#### ****Abstract Kısmında Yapılanlar****

* Bu çalışmada, sınıflandırma doğruluğunu artırmak için k-NN yakınlık grafikleri kullanılarak bir veri filtreleme yöntemi önerildi.
* Gürültülü ve gereksiz özelliklerin çıkarılmasının sınıflandırıcı performansına etkisi araştırıldı.
* Filtreleme parametreleri, aykırı değerlerin en aza indirilmesi problemi olarak ele alındı.
* 6 farklı sınıflandırıcı ve bir heterojen kombiner (DES-LA) ile filtreleme öncesi ve sonrası karşılaştırma yapıldı.
* Sonuçlar, düşük doğruluklu veri setlerinde filtrelemenin faydalı olduğunu, ancak yüksek doğruluklu veri setlerinde etkisinin sınırlı olduğunu gösterdi.

#### ****Keywords (Anahtar Kelimeler):****

1. **Binary Classification:** İki sınıflı sınıflandırma problemleri.
2. **Heterogeneous Combiner:** Farklı performans seviyelerine sahip sınıflandırıcıların birleşimi.
3. **k-NN (k-nearest neighbor):** Veriye en yakın k komşuyu bulma algoritması.
4. **Proximity Graphs:** Noktalar arasındaki mesafeleri temel alan grafik yapıları.
5. **Data Filtering:** Gürültü ve aykırı değerlerin çıkarılması.
6. **Feature Selection:** Önemli özelliklerin seçilmesi.

#### ****Introduction Kısmında Yapılanların Detaylı Açıklamaları:****

1. Sınıflandırıcı performansları, veri setindeki gürültü ve gereksiz özelliklerden etkilenir.
2. k-NN grafikleri kullanılarak, sınıflandırıcılar için ortak bir veri filtreleme yöntemi geliştirildi.
3. Sınıflandırıcı kombinasyonlarında düşük performanslı modellerin etkisini azaltmayı hedefleyen ağırlıklı birleştirme önerildi.

#### ****Research Motivations Kısmında Yapılanların Detaylı Açıklamaları:****

1. Gürültülü veriler, basit sınıflandırıcıların performansını düşürerek sınıflandırıcı kombinasyonlarının etkinliğini zayıflatır.
2. Sınıflandırıcıdan bağımsız bir filtreleme yöntemi geliştirme ihtiyacı vardır.
3. Önerilen yöntem, farklı veri setlerinde basit sınıflandırıcı performansını artırmayı hedefler.

#### ****Literature Review Kısmında Yapılanların Detaylı Açıklamaları:****

1. Gürültülü örneklerin çıkarılmasıyla sınıflandırma doğruluğunu artırma hedeflenir.
2. Filtreleme algoritmaları, karar sınırlarını netleştirir, eğitim setini küçültür ve işlem maliyetlerini azaltır.
3. K-NN ve diğer yaklaşımlar, veri setlerinde filtreleme ve kümeleme için sıklıkla kullanılmıştır.

#### ****Data Filtering Kısmında Yapılanlar:****

* K-NN yakınlık grafikleri ile aykırı değerler tespit edildi.
* Her noktanın komşularının ağırlıklı ortalaması hesaplandı ve eşik değerine göre aykırı değerler işaretlendi.

#### ****Related Studies Kısmında Yapılanların Detaylı Açıklamaları:****

1. K-NN algoritmasını hızlandırmaya yönelik yenilikçi yöntemler incelendi.
2. Çeşitli sınıflandırıcılar üzerindeki filtreleme etkileri analiz edildi.
3. Dinamik birleştirme seçiminde kullanılan farklı yöntemler araştırıldı.

#### ****Methods ve Datasets Kısmında Yapılanlar:****

1. Farklı özelliklere sahip 6 veri seti kullanıldı.
2. Her veri seti için k-NN grafikleri oluşturularak aykırı değerler belirlendi.
3. 6 farklı sınıflandırıcı ve DES-LA combiner test edildi.

#### ****Data Analysis and Selecting Filtering Parameters Kısmında Yapılanlar:****

1. Aykırı değer tespiti için k-NN grafikleri oluşturuldu.
2. Parametreler, en düşük aykırı oranını sağlayacak şekilde optimize edildi.

#### ****Base Classifiers Development Kısmında Yapılanlar:****

* Logistic Regression, Decision Tree, Random Forest gibi modeller eğitildi ve test edildi.
* DES-LA, yerel doğruluğa dayalı ağırlıklı birleştirme yöntemi ile uygulandı.

#### ****Table 1’de Yapılanlar ve Konuyla Bağlantısı:****

* Farklı veri setleri için filtreleme parametreleri (komşu sayısı, güç, eşik değeri) ve aykırı oranları verilmiştir.
* **Bağlantı:** Parametre seçimlerinin doğruluğa etkisi incelenmiştir.

#### ****Properties of Logistic Regression:****

1. Bernoulli dağılımını takip eder.
2. Maksimum olasılık yöntemiyle tahmin yapılır.

#### ****DES-LA Combiner:****

* K-NN grafikleri ile komşular belirlendi, her sınıflandırıcının hata oranı hesaplandı ve ağırlıklar belirlendi.

#### ****Performance Indicator Measures:****

1. **Accuracy:** Doğru sınıflandırma oranı.
2. **Sensitivity:** Pozitif sınıfların doğru sınıflandırılma oranı.
3. **Specificity:** Negatif sınıfların doğru sınıflandırılma oranı.
4. **AUC:** Pozitif sınıfları negatiften ayırma yeteneği.
5. **Brier Score:** Olasılık tahminlerinin doğruluğunu ölçer.

#### ****Experimental Results and Discussion****

1. Filtreleme işlemi, sınıflandırıcıların performansını artırmıştır.
2. Özellikle basit sınıflandırıcılar filtreleme sonrası karmaşık sınıflandırıcılarla benzer sonuçlar vermiştir.
3. DES-LA combiner, filtreleme sonrası daha tutarlı sonuçlar üretmiştir.

#### ****Table 3, Table 4 ve Table 5****

* **Table 3:** Filtreleme sonrası doğruluk artışlarını gösterir.
* **Table 4:** Performans metriklerinde değişiklikleri yüzdelik olarak özetler.
* **Table 5:** Decision Tree’nin diğer sınıflandırıcılarla doğruluk farkını gösterir.

#### ****Figure 1 - Figure 5****

* **Figure 1:** Filtreleme algoritmasının çalışma mekanizmasını açıklar.
* **Figure 2 ve 3:** Sınıflandırıcıların sıralama dağılımlarını görselleştirir.
* \*\*Figure

4:\*\* DES-LA kombinasyonunun etkilerini analiz eder.

* **Figure 5:** KD-tree algoritmasının k-NN seçim sürecini açıklar.

#### ****Real Case Test: Large Defaults Payments Dataset****

1. Taiwan Default Credit Card veri seti kullanılmıştır.
2. KD-tree algoritması ile hızlı komşu seçimi yapılmıştır.
3. DES-LA combiner, filtreleme sonrası test edilmiştir.

#### ****Table 6 ve Table 7****

* **Table 6:** Filtreleme öncesi sınıflandırıcı performansı.
* **Table 7:** Filtreleme sonrası performans artışları.

#### ****Conclusion****

1. Filtreleme yöntemi basit sınıflandırıcıların performansını artırmıştır.
2. DES-LA combiner ile en iyi sonuçlar elde edilmiştir.
3. Yöntem, gelecekte çok sınıflı sınıflandırma problemlerine uyarlanabilir.

SLAYTLAŞTIRILMIŞ HALİ BUNDAN SONRASI ÖNEMLİ DEĞİL

## ****Başlık: k-NN Yakınlık Grafikleriyle Veri Filtreleme Kullanılarak İkili Sınıflandırmanın İyileştirilmesi****

### ****1. Giriş (Introduction)****

Makale, ikili sınıflandırma problemlerindeki doğruluğu artırmayı hedefleyen yenilikçi bir veri filtreleme yöntemini ele alıyor. Filtreleme işlemi, verideki **gürültü ve aykırı değerlerin** çıkarılmasıyla eğitim setinin kalitesini artırıyor. Önerilen yöntem, **k-NN (k-en yakın komşu) grafikleri** ile veriler arasındaki mesafeyi analiz ediyor ve parametre optimizasyonu yapıyor.

**Temel Problem:**

* Gürültülü ve gereksiz özellikler, basit sınıflandırıcıların performansını olumsuz etkiler.
* Örneğin, Decision Tree ve Naïve Bayes gibi basit sınıflandırıcılar, veri setindeki karmaşıklık nedeniyle daha düşük performans gösterir.
* **Çözüm:** Veriyi filtreleyerek aykırı örnekleri azaltmak ve sınıflandırıcıların doğruluk oranını artırmak.

### ****2. Yöntem (Methods)****

Yöntemde, **6 sınıflandırıcı (Decision Tree, Logistic Regression, Naïve Bayes, SVM, Neural Network, Random Forest)** ve bir heterojen birleştirici (**DES-LA combiner**) kullanılmıştır. Veri setine uygulanmadan önce filtreleme yöntemi şu şekilde uygulanmıştır:

#### ****2.1 Veri Filtreleme Süreci (Data Filtering)****

* **k-NN Grafikleri:** Verideki her noktanın k komşusu ile ilişkisi analiz edilir.
* **Aykırı Değerlerin Tespiti:**
  + Aykırılık, her bir noktanın komşularının ağırlıklı ortalaması ile hesaplanır.
  + Formüller:
    - P(i)=∑j∈N(i)l(j)⋅wijP(i) = \sum\_{j \in N(i)} l(j) \cdot w\_{ij}
    - wij=d(i,j)p∑k∈N(i)d(i,k)pw\_{ij} = \frac{d(i, j)^p}{\sum\_{k \in N(i)} d(i, k)^p}
    - **Amaç:** Aykırı değerleri işaretleyerek temiz bir eğitim seti oluşturmak.

#### ****2.2 Görsellerle Yöntemin Açıklaması****

* **Figure 1:** Filtreleme algoritması, k-NN grafiği üzerinde görselleştirilmiştir. Bu şekilde komşuluk ilişkileri ve aykırılıkların nasıl belirlendiği gösteriliyor.  
  **Anlatım:** Bu grafik, her noktanın komşularıyla bağlantısını ve filtreleme sırasında hangi noktaların ayıklandığını net bir şekilde açıklıyor.

### ****3. Veri Setleri ve Deneysel Süreç (Datasets and Experimental Process)****

Çalışmada, farklı zorluk derecelerine sahip 6 veri seti kullanılmıştır:

* **Örnek Veri Setleri:**
  + German Credit Dataset: 1000 giriş, 22 özellik.
  + Banknote Authentication Dataset: 1372 giriş, 4 özellik.
  + Seismic Bumps Dataset: 2584 giriş, 18 özellik.

Filtreleme, bu veri setlerine uygulanmış ve her sınıflandırıcı filtreleme öncesi ve sonrası test edilmiştir.

**Table 1:** Her veri seti için filtreleme parametreleri (komşu sayısı, güç, eşik değeri) ve aykırı oranları sunulmuştur.  
**Anlatım:** Parametrelerin doğru seçimi, filtrelemenin başarısını artırır.

### ****4. Sonuçların Değerlendirilmesi (Experimental Results and Discussion)****

#### ****4.1 Filtrelemenin Etkisi****

1. Basit sınıflandırıcılar (Naïve Bayes, Decision Tree), filtreleme sonrası doğruluklarını belirgin şekilde artırmıştır.
2. Karmaşık sınıflandırıcılar (SVM, Random Forest), filtrelemeden daha az etkilenmiştir, çünkü zaten gürültüye karşı dayanıklıdırlar.
3. **Table 3:** Filtreleme sonrası her sınıflandırıcıda doğruluk artışları karşılaştırılmıştır.
   * **Anlatım:** Bu tablo, filtrelemenin özellikle düşük doğruluklu veri setlerinde faydalı olduğunu göstermektedir.

#### ****4.2 DES-LA Combiner Performansı****

DES-LA, tüm sınıflandırıcıların yerel doğruluğuna dayalı birleştirme yapar ve filtreleme sonrası en yüksek doğruluğa ulaşmıştır.

#### ****4.3 Görsellerin Yorumlanması****

* **Figure 2:** Decision Tree sınıflandırıcısının filtreleme sonrası sıralama dağılımında, kararların daha keskin hale geldiği gözlemlenir.  
  **Anlatım:** Bu, filtrelemenin sınıflandırıcı karar sınırlarını netleştirdiğini gösterir.
* **Figure 3:** Naïve Bayes için benzer şekilde kararların daha kategorik hale geldiği görülür.  
  **Anlatım:** AUC (Area Under the Curve) değerindeki artış, filtreleme sayesinde daha iyi sınıflandırma yapıldığını gösterir.
* **Figure 4:** DES-LA kombinasyonunun filtreleme sonrası artan duyarlılığı vurgulanır.  
  **Anlatım:** Kombinerin baskın sınıflarda daha kesin kararlar verdiği görülüyor.
* **Figure 5:** KD-tree algoritması, k-NN grafiği oluşturma sürecini hızlandırmak için kullanılmıştır.  
  **Anlatım:** Özellikle büyük veri setlerinde filtrelemenin hızlandırılması açısından önemlidir.

### ****5. Gerçek Dünya Verisi ile Test (Real Case Test)****

#### ****Yapılanlar:****

1. **Veri Seti:** Taiwan Default Credit Card veri seti (30.000 giriş).
2. **Amaç:** Filtrelemenin büyük veri setlerinde etkinliğini test etmek.
3. KD-tree algoritması kullanılarak k-NN grafikleri hızla oluşturulmuştur.
4. Her sınıflandırıcı ve DES-LA combiner bu veri setiyle test edilmiştir.

**Table 6 ve Table 7:**

* **Table 6:** Filtreleme öncesi performans değerlerini gösterir.
* **Table 7:** Filtreleme sonrası performans artışlarını karşılaştırır.  
  **Anlatım:** Filtreleme, özellikle basit sınıflandırıcıların doğruluğunu artırmıştır. DES-LA combiner, filtreleme sonrası en iyi sonuçları üretmiştir.

### ****6. Sonuç (Conclusion)****

1. Çalışma, k-NN grafikleri tabanlı filtrelemenin sınıflandırıcı performansını artırmada etkili olduğunu göstermiştir.
2. DES-LA combiner sayesinde, basit sınıflandırıcıların doğruluğu karmaşık modellerle rekabet edebilecek seviyeye getirilmiştir.
3. Filtreleme yöntemi, aykırı değerlerin etkisini azaltarak model genel performansını iyileştirmiştir.
4. Gelecekte bu yöntemin çok sınıflı sınıflandırma problemlerine uygulanması planlanmaktadır.

### ****Sonuç Olarak:****

Bu çalışma, filtrelemenin sınıflandırıcı performansına etkisini açık bir şekilde göstermektedir. Özellikle görseller, filtrelemenin mantığını ve sonuçlarını anlamayı kolaylaştırır. Örneğin, **Figure 1**, k-NN grafikleri ile aykırı değer tespitini görselleştirirken, **Figure 3 ve 4**, filtrelemenin sınıflandırıcıların karar doğruluğunu nasıl artırdığını açıklamaktadır. Bu tür yenilikçi filtreleme yöntemleri, büyük veri setlerinde sınıflandırıcıların doğruluğunu artırmak için oldukça kullanışlıdır.

Sunu1 MAVİ SUNU OLAN

**Başlıklar Altında Yapılanlar:**

1. **Abstract (Özet):**
   * **Ne yapılmış?**  
     Yüksek boyutlu veri setlerinde sınıflandırma performansını artırmak için özellik seçimi yöntemleri (Random Forest, Boruta, RFE) test edilmiştir. RF, en yüksek doğruluk oranına ulaşmıştır.
2. **Introduction (Giriş):**
   * **Ne yapılmış?**  
     Yüksek boyutlu verilerin model performansını düşürdüğü vurgulanmış. Özellik seçiminin önemi ve bu çalışmada kullanılan yöntemlerin (RF, Boruta, RFE) amacı açıklanmıştır.
3. **Material and Method:**
   * **Ne yapılmış?**  
     Bank Marketing, Car Evaluation, Human Activity Recognition veri setleri tanıtılmış. Özellik seçimi yöntemleri (RF, RFE, Boruta) ve sınıflandırma algoritmaları (RF, SVM, KNN, LDA) detaylandırılmış.
4. **Model Performance Evaluation:**
   * **Ne yapılmış?**  
     Hata Matrisi (Confusion Matrix), ROC eğrisi ve doğruluk gibi metrikler ile modellerin performansı değerlendirilmiş. Özellik seçiminin doğruluğa etkisi analiz edilmiştir.
5. **Results and Discussion:**
   * **Ne yapılmış?**  
     Özellik seçimiyle doğruluk artışı ve işlem süresi azalışı gösterilmiş. RF’nin en iyi performansı sunduğu vurgulanmıştır.
6. **Conclusions and Future Work:**
   * **Ne yapılmış?**  
     Özellik seçiminin performansa katkısı özetlenmiş, gelecekte farklı yöntemlerin (Gradient Boosting, CNN) ve veri setlerinin kullanılabileceği önerilmiştir.

**Grafik ve Tabloların Konuyla Bağlantıları:**

1. **Figure 2:**
   * **Ne gösteriyor?**  
     Bank Marketing veri setinde RF ile her bir özelliğin önem dereceleri.
   * **Bağlantı:**  
     "Duration" ve "Balance" gibi özelliklerin sınıflandırma doğruluğuna katkısı vurgulanmıştır.
2. **Figure 3:**
   * **Ne gösteriyor?**  
     RFE ile seçilen kritik özellikler.
   * **Bağlantı:**  
     RFE, veri boyutunu azaltırken doğruluğu artırmıştır.
3. **Figure 4:**
   * **Ne gösteriyor?**  
     Boruta algoritmasıyla önemli ve önemsiz özellikler karşılaştırılmış.
   * **Bağlantı:**  
     Boruta, özellik seçimini daha katı şekilde yaparak "Duration" gibi önemli özelliklerin kritik olduğunu göstermiştir.
4. **Table 3:**
   * **Ne içeriyor?**  
     Kullanılan veri setlerinin tanımları (örneğin, Bank Marketing'de 17 özellik, 45,211 gözlem).
   * **Bağlantı:**  
     Özellik seçimi yöntemlerinin geniş bir veri yelpazesinde test edildiğini gösterir.
5. **Table 8-10:**
   * **Ne içeriyor?**  
     Sınıflandırıcıların doğruluk ve performans karşılaştırmaları.
   * **Bağlantı:**  
     RF'nin diğer yöntemlere göre daha iyi sonuç verdiği kanıtlanmıştır.
6. **ROC ve AUC Grafikleri:**
   * **Ne gösteriyor?**  
     Farklı modellerin ayırt etme gücü.
   * **Bağlantı:**  
     Özellik seçiminin ROC eğrisini iyileştirerek modeli nasıl optimize ettiğini açıklar.