .	227	235	211	223	228	196
5	223	233	181	228	234	190	237	236	205	240	240	214	. . . . .	228	225	182	224	226	189
6	240	241	235	241	242	234	240	243	232	238	245	227	. . . . .	221	225	200	219	225	197
7	237	249	213	236	248	208	236	248	212	234	245	211	. . . . .	220	217	186	223	222	194
8	240	251	221	235	245	220	226	236	212	215	224	203	. . . . .	216	220	219	211	217	217
9	224	225	194	221	225	192	224	232	195	227	242	203	. . . . .	222	218	189	221	217	180
10	230	247	215	230	249	221	226	244	220	226	237	221	. . . . .	217	216	185	221	220	190
11	233	243	191	245	252	210	247	250	229	241	244	235	. . . . .	232	237	197	228	236	199
12	225	239	204	230	233	216	229	231	209	228	229	198	. . . . .	250	248	235	255	252	243
.	.	.	.	.	.	.	.	.	.	.	.	.	. . . . .	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	. . . . .	.	.	.	.	.	.

Şekil 3.9: Bal peteği görüntülerinden veri seti elde etme süreci

5x5 piksel olacak şekilde elde ettiğimiz her bir görüntü için 1. pikselden başlamak üzere 25 pikseldeki RGB değerleri sırasıyla R-G-B olacak şekilde sütunlara kaydedildi. Her bir görüntü için 0 – 255 arasında değerler barındıran 75, etiket değeri (0-1) ile beraber 76 sütuna sahip bir satır verisi elde edilmiştir. Sonuç olarak 1.005.024 satır ve 76 sütundan oluşan bir veri seti elde edilmiştir. (Şekil 3.9)

4. YÖNTEM

Bu çalışmada bal peteği görüntülerinden elde edilen veri setindeki dengesizliğin giderilmesi için sentetik veri üretme (aşırı örnekleme - oversampling) yöntemleri kullanılmıştır. Sentetik veri üretme yöntemleri ile dengeli hale getirilen veri seti, sınıflandırma algoritmaları ile sınıflandırılmıştır. Sınıflandırma sonuçları da belirlenen ve bu alanda en çok kullanılan performans metrikleri ile değerlendirilmiştir.

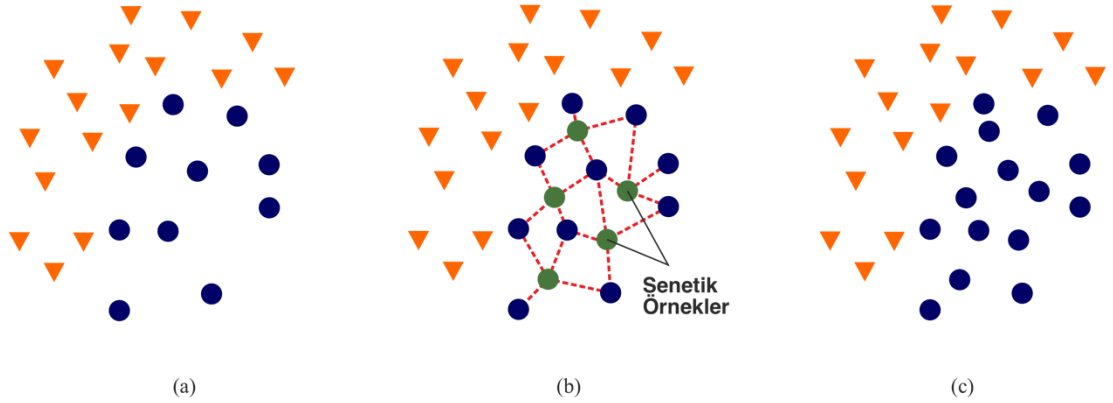
Bu tez çalışmasında, belirlenen aşırı örnekleme yöntemleri ve temel sınıflandırıcılar ile elde edilen sınıflandırma sonuçları orijinal veri dahil olmak üzere kıyaslanmıştır. Detaylı kıyaslamaların sonuçları incelenmiştir.

4.1. Sentetik Veri Üretme Yöntemleri

Veri setlerinde, sınıflara ait örnek sayısındaki dengesizlik problemini giderebilmek için baskın olan sınıfa ait verileri azaltmak ya da azınlık sınıfa ait olan verileri arttırmak gerekmektedir. Veri düzeyinde sentetik veri arttırma işlemleri genel olarak Aşırı örnekleme -Oversampling, veri azaltma işlemleri de Az Örnekleme - Undersampling başlıkları altında toplanmaktadır. Az örnekleme (Undersampling) metotlarında var olan gerçek veriler azaltılırken baskın sınıfa ait veriler kaybolmaktadır. Aşırı örnekleme (Oversampling) için en büyük avantaj gerçek verilerin kaybolmamasıdır (Uyanık ve Kasapbaşı, 2021). Bu çalışmada sentetik veri üretme yaklaşımları olarak SMOTE, Borderline-SMOTE1, Borderline-SMOTE2, Safe-Level-SMOTE ve DEBOHID kullanılmıştır.

4.1.1. SMOTE (Synthetic Minority Oversampling Technique)

Chawla ve arkadaşları (Chawla ve ark., 2002) tarafından geliştirilen SMOTE algoritması, azınlık verisi örneklerinden yeni veriler üreterek dengesizliği gidermeye çalışan, literatürde de en çok bilinen ve kullanılan veri örnekleme yöntemlerinden bir tanesidir. SMOTE yaklaşımına süreçler Şekil 4.1’te verilmiştir.



Şekil 4.1.: SMOTE sentetik veri üretme yöntemi. (a) Dengesiz veri seti. (b) Sentetik veri oluşturma süreci. (c) Dengeli hale getirilen veri seti.

SMOTE algoritmasında azınlık sınıfına ait her bir gözlem ($\overrightarrow{x_i}$) için Öklid mesafesine göre azınlık sınıfına ait k en yakın komşudan komşular rastgele ($\overrightarrow{x_j}$) alınır. Örneklem ile seçilen komşu arasındaki fark hesaplanır. Sonrasında 0 ile 1 arasında rastgele bir sayı ($\vec{\alpha}$) seçilir. Örneklem ile k en yakın komşusu arasındaki fark, elde edilen rastgele sayı ile çarpılır. Sonrasında Denklem 4.1 kullanılarak yeni sentetik veri $\overrightarrow{x_{snt}}$ elde edilmiş olur (Aydin Hakli, 2018).

$$\overrightarrow{x_{snt}} = \overrightarrow{x_i} + \left(\overrightarrow{x_j} - \overrightarrow{x_i}\right) * \vec{\alpha} \qquad (4.1)$$

Denklem 4.1.' de görüleceği üzere SMOTE yöntemi ile her azınlık örneği için örnekleme oranına göre seçilen azınlık sınıfına ait komşu örnekler arasında yeni örnekler üretilir.

4.1.2. Borderline-SMOTE

Sınıflandırma algoritmalarının çoğunun eğitim sürecindeki ilk amacı her sınıfın sınır çizgisini öğrenmeye çalışmaktır. Sınır çizgilerinin doğru şekilde tespiti daha başarılı bir sınıflandırma anlamına gelecektir. Sınırlara yakın örnekler uzaktaki örneklere göre yanlış sınıflandırılmaya daha yatkın örneklerdir. Bu nedenle sınıflandırma aşamasında bu örnekler daha önemli hale gelmektedir (Han ve ark., 2005).

Yukarıdaki analize dikkat çeken Han ve arkadaşları (Han ve ark., 2005) temelde SMOTE (Sentetik Azınlık Örneklem Arttırma Yöntemi) algoritmasına dayanan iki yeni yöntem önermişlerdir. Bu yöntemleri Borderline-SMOTE1 ve Borderline-SMOTE2 olarak isimlendirilmiştir. SMOTE yönteminde her azınlık örneği için seçilen komşu

örnekler arasında yeni örnekler üretilirken, Borderline-SMOTE yalnızca sınırdaki azınlık örneklerini aşırı örneklemektedir. Sınır çizgisine yakın azınlık sınıfı örnekleri tespit edildikten sonra bunlardan sentetik veriler üretilir ve gerçek eğitim setine eklenir.

Borderline-SMOTE1 yöntemi için; tüm eğitim setinin T olduğunu, azınlık sınıfının P, çoğunluk sınıfının N olduğunu varsayalım.

$$P = (p_1, p_2, \dots, p_{pnum}), \qquad N = (n_1, n_2, \dots, n_{nnum}) \tag{4.2}$$

Denklem 4.2' de p_{pnum} azınlık sınıfı örnek sayısı, n_{nnum} çoğunluk sınıfı örnek sayısıdır. Azınlık sınıfındaki her örnek için (p_i $(i = 1,2, \dots, p_{pnum})$) tüm eğitim setinden en yakın m adet komşu hesaplanır. En yakın m komşu arasındaki çoğunluk örneklerinin sayısını m' (0≤m≤ m') olarak gösterirsek;

m = m' ise, yani p_i' nin en yakın komşularının hepsi çoğunluk sınıfına ait ise p_i gürültülü kabul edilir ve bu örnekler göz ardı edilir.

m/2 ≤ m' < m ise, yani p_i' nin çoğunluk sınıfına ait komşu sayısı azınlık sınıfa ait komşu sayısından fazla ise yanlış sınıflandırmaya yatkındır denir ve TEHLİKE grubu olarak kabul edilir.

0 ≤ m' < m/2 ise, yani p_i' nin tüm komşuları azınlık sınıfından ise veya azınlık sınıfına ait komşu sayısı çoğunluk sınıfına ait komşu sayısından fazla ise p_i güvenli olarak kabul edilir ve yine göz ardı edilir.

TEHLİKE sınıfı örnekleri *P* sınıfının sınır örnekleridir (Denklem 4.3).

$$TEHLİKE = \{p'_1, p'_2, \dots p'_{dnum}\},\ 0 \le dnum \le pnum \tag{4.3}$$

TEHLİKE sınıfına ait her örnek için P sınıfına yani azınlık sınıfa ait k adet komşu hesaplanır.

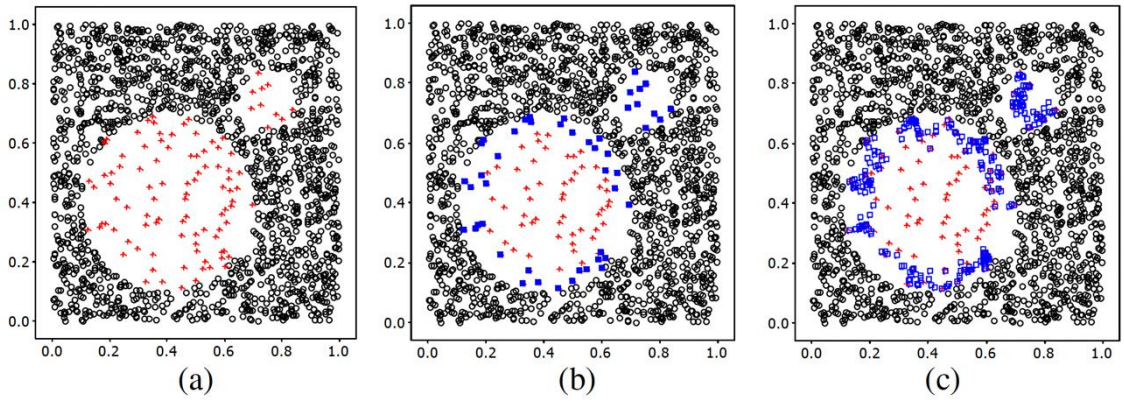
Sonraki adımda TEHLİKE gurubu örneklerinden s x dnum kadar sentetik pozitif örnek üretilir. Burada s değeri 1 ile k arasında bir değerdir. Her bir p'_1 için azınlık sınıfındaki k komşusundan rastgele s en yakın komşusu seçilir. İlk olarak p'_1ve onun azınlık sınıfına ait s en yakın komşusu arasındaki fark dif_j ($j = 1,2, \dots, s$) hesaplanır. Sonrasında dif_j, 0 ve 1 arasında rastgele seçilen bir sayı r_j ($j = 1,2, \dots, s$) ile çarpılır. Sonuç olarak p'_1 ve onun komşuları arasında s sentetik veri üretilir. Bu şekilde yeni

sentetik veri üretilmiş olur. Denklem 4.4.'de sentetik veri üretme formülasyonu gösterilmiştir. Bu prosedür tüm örnekler için uygulanır ve yeni sentetik veriler üretilir.

$$sentetik_j = p'_i + r_j \; x \quad dif_j, \quad j = 1,2,\dots,s \tag{4.4}$$

TEHLİKE sınıfı örnekleri tespit edildikten sonra SMOTE' a benzer işlemler gerçekleştirilir ve $s \; x \; dnum$ kadar sentetik veri elde edilir.

Bu yöntemle azınlık sınır örnekleri ile en yakın komşuları arasında bir hat boyunca yeni sentetik verilerin üretildiği ve böylece azınlık sınıfının sınırında bulunan verilerin güçlendiği görülecektir. Böylece, sınıflar arasındaki ayrımın daha rahat yapılabilmesi sağlanmaya çalışılır. Borderline-SMOTE' a ait gösterim Şekil 4.2' de görülmektedir.



Şekil 4.2.: (a) Veri setinin gerçek dağılımı. (b) Sınırda azınlık örnekleri (dolu kareler). (c) Sınırda sentetik azınlık örnekleri (içi boş kareler) (Han ve ark., 2005)

Bordeline-SMOTE1 ile Borderline-SMOTE2 arasındaki temel fark Borderline-SMOTE1 algoritmasının TEHLİKE sınıfındaki örneklerin komşularını sadece azınlık sınıfından seçmesidir. Bir başka deyiş ile Borderline-SMOTE2 azınlık sınıfına ait örneğin en yakın k komşusunu seçerken çoğunluk sınıfındaki verileri de seçebilir. Seçilen örnek ile komşusu arasındaki fark 0 ile 0,5 arasında rastgele seçilen bir sayı ile çarpılır. Böylece yeni üretilen örneklerin azınlık sınıfına daha yakın olmuş olması sağlanır (Han ve ark., 2005).

4.1.3. Safe-Level-SMOTE

Bunkhumpornpat ve ark. (Bunkhumpornpat ve ark., 2009), temelde SMOTE algoritmasına dayanan Safe-Level-SMOTE algoritmasını önermişlerdir. Sentetik veriler

oluşturulmadan önce azınlık sınıfı örneklerine kendi güvenli seviyesini atamışlardır. Tüm sentetik örnekler güvenli seviyeye yakın konumlandırılır, böylece sentetik veriler sadece güvenli bölgelerde oluşturulur.

Güvenli seviye (Safel Level, sl) Denklem 4.5 ile hesaplanır. Bir örneğin güvenli seviye değeri 0' a yakınsa örnek gürültülü kabul edilir. k' ye yakınsa örnek güvenli kabul edilir. Güvenli seviye oranı (Safe Level Ratio, sl_ratio) ise formül Denklem 4.6 ile hesaplanmıştır.

sl = k en yakın komşu arasında ki pozitif örnek sayısı (4.5)

sl_ratio = bir pozitif örneğin sl değeri / en yakın komşusunun sl değeri (4.6)

Safe-Level-SMOTE yöntemi için tüm orijinal pozitif örnekler kümesi D olarak tanımlanırsa, p bu kümedeki bir örnektir. n, p'nin seçilmiş en yakın komşularıdır. s ise tüm sentetik pozitif örnekler kümesi olan D' kümesine dahil edilen sentetik örnektir. sl_p ve sl_n sırasıyla p ve n' nin güvenli seviyesidir. p ve n' ye güvenli seviye değerleri atandıktan sonra güvenli seviye oranı (sl_ratio) hesaplanır. n' ye ait güvenli seviye oranı 0 olursa (sl_ratio) sonsuza eşittir denir, 0' eşit olmazsa sl_n / sl_p olacak şekilde hesaplanır. Güvenli seviye oranına göre beş farklı durum ortaya çıkar.

Birinci durum; güvenli seviye oranının sonsuza eşit olması ve p' nin güvenli seviyesinin sıfıra eşit olmasıdır. Bu durumda p ve n gürültülü kabul edilir ve algoritma gürültülü bölgede sentetik veri üretmez.

İkinci durum; güvenli seviye oranının sonsuza eşit olması ve p' nin güvenli seviyesinin sıfıra eşit olmamasıdır. Bu durumda n gürültülü kabul edilir ve algoritma n gürültülü örneğinden kaçınmak için p' yi kopyalayarak n gürültülü örneğinden uzakta bir sentetik veri üretecektir.

Üçüncü durum; güvenli seviye oranının 1' e eşit olmasıdır. Bu p ve n' nin güvenli seviyesinin aynı olduğu anlamına gelir. Bu durumda p ve n arasındaki hat boyunca sentetik bir örnek oluşturulacaktır. Çünkü bu durumda p, n kadar güvenlidir.

Dördüncü durum; güvenli seviye oranının 1' den büyük olduğu durumdur. Bu durumda sentetik veri p' ye daha yakın konumlandırılır. Çünkü p, n' den daha güvenlidir denir. Sentetik örnek ($0 - 1/sl\_ratio$) aralığında oluşturulacaktır.

Beşinci durum; güvenli seviye oranının 1'den küçük olduğu durumdur. Bu durumda p'nin güvenli seviyesinin n' ninkinden daha az olduğu anlamına gelir. Bu durum

oluşursa sentetik veri n' ye daha yakın konumlandırılır. Çünkü n, p' den daha güvenlidir. Sentetik örnek $(1 - sl\_ratio\ ,\ 1)$ aralığında oluşturulacaktır.

Her pozitif örnek için bu beş durum kontrol edilir ve ilk durum oluşmazsa p ve n arasında belirli bir aralıklı çizgi boyunca s örneği oluşturulur ve D' kümesine eklenir.

Bunkhumpornpat ve ark. (Bunkhumpornpat ve ark., 2009), iki sınıflı dengesiz problemi inceledikleri çalışmada Safe-Level-SMOTE yönteminin SMOTE ve Borderline-SMOTE yöntemlerinden daha iyi sınıflandırma sonuçları verdiğini deneysel çalışmalarıyla göstermiştir.

4.1.4. DEBOHID

Kaya ve ark. (Kaya ve ark., 2021), azınlık sınıfı için yeni örnekler oluşturan 'Yüksek düzeyde dengesiz veri kümeleri için bir diferansiyel evrim tabanlı aşırı örnekleme yaklaşımı (DEBOHID: Differential Evolution Based Oversampling approach for Highly Imbalanced Datasets)' adını verdikleri bir aşırı örnekleme yaklaşımı önermişlerdir. Önerilen DEBOHID yaklaşımında SMOTE metodundan esinlenilmiş ve Diferansiyel Evrim (DE) algoritmasının temel donör üretme stratejisi kullanılmıştır.

DE, Storn ve Price (Storn ve Price, 1997) tarafından geliştirilen sürekli optimizasyon problemler için önerilen etkin bir metasezgisel algoritmadır. DE 4 ana adımdan oluşur. Bu aşamalar: Vektörlerin başlatılması, vektörlerdeki fark, çaprazlama işlemi ve seçim aşamalarıdır. SMOTE metodu, sentetik veri üretme aşamasında rastgele seçilen bir komşu örnekten faydalanır. DEBOHID yaklaşımı ise sentetik veri üretiminde çeşitliliği arttırabilmek amacıyla üç adet farklı komşudan örnekten faydalanır. Bu işlemin gerçekleşebilmesi için DE' nin donör vektör üretme formülünden faydalanılır (Denklem 4.7).

$$V_i = X_{r1} + F\ x\ (X_{r2} - X_{r3}) \quad r1 \neq r2 \neq r3 \tag{4.7}$$

Burada V_i donör vektördür. r1, r2, r3 birbirinden farklı indekslerdir. X popülasyon vektörleridir. F, 0 ile 2 arasında belirlenen ölçeklendirme faktörüdür. Sentetik veri üretilirken, mevcut veri mi yoksa Denklem 4.7 ile üretilen donör verinin mi sentetik veri havuzuna ekleneceğini belirlenirken çaprazlama oranından (CR - crossover rate) faydalanılır (Denklem 4.8).

$$T_{i,j} = \begin{cases} V_{i,j}\,, if\ rand\ (0,1) \leq CR\ or\ j\ = rj \\ X_{i,j}\,, if\ rand\ (0,1) > CR\ or\ j\ \neq rj \end{cases} \quad (4.8)$$

Burada $T_{i,j}$ deneme (trial) vektörünün j. boyutudur. $rand(0,1)$, 0 ve 1 arasında rastgele bir sayıdır. rj rastgele seçilen boyut, CR kullanıcı tarafından başlatma aşamasında belirlenen çaprazlama oranı değeri, $V_{i,j}$, i. donör vektörün j. boyutudur. $X_{i,j}$, i. vektörün j. boyutudur. j boyut indeksidir.

Yeni oluşturulan aday çözüm, temel alınan aday çözüme yakın veya benzer olması beklenir. Böylece çoğunluk sınıfındaki bir örneğe benzemeyen bir sentetik üretme işlemi gerçekleştirilmiş olur.

4.2. Sınıflandırmada Kullanılan Algoritmalar

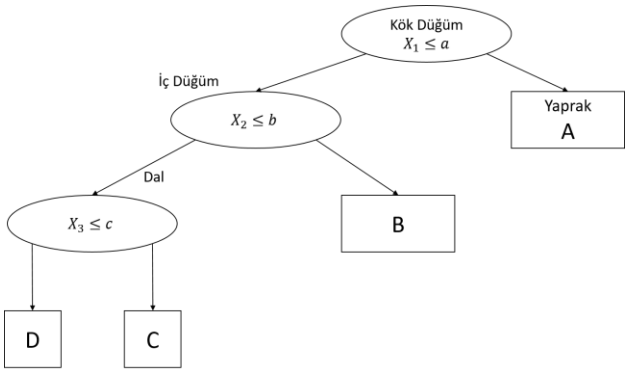
4.2.1. Karar ağacı (Decision Tree)

Karar ağacı (KA) tümevarım ile çıkarım yapan bir denetimli öğrenme algoritmasıdır. Karar ağacı sınıflandırması ağaç şeklinde bir yapı oluşturmaktadır. Baştan sona bir kural yapısına sahiptir. Büyük verileri küçük alt parçalara böler ve bu yapısı nedeniyle eğitilmesi ve yorumlanması kolay bir algoritmadır (Gümüştaş, 2019).

Karar ağaçları kolay görselleştirilebilmeleri ve anlaşılırlıkları sebebiyle en popüler makine öğrenimi algoritmalarından biridir. Karar ağaçları yapısı düğüm, dal ve yapraklardan oluşmaktadır. Ana düğüm (root) ağacın ilk elemanıdır ve iki veya daha fazla alt düğüme ayrılabilir. Bu alt düğüm karar düğüm veya yaprak düğüm olabilir. Karar düğüm, tahminde bulunur ve sınıflandırma yapar. Yaprak düğüm bir düğümün son elemanıdır ve yaprağa delindiğinde yaprağın temsil ettiği sınıf verinin sınıfıdır.

Veriler sınıflandırılmak istendiğinde en tepedeki kök düğümden başlayarak bir yaprak düğüme ulaşana kadar karar düğümlerdeki yönlendirmeye göre ilerler. Yaprağa ulaştığında yaprağın temsil ettiği sınıf verinin sınıfıdır (Amasyalı ve ark., 2006). Ağacın devam ettirilmesi belli kriterler göz önünde bulundurularak yapılmaktadır. Bu kriterlere göre seçim yapılmakta ve ağaç oluşturulmaktadır (Çelik, 2009). Şekil 4.3' te karar ağacının görünümü verilmiştir.

Kapsamlı veri hazırlığı gerektirmemesi, hem sayısal hem kategorik verilerde çalışması, ikili ve çoklu tahminlerde iyi çalışması, işleyişinin kolayca anlaşılması Karar Ağacı' nın avantajları arasındadır (Saihood, 2021).



Şekil 4.3 Karar ağacı görünümü (Gümüştaş, 2019)

4.2.2. K-En yakın komşu algoritması (K-Nearest Neighbors)

K-en yakın komşu algoritması (kNN) 1952 yılında Evelyn ve Joseph Hodges tarafından geliştirilen ve Thomas Cover tarafından genişletilen bir sınıflandırma yöntemidir (Fix ve Hodges, 1989) (Cover ve Hart, 1967). kNN makine öğrenmesi teknikleri arasında kullanılan en basit sınıflandırıcılardan bir tanesidir. Bu algoritmada verinin sınıfı belirlenirken bu veriye en yakın, sınıfı belli olan k adet komşusu tespit edilir. Bu komşuluk ilişkisinde baskın olan sınıf etiketini test verisinin sınıf etiketi olarak belirler. Dolayısıyla algoritma başarısını etkileyen en büyük etken k değerinin seçimidir. K sayısının belirlenmesine, geçmiş tecrübeler ve denemelere göre karar verilir (Sarıbacak, 2021). K sayısının küçük bir değer seçilmesi aşırı öğrenmeye sebep olurken k sayısının büyük bir değer seçilmesi genellemeye sebep olabilir (Pir, 2022). k sayısının çift olması bazı noktalarda yakın komşu sınıf etiketlerinin eşit sayıda olmasına sebep olabileceği için genellikle tek sayı seçilir. Literatürde k değerinin 1, 3, 5 değerlerini aldığında en iyi sonuçları verdiği görülmüştür (Başer ve ark., 2021). Bununla birlikte en iyi yol algoritmayı farklı k değerleriyle birkaç kez çalıştırmaktır ve kNN algoritması ile elde edilen en iyi sonuca göre k değerini seçmektir (Saihood, 2021).

k adet verinin uzaklık hesabı yapılırken birçok yöntem kullanılır. Bunlardan bazıları Öklid, Manhattan, Minkowski, Chebyschev, Dilca gibi yöntemlerdir. Bunlardan en sık kullanılan ölçüt Öklid uzaklık ölçütüdür (Taşcı ve Onan, 2016). P ve Q gibi herhangi iki nokta arasındaki Öklid uzaklık ($d_{öklid}$), $P = (x_1, x_2, \dots, x_n)$ ve $Q = (y_1, y_2, \dots, y_n)$ olmak üzere Denklem 4.9' daki gibi hesaplanır.

$$d_{\text{öklid}} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (4.9)$$

Öğrenme sürecinin basit ve hızlı olması, gürültülü eğitim verilerine karşı başarılı olması bu sınıflandırmanın en büyük avantajlarındandır (Bhatia, 2010). Veri setinin büyüklüğüne göre yüksek düzeyde bellek alanına gereksinim vardır. Bu durum özellikle fazla öznitelik içeren verilerde çalışmayı zorlaştırmaktadır. Ayrıca dışardan girilen k değerine ve belirlenen uzaklık yöntemine karşı olan hassasiyetinin performansı etkilemesi dezavantajları olarak görülebilir (Liu ve Zhang, 2012).

4.2.3. Destek vektör makineleri (DVM)

Destek vektör makineleri (DVM) sınıflandırma ve regresyon analizi problemlerinin çözümünde kullanılan istatiksel öğrenme teorisine dayanan bir denetimli öğrenme modelidir. Destek vektör makineleri, örüntü tanıma ve sınıflandırma problemlerinin çözümü için Vladimir Vapnik ve Alexey Yakovlevich Chervonenkis tarafından 1963 yılında ortaya atılmıştır (Cortes ve Vapnik, 1995; Sarıbacak, 2021).

DVM, ilk olarak doğrusal verilerin sınıflandırılmasında kullanılsa da sonraki yıllarda doğrusal olmayan verilerin sınıflandırılma problemlerinde de kullanılmıştır (Sarıbacak, 2021). DVM algoritması, temelde iki veri sınıfını birbirinden ayıran en iyi sınır veya hiper düzlemi bulmayı amaçlar. Bu hiper düzlemin geniş olması iki sınıfın birbirinden ayrılmasının daha kolay olmasını sağlar (Pir, 2022).

4.2.3.1. Doğrusal ayrılabilen DVM

Doğrusal ayrılabilen verileri sınıflandırırken bu iki sınıfı birbirinden ayıran doğrusal bir hiper düzlem vardır. Bu durumda DVM, en büyük sınıra sahip hiper düzlemi bulmayı amaçlar.

x girdi vektörü, w ağırlık vektörü ve b sapma olmak üzere hiper düzlem karar sınırları denklem 4.10’ da gösterildiği gibi bulunmaktadır.

$$wx + b = 0 \quad (4.10)$$

Marjini en büyükleme denklem 4.11 ile bulunmaktadır.

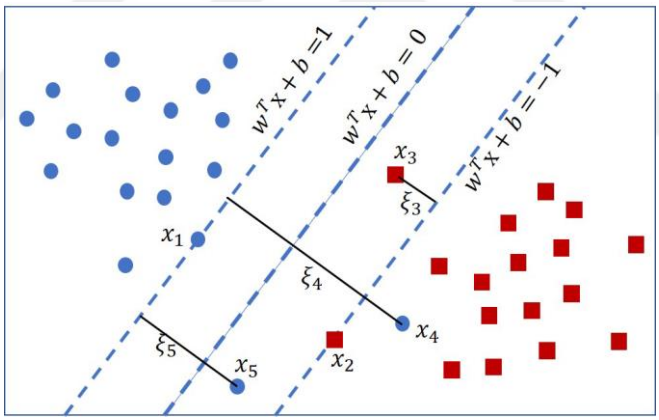
$$d = \frac{2}{\| w \|^2} \tag{4.11}$$

Doğrusal ayrımın gerçekleşebilmesi için tüm verilerin denklem 4.12' deki eşitliği sağlaması gerekir.

$$f(x_i) = \begin{cases} 1, eğer\ w.x_i + b \geq 1 \\ -1, eğer\ w.x_i + b \leq -1 \end{cases} \tag{4.12}$$

4.2.3.2. Belirli bir hata oranıyla doğrusal ayrılabilen DVM

Veri setinin gürültülü veri içermesi, çok boyutlu olması veya karmaşık bir yapıya sahip olması durumunda belirli bir hata oranıyla ayrılma durumu ortaya çıkabilmektedir. (Li ve ark., 2009), (Sarıbacak, 2021). Bu durumda iki sınıfı birbirinden ayırmak için gevşek sınır (soft margin) yöntemi kullanılmaktadır.



Şekil 4.4: Belirli bir hata oranıyla doğrusal ayrılma durumu (Le ve ark., 2018)

Gevşek sınır yönteminde Şekil 4.4' te görüleceği gibi bir örneğin yanlış sınıflandırılması durumunda ait olduğu karar sınırına olan uzaklığının ölçüsü olan ε aylak değişkeni eklenir (Cortes ve Vapnik, 1995). Bu durum için ayırma hiper düzleminin bulunabilmesi için veri setindeki tüm örneklerin Denklem 4.13'teki eşitsizlikleri sağlaması gerekir.

$$f(x_i) = \begin{cases} 1, eğer\ w.x_i + b \geq 1 - \varepsilon \\ -1, eğer\ w.x_i + b \leq -1 - \varepsilon \end{cases} \tag{4.13}$$

ε aylak değişkeninin seçimi marjini etkiler. Bu durumda ε ile marjin arasındaki dengenin sağlanabilmesi ve yanlış sınıflandırma ihtimalini düşürmek için denklem 4.14' te görüldüğü gibi bir kareli optimizasyon problemine dönüştürülür ve bir C parametresi kullanılır (Sarıbacak, 2021).

$$\phi(w) = \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \varepsilon_i \tag{4.14}$$

C, Lagrange çarpanının alabileceği üst sınırı ifade eden ceza parametresidir ve Langrange çarpanının üst sınırını gösteren ceza parametresini ifade eder.

$$y_i(wx_i + b) - 1 + \varepsilon \geq 0 \tag{4.15}$$

Lagrange fonksiyonunun kullanıldığı kareli optimizasyon probleminin çözümü sonucunda elde edilen hiper düzleme bağlı elde edilen sınıflandırıcı denklem 4.16' da gösterilmiştir.

$$f(x) = sgn\big((wx_i) + b\big) = sign(. \sum_{i=1}^{\iota} \big(y_i \alpha_i \big(x_i x_j\big)\big) \tag{4.16}$$

4.2.3.3. Doğrusal olmayan DVM

Çoğunlukla gerçek dünya problemlerinde verilerin doğrusal olması çok mümkün olmamaktadır. Bu nedenle sınıfları ayırma işlemi, ayırma işleminin tahmini ile mümkün olmaktadır. Bu gibi durumlarda doğrusal olmayan DVM algoritmaları kullanılmaktadır. Uygulamada bu ayırma işlemi çok zor olmaktadır (Ayhan ve Erdoğmuş, 2014). Bu durumda doğrusal olmayan bir çizgiye ihtiyaç duyulacağından DVM başka bir uzaya taşınarak tutarlı bir ayrım sağlanmaya çalışılır (Karakoyun ve Hacıbeyoğlu, 2014).

4.3. Değerlendirme Ölçütleri

Sınıflandırma sonuçlarının doğru değerlendirmesi için doğru değerlendirme ölçütünün seçimi önemli bir noktadır. Özellikle dengesiz verilerin sınıflandırılma sonuçlarının değerlendirilmesi için doğru ölçüt seçimi daha fazla önem kazanmaktadır. Sınıflandırma sonuçlarının değerlendirilmesinde genellikle doğruluk ölçütü

değerlendirme ölçütü olarak kullanılır. Ancak dengesiz veri setlerinde doğruluk ölçütü ile değerlendirme yanlış sonuçlara sebep olacaktır. Azınlık sınıfına ait verilerin hepsi yanlış da değerlendirilse çoğunluk sınıfına ait verilerin doğru tahmin edilmesi durumunda ölçüm sonucunun doğruluk oranı yine de yüksek olacaktır. Ancak bu ölçümün doğru olduğu anlamına gelmeyecektir. Sadece çoğunluk sınıfının sınıflandırıcılar tarafından daha iyi öğrenildiği anlamına gelebilir. Azınlık sınıfı ise sınıflandırıcılar tarafından göz ardı edilmiştir diyebiliriz.

Performans ölçümü, veri setine uygulanan sınıflandırıcının hem etkinliğini değerlendirmede hem de öğrenme sürecine rehberlik için temel göstergedir (Haixiang ve ark., 2017). Sınıflandırma sonuçlarının tek bir ölçüt ile değerlendirilmesi de eksik olacaktır. Farklı ölçütlerden alınan sonuçlara göre farklı değerlendirmeler yapılabilir ve öğrenme sürecine rehberlik sağlanabilir.

Sınıflandırma sonuçlarının değerlendirilmesinde çoğunlukla karmaşıklık matrisi kullanmaktadır. Karmaşıklık matrisi Çizelge 4.1.' de gösterilmiştir.

Çizelge 4.1: Karmaşıklık Matrisi (Confusion Matrix)

		Tahmin Edilen Değerler	
		Pozitif	Negatif
Gerçek Değerler	Pozitif	TP	FN
	Negatif	FP	TN

TP (True Positive) : Pozitif sınıfa ait doğru (pozitif) sınıflandırılan veri sayısı.
FP (False Positive) : Negatif sınıfa ait yanlış (pozitif) sınıflandırılan veri sayısı.
TN (True Negative) : Negatif sınıfa ait doğru (negatif) sınıflandırılan veri sayısı
FN (False Negative) : Pozitif sınıfa ait yanlış (negatif) sınıflandırılan veri sayısı

4.3.1. Doğruluk (Accuracy)

Sınıflandırma sonuçlarının değerlendirilmesinde en çok kullanılan ve basit yöntemlerden biridir (Coşkun ve Baykal, 2011). Sadece baskın olan sınıfa ait örneklemlerin doğru sınıflandırılması bile doğruluk oranının yüksek çıkmasına sebep olacağı için dengesiz veri setlerinde kullanılması sonuçların yanıltıcı olmasına sebep olacaktır (Gümüştaş, 2019). Sonuç olarak doğruluk, dengesiz veri setlerinde tek başına bir başarı ölçütü olmayacaktır (He ve Garcia, 2009). Denklem 4.17' de görüleceği üzere doğruluk doğru sınıflandırılan sonuçların tüm sonuçlara oranı ile hesaplanır.

$$Doğruluk = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \quad (4.17)$$

4.3.2. Kesinlik (Precision)

Kesinlik, gerçek pozitiflerin (TP), pozitif tahmin edilen örnek sayısına oranı olarak tanımlanır. Pozitif olarak tahmin edilen değerlerin yüzde kaçının doğru olduğunu hesaplar. Denklem 4.18 ile hesaplanmaktadır. Doğru tahmin edilen pozitif örnek sayısı arttıkça kesinlik değeri artarken, yanlış tahmin edilen negatif örnek sayısı arttıkça kesinlik değeri azalacaktır.

$$Kesinlik = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (4.18)$$

4.3.3. Duyarlılık (Recall)

Duyarlılık, pozitif olarak tahmin edilen örneklerin gerçek pozitiflere oranını hesaplar. Bir başka ifade ile gerçek pozitiflerin ne kadarının pozitif olarak işaretlendiğini gösterir. Doğru tahmin edilen pozitif örnek sayısı arttıkça duyarlılık artarken, yanlış tahmin edilen pozitif örnek sayısı arttıkça duyarlılık azalır. "Gerçek Pozitif Oranı (True Positive Rate - TPR)" olarak da bilinir. Duyarlılık Denklem 4.19' da gösterilen denklem ile hesaplanır.

$$Duyarlılık = \frac{TP}{TP + FN} \quad (4.19)$$

4.3.4. Özgüllük (Specificity)

Özgüllük, doğru tahmin edilen negatif örnek sayısının gerçek negatif örnek sayısına oranı ile hesaplanmaktadır. Gerçek negatiflerin ne kadarının doğru tahmin edildiğini göstermek için kullanılır. Doğru tahmin edilen negatif örnek sayısı arttıkça özgüllük değeri artarken yanlış tahmin edilen örneklem sayısı arttıkça özgüllük değer düşmektedir (Uyanık ve Kasapbaşı, 2021). Aynı zamanda "Gerçek Negatif Oran (True Negative Rate - TNR)" olarak da bilinir.

$$Özgüllük = \frac{TN}{TN + FP} \quad (4.20)$$

4.3.5. F1-Skor (F1-Score)

Kesinlik ve duyarlılık değerleri tek başına anlamlı bir değerlendirme için yeterli değildir. Bu iki ölçütün beraber değerlendirildiği F1-Skor, daha doğru sonuçlar için doğru bir değerlendirme ölçütüdür (Coşkun ve Baykal, 2011). F1-Skor kesinlik ve duyarlılık ölçütlerinin harmonik ortalaması alınarak hesaplanır. Değerlendirme sonuçları 0 ve 1 arasındadır ve bu değerin 1' e yaklaşması sınıflandırma başarısının iyi olduğu anlamına gelmektedir.

$$F1 - \text{Ölçütü} = 2\,x\,\frac{Kesinlik * Duyarlılık}{Kesinlik + Duyarlılık} \tag{4.21}$$

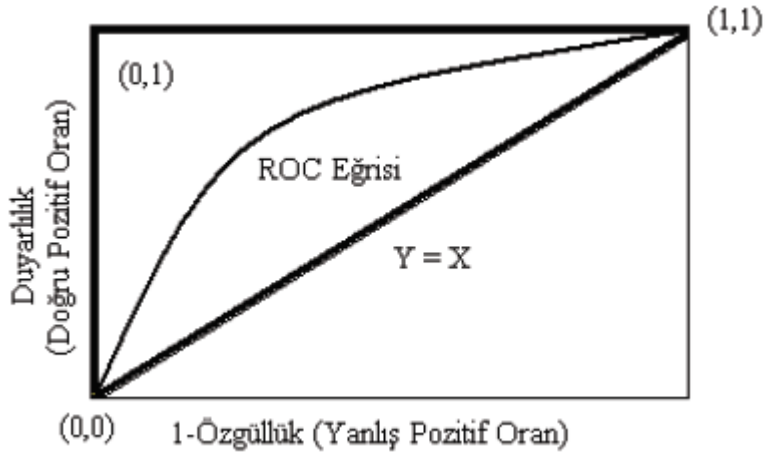
4.3.6. G-Ortalama (G-Mean)

G-ortalama ölçütü ile pozitif ve negatif sınıf performansları bir arada değerlendirilir. Bu değerlendirme sağlanırken geometrik ortalama kullanılır. G-ortalama değerinin yüksek çıkması hem pozitif sınıf doğru tahmin oranının hem de negatif sınıf doğru tahmin oranının yüksek olduğu anlamına gelmektedir.

$$G - Ortalama = \sqrt{\frac{TP}{TP + FN}\,X\,\frac{TN}{TN + FP}} \tag{4.22}$$

4.3.7. Eğri altında kalan alan (Area under the curve - AUC)

Eğri altında kalan alan (AUC) – Alıcı İşlem Karakteristiği (ROC) sınıflandırma problemlerinin performansını değerlendirmede özellikle de dengesiz veri setlerinin performansını değerlendirmede önemli bir ölçüttür. ROC eğrisi ikili sınıflandırma problemlerinde sıklıkla kullanılır. ROC eğrisi dikey y ekseninde doğru pozitiflerin (Duyarlılık - TPR), yatay x ekseninde yanlış pozitiflerin (Özgüllük - FPR) oranlarının bulunduğu bir eğridir. Bu oranlar yatay ve dikey düzlemde [0,1] arasında gösterilir (Bulut, 2016). ROC eğrisinde, eğri altında kalan alan AUC değerini verir (Şekil 4.5)



Şekil 4.5.: ROC eğrisi.

5. ARAŞTIRMA SONUÇLARI VE TARTIŞMA

Bu tez çalışmasında kullanılan veri seti 19 adet bal peteğinin iki yüzeyinden alınan 38 bal peteği görüntüsünden elde edilmiştir. Bal peteklerine ait görüntüler 5x5 piksel boyutunda parçalara bölünmüştür. Bu şekilde her petek görüntüsünden 26.448 adet resim parçası elde edilmiştir. Her parçadaki her bir piksel için sırasıyla R-G-B değerleri okunmuştur. 38 görüntü için aynı şekilde yapılan işlem sonucunda 1.005.024 satır ve 76 sütundan oluşan veri seti elde edilmiştir. Veri setinde 188.147 adet görüntü içinde larva bulunan kapalı hücreyi temsilen "1" olarak etiketlenmiştir. Geriye kalan 816.877 adet veri "0", yani içinde larva olan kapalı hücre dışındaki tüm hücreler olarak etiketlenmiştir.

Veri seti 5-Çapraz Doğrulama ile farklı alt kümelere bölünmüştür. Bölünme işlemi satırlardaki veriler rastgele seçilerek yapılmıştır. Rastgele seçilen alt kümelere ait indisler tüm sınıflandırma işlemlerinde kullanılmak üzere saklanmıştır.

Sentetik aşırı örnekleme algoritmaları olarak SMOTE, Borderline-SMOTE1, Borderline-SMOTE2, Safe-Level-SMOTE, DEBOHID kullanılmıştır. Sınıflandırma algoritmaları olarak kNN, Karar Ağacı ve DVM algoritmaları kullanılmıştır. Değerlendirme ölçütü olarak F1-Skor, G-Ortalama ve AUC metrikleri kullanılmıştır. Çalışmada bilgisayar olarak Macbook Air (M1, 2020, 16 GB RAM) kullanılmıştır. Tüm deneysel çalışmalarda adil bir karşılaştırma yapabilmek için MATLAB® R2019b yazılımı kullanılmıştır. MATLAB® yazılımında, DVM için fitcsvm; kNN için fitcknn ve KA için fitctree komutları varsayılan ayarları ile kullanılmıştır. İki adet temel parametre kullanan DEBOHID yaklaşımında, F parametresi 0,3 ve CR parametresi 0,6 olarak belirlenmiştir.

K-Çapraz doğrulamayla farklı alt kümeye bölünen veri setinin her alt kümesi için sınıflandırma sonuçları kaydedilmiş ve bu sınıflandırma sonuçlarının ortalamaları sınıflandırma sonucu olarak ele alınmıştır.

Gerçekleştirilen deneysel çalışmalar sonucunda aşırı örnekleme yaklaşımlarının G-Ortalamaya göre değerlendirmeleri Çizelge 5.1' de, F1-Skora göre sonuçları Çizelge 5.2' de, AUC' a göre sonuçları Çizelge 5.3' te karşılaştırmalı bir şekilde verilmiştir. Genel olarak sentetik veri üretimi ile sınıf dengesizliği giderildikten sonra sınıflandırma başarısının arttığı gözlemlenmiştir. Çizelgeler incelendiğinde SMOTE, Borderline-SMOTE2 ve DEBOHID aşırı örnekleme tekniklerinin ön plana çıktığı görülmektedir.

Çizelge 5.1: Tüm sınıflandırıcılar için G-Ortalama metriğine ait sonuçlar.

	ORJİNAL VERİ		SMOTE		BORDERLINE-SMOTE1		BORDERLINE-SMOTE2		SAFE-LEVEL-SMOTE		DEBOHID	
	Ort.	**Std.**	**Ort.**	**Std.**	**Ort.**	**Std.**	**Ort.**	**Std.**	**Ort.**	**Std.**	**Ort.**	**Std.**
kNN	0,9697	0,0004	**0,9747**	0,0003	0,9684	0,0003	0,9662	0,0003	0,9721	0,0004	0,9741	0,0003
KA	0,9226	0,0004	0,9654	0,0034	0,9551	0,0005	0,9607	0,0002	0,9622	0,0004	**0,9718**	0,0003
DVM	0,3394	0,0514	0,4391	0,1778	0,5219	0,1075	**0,5519**	0,1147	0,2710	0,0964	0,4813	0,0834
Ort.	0,7439		0,7931		0,8151		**0,8263**		0,7351		0,8091	

Çizelge 5.2: Tüm sınıflandırıcılar için F1-Skor metriğine ait sonuçlar.

	ORJİNAL VERİ		SMOTE		BORDERLINE-SMOTE1		BORDERLINE-SMOTE2		SAFE-LEVEL-SMOTE		DEBOHID	
	Ort.	**Std.**	**Ort.**	**Std.**	**Ort.**	**Std.**	**Ort.**	**Std.**	**Ort.**	**Std.**	**Ort.**	**Std.**
kNN	0,9236	0,0006	**0,9756**	0,0003	0,9698	0,0003	0,9678	0,0003	0,9731	0,0003	0,9708	0,0003
KA	0,8771	0,0006	0,9655	0,0035	0,9552	0,0005	0,9609	0,0002	0,9622	0,0004	**0,9691**	0,0003
DVM	0,1690	0,0493	0,3869	0,1829	0,5025	0,0650	**0,5309**	0,0544	0,2457	0,1315	0,4553	0,0294
Ort.	0,6566		0,7760		0,8092		**0,8199**		0,7270		0,7984	

Çizelge 5.3: Tüm sınıflandırıcılar için AUC metriğine ait sonuçlar.

	ORJİNAL VERİ		SMOTE		BORDERLINE-SMOTE1		BORDERLINE-SMOTE2		SAFE-LEVEL-SMOTE		DEBOHID	
	Ort.	**Std.**	**Ort.**	**Std.**	**Ort.**	**Std.**	**Ort.**	**Std.**	**Ort.**	**Std.**	**Ort.**	**Std.**
kNN	0,9697	0,0004	**0,9750**	0,0003	0,9689	0,0003	0,9668	0,0003	0,9724	0,0004	0,9743	0,0003
KA	0,9239	0,0004	0,9654	0,0034	0,9551	0,0005	0,9607	0,0002	0,9622	0,0004	**0,9718**	0,0003
DVM	0,4216	0,0754	0,5306	0,1112	0,5554	0,0956	**0,5805**	0,1119	0,3439	0,0924	0,5224	0,0984
Ort.	0,7717		0,8237		0,8265		**0,8360**		0,7595		0,8228	

Çizelge 5.4: Her bir sınıflandırıcının ilgili metrikteki Friedman test sonuçları

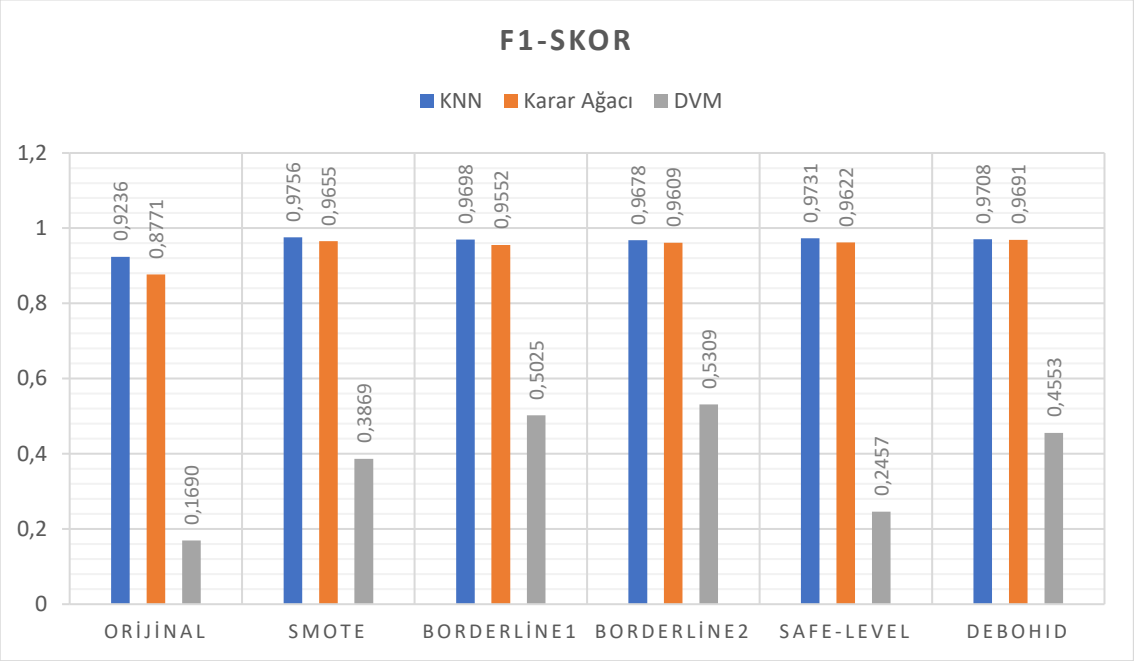
Metrikler	Sınıflandırıcılar	Orijinal	Smote	Border1	Border2	SafeLevel	Debohid	p-Değeri
G-Ort.	**kNN**	3,00	**5,80**	2,00	1,00	4,00	5,20	1,71E-04
	KA	1,00	5,00	2,00	3,00	4,00	**6,00**	1,39E-04
	DVM	2,20	4,20	4,60	**5,00**	1,40	3,60	1,34E-02
F1-Skor	**kNN**	1	**6**	3	2	5	4	1,39E-04
	KA	1	5,2	2	3	4	**5,8**	1,71E-04
	DVM	1,6	3,8	5	**5,4**	1,6	3,6	2,07E-03
AUC	**kNN**	2,8	**5,8**	2,2	1	4	5,2	2,09E-04
	KA	1	5	2	3	4	**6**	1,39E-04
	DVM	3	3,8	4	**5**	1,4	3,8	6,26E-02
Ortalama		1,84	**4,96**	2,98	3,16	3,27	4,80	8,78E-03
Kazanan/Toplam		0/9	**3/9**	0/9	**3/9**	0/9	**3/9**	

Çizelge 5.5: Her bir metrik için aşırı örnekleme yaklaşımı ve sınıflandırıcıların Friedman test sonucu.

Metrikler	O_kNN	O_KA	O_DVM	S_kNN	S_KA	S_DVM	B1_kNN	B1_KA	B1_DVM	B2_kNN	B2_KA	B2_DVM	SL_kNN	SL_KA	SL_DVM	D_kNN	D_KA	D_DVM	p-Değeri
G-Ort.	13,8	**17,8**	12,8	11,8	15,4	17,2	7	11,8	8	9	10	15,4	2,2	4,2	4,6	5	1,4	3,6	1,142E-10
F1-Skor	8	7	1,6	**18**	12,8	3,8	14,8	9	5	12,8	10	5,4	17	11	1,6	15,8	13,8	3,6	8,452E-11
AUC	13,6	7	3	**17,8**	11,8	3,8	13	8	4	11,8	9	5	15,6	10	1,4	17,2	15,2	3,8	1,414E-10
Ortalama	11,80	10,60	5,80	**15,87**	13,33	8,27	11,60	9,60	5,67	11,20	9,67	8,60	11,60	8,40	2,53	12,67	10,13	3,67	1,13E-10

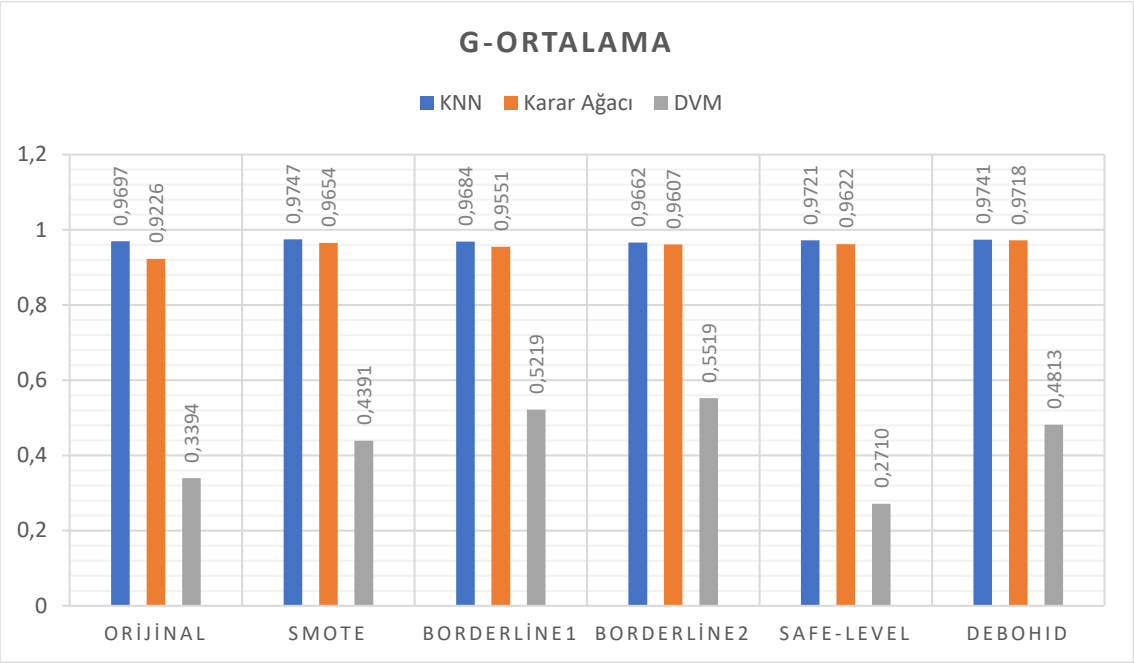
Friedman testi (Friedman, 1940), deneysel sonuçları analiz etmek için sıklıkla kullanılan parametrik olmayan istatistiksel bir testtir. İlişkili grupların olduğu durumlarda ve birden fazla sınıflandırıcının sonuçlarının değerlendirildiği durumlar için uygundur. Bu çalışmada, aşırı örnekleme yöntemleri ile ön işleme tutulan verilerin sınıflandırılması ile elde edilen sonuçlar Friedman testine tabi tutulmuştur.

Kullanılan sınıflandırıcılar bazında yapılan Friedman testi sonuçları Çizelge 5.4' te görülmektedir. Friedman testine göre sonuçlar incelendiğinde SMOTE, Bordeline-SMOTE2 ve DEBOHID yöntemlerinin etkinliği burada da görülmektir. Çizelge 5.5' te aşırı örnekleme yaklaşımı ve sınıflandırıcılar bazında Friedman Testi sonuçları görülmektedir.



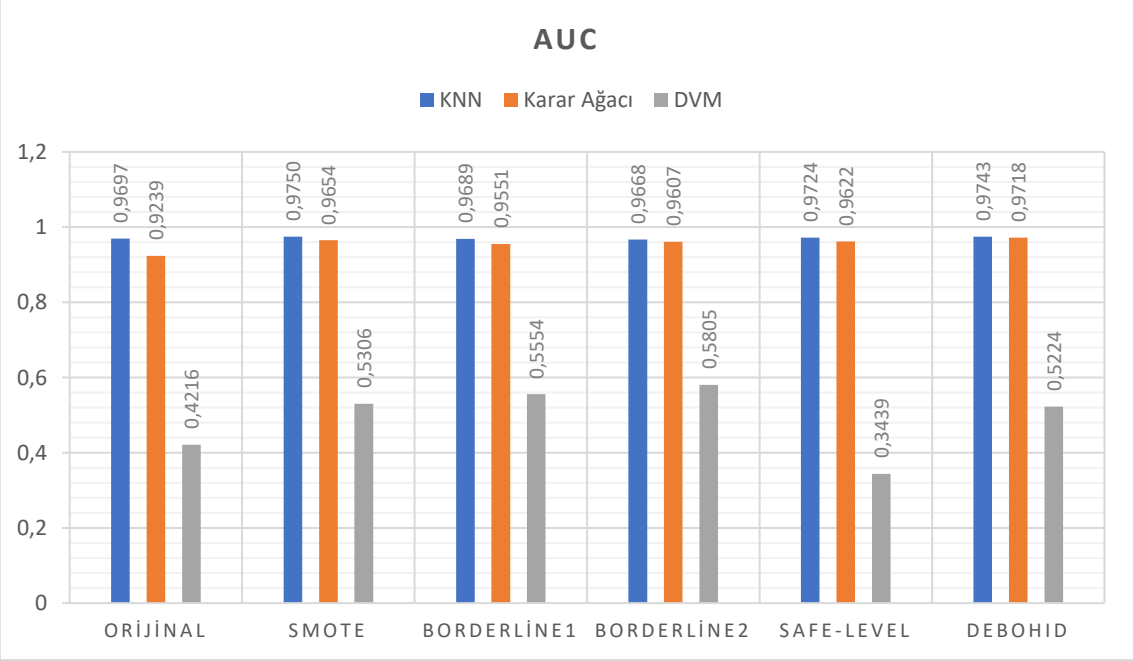
Şekil 5.1.: Tüm metotlar için F1 Skor ölçütüne göre sonuçlar

Tüm metotlar için F1-Skora dayalı değerlendirme sonuçları Şekil 5.1' de görülmektedir. kNN için en başarılı sınıflandırma sonucunun SMOTE yöntemi ile, Karar Ağacı için en başarılı sonucun DEBOHID, DVM için en başarılı sonucun Borderline-SMOTE2 yöntemi ile elde edildiği görülmektedir.



Şekil 5.2.: Tüm metotlar için G-Ortalama ölçütüne göre sonuçlar

Tüm metotlar için G-Ortalama ölçütü sonuçları Şekil 5.2' de görülmektedir. G-Ortalama ile yapılan sınıflandırma sonuçlarında kNN için en başarılı yöntemin SMOTE, Karar Ağacı için en başarılı yöntemin DEBOHID, DVM için en başarılı yöntemin Borderline-SMOTE2 olduğu görülmektedir.



Şekil 5.3.: Tüm metotlar için AUC ölçütüne göre sonuçları

Şekil 5.3’ te görülen tüm metotlar için AUC ölçütüne göre sonuçlar değerlendirildiğinde yine kNN için en başarılı sınıflandırma sonucunun SMOTE, Karar Ağacı sınıflandırma yöntemi için en başarılı sonucun DEBOHID, DVM için en başarılı sınıflandırma sonucunun Borderline-SMOTE2 yönteminde ortaya çıktığı görülmektedir.

kNN sınıflandırma algoritması değerlendirildiğin de F1-Skora göre tüm sentetik veri üretim yöntemlerinde sınıflandırma başarısının artmış olduğu görülmektedir. En yüksek sınıflandırma başarısının da %97,56 ile SMOTE yönteminde olduğu görülmektedir. En düşük başarıya sahip sentetik veri üretme yönteminin %96,78 başarı ile Borderline-SMOTE2 olduğu görülmesine rağmen bu yöntemde de gerçek veriye göre sınıflandırma başarısının arttığı görülmektedir. G-Ortalamada ölçütüne göre değerlendirildiğinde Borderline-SMOTE1 ve Borderline-SMOTE2 yöntemlerinde gerçek veriye göre sınıflandırma başarısının çok az olsa da düştüğü görülmektedir. G-Ortalama ölçütüne göre en yüksek sınıflandırma oranı %97.47 ile SMOTE yönteminde olmuştur. DEBOHID ve Safe-Level-SMOTE algoritmalarında da ciddi bir değişim görülmektedir. AUC ölçütüne göre değerlendirildiğin yine Borderline-SMOTE1 ve Borderline-SMOTE2 algoritmalarında başarı oranında düşüş görülmektedir. En yüksek başarıyı ise %97,50 ile SMOTE algoritması sağlamıştır.

Karar Ağacı sınıflandırma algoritmasına göre değerlendirildiğinde tüm sentetik veri üretme yöntemlerinde, tüm değerlendirme sonuçlarına göre başarı düzeyinin arttığı görülmektedir. F1-Skora göre değerlendirildiğinde gerçek veride %87,71 olan başarı oranı %96,91 ile DEBOHID algoritmasında en başarılı duruma ulaşmıştır. G-Ortalama ölçütüne göre değerlendirildiğinde en yüksek başarıyı %97,18 ile DEBOHID algoritmasının gösterdiği görülmektedir. AUC ölçütüne göre yine en iyi başarıyı %97,18 ile DEBOHID algoritması sağlamıştır. Kısaca Karar ağacı sınıflandırma algoritması ile değerlendirildiğinde en iyi sınıflandırma başarısını DEBOHID yöntemi sağlamıştır.

DVM sınıflandırması algoritması ile değerlendirildiğinde F1-Skora göre en yüksek başarı düzeyini %55,09 ile Borderline-SMOTE2 yönteminin gerçekleştirdiği görülmektedir. SMOTE yönteminin de sınıflandırma başarısını arttırdığı görülse de %38,69 ile diğer yöntemlerin gerisinde kaldığı görülmektedir. G-Ortalama ve AUC metriklerine göre değerlendirildiğinde sırasıyla %55.19 ve %58.05 başarı oranı ile en yüksek başarının Borderline-SMOTE2 algoritması ile sağlandığı görülmektedir. G-Ortalama ve AUC ile değerlendirildiğinde Safe-Level-SMOTE algoritmasında başarı düzeyinin düştüğü görülmektedir.

Genel olarak değerlendirildiğinde kNN sınıflandırma yöntemi ile yapılan sınıflandırmalarda SMOTE algoritmasının ön plana çıktığı görülmektedir. Borderline-SMOTE algoritmalarının geride kaldığı görülmektedir. Karar Ağacı sınıflandırma yöntemine göre tüm ölçütlerde DEBOHID yönteminin başarı sağladığı görülmektedir. Tüm sentetik veri üretme yöntemlerinde başarı sağladığı görülse de Borderline-SMOTE algoritmalarının az bir oranla diğer algoritmaların gerisinde kaldığı görülmektedir. DVM algoritmasına göre değerlendirildiğinde Borderline-SMOTE2 algoritmasının tüm ölçütler için en yüksek başarıyı sağladığı görülmektedir. Safe-Level-SMOTE yönteminin ise mevcut veri setinde diğer yaklaşımlara göre daha az başarılı olduğu görülmektedir.

6. SONUÇLAR VE ÖNERİLER

6.1. Sonuçlar

Arıcılık dünya ve ülkemiz için önemli bir tarımsal faaliyet ve gelir kaynağıdır. Arıcılık, gelir kaynağı olması açısından kırsal kalkınmaya etkisi ile sosyo-ekonomik anlamda önemli bir yere sahiptir. Arıcılık faaliyetleri sonucunda ana çıktı bal olmaktadır. Bunun yanı sıra hem besin kaynağı olan hem ekonomik değeri bulunan hem de sağlık gibi farklı alanlarda kullanılabilen arı sütü, polen ve propolis gibi yan ürünlerin de üretilmesi önem arz etmektedir. Arıcılık faaliyetleri ile ekosistemdeki tozlaşma da önemli ölçüde sağlanmaktadır. Bu durum da biyolojik hayatın devam ettirilmesi hususunda ciddi katkı sağlamaktadır.

Arıcılık faaliyetleri ile bal üretimi hem ülkemizde hem de dünyada önemli bir yere sahiptir. Gıda üretiminde, gıda güvenliği ve üretimdeki sürekliliğin korunabilmesi iki önemli başlık olarak ele alınmaktadır. Bal üretimi meşakkatli birçok süreci içinde barındırmaktadır. Var olan süreçler doğru bir şekilde ele alınmadığında mevcut kaynaklar uygun kullanılamayacaktır. Bu da hem üretimdeki verim kayıplarını hem de ilerleyen zamanlarda ortaya çıkabilecek üretimdeki süreklilik problemlerini ortaya çıkaracaktır. Türkiye'de yeterli sayıda arı kovanı bulunmaktadır, fakat diğer ülkeler ile kıyaslandığında ciddi bir verim problemi olduğu ortadadır. Bu noktada üzerinde durulacak konu üretim odaklı anlayışın bir kenara bırakılıp verimlilik odaklı anlayışın benimsenmesidir. Arıcılık faaliyetinin her safhası için teknolojik gelişmelerde göz önünde bulundurularak verimliliği arttırmak için iktiza eden önlemlerin alınması gerekmektedir.

Her alanda olduğu gibi gıda üretiminde de teknolojik gelişmeler fazlasıyla yaşanmaktadır. Bu kapsamda bal hasadı aşamasında koloninin sürdürülebilirliği ve bal üretiminde yüksek düzeyde verim sağlama hedefi bu tez çalışması için ilham kaynağı olmuştur. Bal hasadı sürecinde, üzerinde arı larvası olan petekler yavru kaybının olması veya bu yavruların bala karışarak bal homojenliğinin bozulması gibi kaygılarla bal süzme aşamasında değerlendirilememektedirler. Lakin, hasat aşamasına gelen petekler incelendiğinde içinde yavru bulunan alanların, bal içeren alana göre daha az yer kapladığı görülmektedir. Bu peteklerin hasat sürecinde değerlendirilememesi uzun vadede verim kaybına sebep olmaktadır.

Bu çalışmada, bal peteğindeki yavru arılara ait alanların tespiti edilmesi bir sınıflandırma problemi olarak ele alınmıştır. Çalışmada endüstriyel kamera ile elde

edilmiş 38 farklı bal peteği görüntüsü kullanılarak bir veri seti elde edilmiştir. Elde edilen görüntüler incelendiğinde içinde yavru bulunan alanların diğer alanlara göre oranla daha az (yaklaşık 1/5 oranında) olduğu görülmüştür. Bu durum veri setindeki sınıflar arasında dengesizlik problemi yaratmaktadır. Tez çalışmasında oluşturulmuş olan dengesiz veri seti, veri düzeyinde aşırı örnekleme yaklaşımları (SMOTE, Borderline-SMOTE1, Borderline-SMOTE2, Safe-Level SMOTE ve DEBOHID) ile dengeli hale getirilmiştir. Bunun yanı sıra farklı sınıflandırıcılar (kNN, Karar Ağacı ve Destek Vektör Makineleri) ve metrikler (F1-Skor, G-Ortalama ve AUC) kullanılarak hem sınıflandırıcıların hem de aşırı örnekleme yaklaşımlarının bu veri setindeki performansı detaylı olarak kıyaslanmıştır. Yapılan deneysel çalışmalarda sentetik veri üretme yöntemleri kullanılarak dengeli hale getirilen veri setindeki sınıflandırma başarısının genel olarak arttığı görülmektedir.

kNN sınıflandırma algoritması ile elde edilen sonuçlarda SMOTE, Safe-Level-SMOTE ve DEBOHID yöntemlerinin başarısı göze çarpmaktadır. Karar Ağacı sınıflandırma algoritması ile yapılan sınıflandırma sonucunda tüm metriklerde başarının arttığı, DEBOHID yönteminin diğer yöntemlere göre daha iyi olduğu görülmektedir. DVM algoritması ile gerçekleştirilen sınıflandırma sonuçlarında Borderline-SMOTE2 yönteminin ön plana çıktığı görülmüştür.

6.2. Öneriler

Bu çalışmada yalnızca aşırı örnekleme yaklaşımları kullanılıp sentetik veri üretilmiş ve veri seti dengeli hale getirilerek deneysel çalışmalar gerçekleştirilmiştir. Sonraki çalışmalarda, bu veri setinde, yine veri düzeyinde az örnekleme yaklaşımları kullanılabilir. Ayrıca veri düzeyi yaklaşımlarına odaklanmak yerine algoritmik düzey veya melez yaklaşımlarda göz önünde bulundurulabilir.

Yapılan çalışmada bal peteği üzerindeki alanlar iki sınıf olacak şekilde etiketlenmiştir. Gelecek çalışmalarda petekler üzerindeki alanlar çok sınıflı olacak şekilde etiketlenebilir. Çalışma farklı sınıflandırma algoritmalarını da kapsayacak şekilde genişletilebilir.

KAYNAKLAR

Alhakbani, H., 2019, Handling class imbalance using swarm intelligence techniques, hybrid data and algorithmic level solutions, *Goldsmiths, University of London*.

Alves, T. S., Pinto, M. A., Ventura, P., Neves, C. J., Biron, D. G., Junior, A. C., De Paula Filho, P. L. ve Rodrigues, P. J., 2020, Automatic detection and classification of honey bee comb cells using deep learning, *Computers and electronics in agriculture*, 170, 105244.

Amasyalı, M. F., Diri, B. ve Türkoğlu, F., 2006, Farklı özellik vektörleri ile Türkçe dokümanların yazarlarının belirlenmesi, *15th Turkish Symposium on Artificial Intelligence and Neural Networks*.

Aydilek, İ. B., 2018, Yazılım hata tahmininde kullanılan metriklerin karar ağaçlarındaki bilgi kazançlarının incelenmesi ve iyileştirilmesi, *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, 24 (5), 906-914.

Aydin Hakli, D., 2018, Sınıf dengesizliği sorununu çözmek için kullanılan algoritmaların farklı sınıflandırma yöntemlerinde performanslarının karşılaştırılması.

Ayhan, S. ve Erdoğmuş, Ş., 2014, Destek vektör makineleriyle sınıflandırma problemlerinin çözümü için çekirdek fonksiyonu seçimi, *Eskişehir Osmangazi Üniversitesi İktisadi ve İdari Bilimler Dergisi*, 9 (1), 175-201.

Başer, B. Ö., Yangın, M. ve Sarıdaş, E. S., 2021, Makine öğrenmesi teknikleriyle diyabet hastalığının sınıflandırılması, *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 25 (1), 112-120.

Bhatia, N., 2010, Survey of nearest neighbor techniques, *arXiv preprint arXiv:1007.0085*.

Bulut, F., 2016, Dengesiz veri setlerinde denetimli öğrenicilerin başarım değerlendirmesi.

Bunkhumpornpat, C., Sinapiromsaran, K. ve Lursinsap, C., 2009, Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem, *Pacific-Asia conference on knowledge discovery and data mining*, 475-482.

Burucu, V. ve Gülse Bal, H. S., 2017, Türkiye'de arıcılığın mevcut durumu ve bal üretim öngörüsü, *Tarım ekonomisi araştırmaları dergisi*, 3 (1), 28-37.

Burucu, V. ve Gülse Bal, H. S., 2018, Arıcılık İşletmelerinin Pazarlama Olanakları: Kastamonu İli Azdavay İlçesi Örneği, *Tarım ekonomisi araştırmaları dergisi*, 4 (1), 23-35.

Cao, L. ve Zhai, Y., 2015, Imbalanced data classification based on a hybrid resampling svm method, *2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom)*, 1533-1536.

Cao, P., Zhao, D. ve Zaiane, O., 2013, An optimized cost-sensitive SVM for imbalanced data learning, *Pacific-Asia conference on knowledge discovery and data mining*, 280-292.

Cateni, S., Colla, V. ve Vannucci, M., 2014, A method for resampling imbalanced datasets in binary classification tasks for real-world problems, *Neurocomputing*, 135, 32-41.

Chawla, N., Japkowicz, N. ve Kolcz, A., 2003, Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Data Sets.

Chawla, N. V., Bowyer, K. W., Hall, L. O. ve Kegelmeyer, W. P., 2002, SMOTE: synthetic minority over-sampling technique, *Journal of artificial intelligence research*, 16, 321-357.

Chawla, N. V., Japkowicz, N. ve Kotcz, A., 2004, Special issue on learning from imbalanced data sets, *ACM SIGKDD explorations newsletter*, 6 (1), 1-6.

Chen, Z., Duan, J., Kang, L. ve Qiu, G., 2021, A hybrid data-level ensemble to enable learning from highly imbalanced dataset, *Information Sciences*, 554, 157-176.

Cieslak, D. A., Chawla, N. V. ve Striegel, A., 2006, Combating imbalance in network intrusion datasets, *GrC*, 732-737.

Cortes, C. ve Vapnik, V., 1995, Support-vector networks, *Machine learning*, 20 (3), 273-297.

Coşkun, C. ve Baykal, A., 2011, Veri madenciliğinde sınıflandırma algoritmalarının bir örnek üzerinde karşılaştırılması, *Akademik Bilişim*, 2011, 1-8.

Cover, T. ve Hart, P., 1967, Nearest neighbor pattern classification, *IEEE transactions on information theory*, 13 (1), 21-27.

Çelik, M., 2009, Veri madenciliğinde kullanılan sınıflandırma yöntemleri ve bir uygulama, *Yayımlanmamış Yüksek Lisans Tezi, İstanbul Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul*.

Çürükoğlu, N., 2019, Imbalanced Dataset Problem in Classification Algorithms, *2019 1st International Informatics and Software Engineering Conference (UBMYK)*, 1-5.

Dhar, S. ve Cherkassky, V., 2014, Development and evaluation of cost-sensitive universum-SVM, *IEEE transactions on cybernetics*, 45 (4), 806-818.

Farahmand, M., 2022, Bal peteğindeki hücrelerin tespit edilmesi için derin öğrenme yaklaşımlarının kullanılması, *Selçuk Üniversitesi Fen Bilimleri Enstitüsü*.

Fernández, A., Garcia, S., Herrera, F. ve Chawla, N. V., 2018, SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary, *Journal of artificial intelligence research*, 61, 863-905.

Fix, E. ve Hodges, J. L., 1989, Discriminatory analysis. Nonparametric discrimination: Consistency properties, *International Statistical Review/Revue Internationale de Statistique*, 57 (3), 238-247.

Friedman, M., 1940, A comparison of alternative tests of significance for the problem of m rankings, *The Annals of Mathematical Statistics*, 11 (1), 86-92.

Gao, M., Hong, X., Chen, S. ve Harris, C. J., 2011, A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems, *Neurocomputing*, 74 (17), 3456-3466.

García, S. ve Herrera, F., 2009, Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy, *Evolutionary computation*, 17 (3), 275-306.

Genç, F., Aksakal, V., Göktaş, B., Cengiz, M. M., Memiş, S., Korucuk, S., Erdoğan, Ü., Erdoğan, Y. ve Yılmaz, Y. Y., 2020, Arıcılık Üzerine Bilimsel Araştırmalar.

Grzymala-Busse, J. W., Stefanowski, J. ve Wilk, S., 2005, A comparison of two approaches to data mining from imbalanced data, *Journal of Intelligent Manufacturing*, 16 (6), 565-573.

Gümüştaş, E., 2019, Kayıp gözlem içeren dengesiz veri setlerinin topluluk öğrenme algoritmaları ile sınıflandırılması.

Güngörmüş, A., 2020, Görüntü işleme teknikleri kullanarak petek üzerindeki arı larvasının konumunun ve özelliklerinin tespiti, *Balıkesir Üniversitesi Fen Bilimleri Enstitüsü*.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H. ve Bing, G., 2017, Learning from class-imbalanced data: Review of methods and applications, *Expert Systems with Applications*, 73, 220-239.

Han, H., Wang, W.-Y. ve Mao, B.-H., 2005, Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning, *International conference on intelligent computing*, 878-887.

Hand, D. J., 2005, Good practice in retail credit scorecard assessment, *Journal of the Operational Research Society*, 56 (9), 1109-1117.

He, H., Bai, Y., Garcia, E. A. ve Li, S., 2008, ADASYN: Adaptive synthetic sampling approach for imbalanced learning, *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, 1322-1328.

He, H. ve Garcia, E. A., 2009, Learning from imbalanced data, *IEEE Transactions on knowledge and data engineering*, 21 (9), 1263-1284.

Japkowicz, N., 2000, Learning from imbalanced data sets: a comparison of various strategies, *AAAI workshop on learning from imbalanced data sets*, 10-15.

Karakoyun, M. ve Hacıbeyoğlu, M., 2014, Biyomedikal veri kümeleri ile makine öğrenmesi siniflandirma algoritmalarinin istatistiksel olarak karşilaştirilmasi, *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi*, 16 (48), 30-42.

Karlıdağ, S. ve Köseman, A., 2015, Türkiye ve Malatya'da arıcılığın yeri ve önemi, *Araştırma*.

Kaur, H., Pannu, H. S. ve Malhi, A. K., 2019, A systematic review on imbalanced data challenges in machine learning: Applications and solutions, *ACM Computing Surveys (CSUR)*, 52 (4), 1-36.

Kaya, E., Korkmaz, S., Sahman, M. A. ve Cinar, A. C., 2021, DEBOHID: A differential evolution based oversampling approach for highly imbalanced datasets, *Expert Systems with Applications*, 169, 114482.

Kekeçoğlu, M., Gürcan, E. K. ve Soysal, M. İ., 2007, Türkiye arı yetiştiriciliğinin bal üretimi bakımından durumu, *Tekirdağ Ziraat Fakültesi Dergisi*, 4 (2), 227-236.

Krawczyk, B., 2016, Learning from imbalanced data: open challenges and future directions, *Progress in Artificial Intelligence*, 5 (4), 221-232.

Kubat, M., Holte, R. C. ve Matwin, S., 1998, Machine learning for the detection of oil spills in satellite radar images, *Machine learning*, 30 (2), 195-215.

Langstroth, L. L., 1857, A Practical Treatise on the Hive and Honey-bee, CM Saxton & Company, p.

Le, H. M., Tran, T. D. ve Van Tran, L., 2018, Automatic heart disease prediction using feature selection and data mining technique, *Journal of Computer Science and Cybernetics*, 34 (1), 33-48.

Li, S., Li, H., Li, M., Shyr, Y., Xie, L. ve Li, Y., 2009, Improved prediction of lysine acetylation by support vector machines, *Protein and peptide letters*, 16 (8), 977-983.

Li, Y., Sun, G. ve Zhu, Y., 2010, Data imbalance problem in text classification, *2010 Third International Symposium on Information Processing*, 301-305.

Liu, H. ve Zhang, S., 2012, Noisy data elimination using mutual k-nearest neighbor for classification mining, *Journal of Systems and Software*, 85 (5), 1067-1074.

Mazurowski, M. A., Habas, P. A., Zurada, J. M., Lo, J. Y., Baker, J. A. ve Tourassi, G. D., 2008, Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance, *Neural networks*, 21 (2-3), 427-436.

Pir, M. Ş., 2022, Dengesiz veri setlerinde sınıflandırma problemlerinin çözümünde melez yöntem uygulaması, *Bursa Uludağ Üniversitesi*.

Prati, R. C., Batista, G. E. ve Monard, M. C., 2009, Data mining with imbalanced class distributions: concepts and methods, *IICAI*, 359-376.

Pristyanto, Y., Pratama, I. ve Nugraha, A. F., 2018, Data level approach for imbalanced class handling on educational data mining multiclass classification, *2018 International Conference on Information and Communications Technology (ICOIACT)*, 310-314.

Qiu, C., Jiang, L. ve Li, C., 2017, Randomly selected decision tree for test-cost sensitive learning, *Applied Soft Computing*, 53, 27-33.

Sağlam, F., Sözen, M. ve Cengiz, M. A., 2021, Optimization Based Undersampling for Imbalanced Classes, *Adıyaman University Journal of Science*, 11 (2), 385-409.

Sahare, M. ve Gupta, H., 2012, A review of multi-class classification for imbalanced data, *International Journal of Advanced Computer Research*, 2 (3), 160.

Saihood, Q. L., 2021, Exploratıon Of Machıne Learnıng Technıques In Predıctıng The Chıldhood Anemıa.

Sarıbacak, B., 2021, Makine öğrenmesi sınıflandırma yöntemleri ile hematoloji hastalıklarından demir eksikliği anemisinin erken teşhis edilmesi.

Semerci, A., 2017, Türkiye arıcılığının genel durumu ve geleceğe yönelik beklentiler, *Mustafa Kemal Üniversitesi Ziraat Fakültesi Dergisi*, 22 (2), 107-118.

Sowah, R. A., Agebure, M. A., Mills, G. A., Koumadi, K. M. ve Fiawoo, S. Y., 2016, New cluster undersampling technique for class imbalance learning, *International Journal of Machine Learning and Computing*, 6 (3), 205-214.

Sparavigna, A. C., 2016, Analysis of a natural honeycomb by means of an image segmentation, *Philica*, 2016 (897).

Storn, R. ve Price, K., 1997, Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces, *Journal of global optimization*, 11 (4), 341-359.

Sun, Y., Wong, A. K. ve Kamel, M. S., 2009, Classification of imbalanced data: A review, *International journal of pattern recognition and artificial intelligence*, 23 (04), 687-719.

Şeker, İ., Köseman, A., Karlıdağ, S. ve Aygen, S., 2017, Arıcılık faaliyetleri II: Malatya ilinde arıcılık faaliyetlerinin yetiştirici tercihleri, üretim nitelikleri ve arı hastalıkları kapsamında değerlendirilmesi, *Tekirdağ Ziraat Fakültesi Dergisi*.

Tahir, M. A., Kittler, J. ve Yan, F., 2012, Inverse random under sampling for class imbalance problem and its application to multi-label classification, *Pattern Recognition*, 45 (10), 3738-3750.

Tang, Y., Zhang, Y.-Q., Chawla, N. V. ve Krasser, S., 2008, SVMs modeling for highly imbalanced classification, *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39 (1), 281-288.

Taşcı, E. ve Onan, A., 2016, K-en yakın komşu algoritması parametrelerinin sınıflandırma performansı üzerine etkisinin incelenmesi, *Akademik Bilişim*, 1 (1), 4-18.

Topal, A. ve Amasyalı, M. F., 2021, Yapay Örnek Üretimi Ne Zaman İşe Yarar? When does Synthetic Data Generation Work?, *Conference: 29. IEEE Conference on Signal Processing and Communications*.

Uyanık, F. ve Kasapbaşı, M. C., 2021, Telekomünikasyon sektörü için veri madenciliği ve makine öğrenmesi teknikleri ile ayrılan müşteri analizi, *Düzce Üniversitesi Bilim ve Teknoloji Dergisi*, 9 (3), 172-191.

Van Hulse, J., Khoshgoftaar, T. M. ve Napolitano, A., 2007, Experimental perspectives on learning from imbalanced data, *Proceedings of the 24th international conference on Machine learning*, 935-942.

Wang, S. ve Yao, X., 2013, Using class imbalance learning for software defect prediction, *IEEE Transactions on Reliability*, 62 (2), 434-443.

Yavaş, M., Güran, A. ve Uysal, M., 2020, Covid-19 veri kümesinin SMOTE tabanlı örnekleme yöntemi uygulanarak sınıflandırılması, *Avrupa Bilim ve Teknoloji Dergisi*, 258-264.

Zareapoor, M. ve Yang, J., 2017, A novel strategy for mining highly imbalanced data in credit card transactions, *Intelligent Automation & Soft Computing*, 1-7.

Zhang, D., Ma, J., Yi, J., Niu, X. ve Xu, X., 2015, An ensemble method for unbalanced sentiment classification, *2015 11th international conference on natural computation (ICNC)*, 440-445.

Zhang, Y.-P., Zhang, L.-N. ve Wang, Y.-C., 2010, Cluster-based majority under-sampling approaches for class imbalance learning, *2010 2nd IEEE International Conference on Information and Financial Engineering*, 400-404.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı	**:**	Serkan ÖZGÜN
Uyruğu	**:**	Türkiye

EĞİTİM

Derece		**Adı, İlçe, İl**	**Bitirme Yılı**
Lise	:	Selçuklu Lisesi	2004
Üniversite	:	Selçuk Üniversitesi	2010
Yüksek Lisans	:	Selçuk Üniversitesi	Devam ediyor.

İŞ DENEYİMLERİ

Yıl	**Kurum**	**Görevi**
2012-Devam Ediyor.	Milli Eğitim Bakanlığı	Bilişim Teknolojileri Öğretmeni

UZMANLIK ALANI
Görüntü İşleme, Veri Dengeleme Yöntemleri
YABANCI DİLLER
İngilizce
BELİRTMEK İSTEĞİNİZ DİĞER ÖZELLİKLER
-
YAYINLAR

Özgün, S. ve Şahman, M. A., 2021, Boosting The Classification Success Of Bee Larva Cells In The Imbalanced Dataset, *International Symposiumon Implementations of Digital Industry and Management of Digital, pp:17, 10-11 Kasım 2021, Konya*