**Kredi Risk Analizi**

**Özet**

Risklerin gerçekçi bir şekilde tespit edilmesi ve olası kayıpların görünür hale getirilmesi amacıyla kredi riskinin ölçülmesine ihtiyaç duyulmaktadır. Kredi riskinin ölçülmesine yönelik olarak gelişti rilen çeşitli modellerde temel amaç, temerrüde düşmesi olas borçlunun belirlenerek risk miktarının tahmin edilmesidir.

Bu çalışmada, dört farklı makine öğrenmesi algoritması kullanarak modeller oluşturacağız: **Logistic Regression**, **Decision Tree**, **Gradient Boosting ve Malatya Centrality Algoritması**.

Öncelikle, modelleri eğitmek için Kaggle sitesinden bir veri seti seçiyoruz. Seçtiğimiz veri seti, bir Alman bankasına ait gerçek verilerden oluşmaktadır. Daha sonra, sonuca en çok etki eden ve en çok ilişkili özellikleri seçiyoruz. Veriler işlenirken, bilgi kaybı fazla olan veriler temizlenmiştir ve bu süreçte Standardizasyon kullanılmıştır. Veriler işlenip eğitim için hazırlandıktan sonra, üç gruba ayrılır: eğitim, doğrulama ve test.

Veri hazırlama süreci tamamlandıktan ve modellere ana girdi olacak özellikler seçildikten sonra eğitim süreci başlar.

Eğitim işlemi tamamlandıktan sonra, modellerin performansını test grubuyla değerlendiriyoruz. Bu aşamada, doğruluk, hasasiyet gibi çeşitli performans ölçütlerini kullanarak üç modelin sonuçlarını karşılaştırıyoruz.

Sonuç olarak, genel doğruluk ve yüksek geri çağırma açısından **Gradient Boosting** algoritmasının üstün performans gösterdiğini gözlemliyoruz. **Decision Tree** iyi bir performans sergilemiştir ancak Gradient Boosting'e kıyasla daha düşük bir verim sunmuştur; makul bir doğruluğa sahip olsa da geri çağırma değeri biraz daha düşüktür. **Malatya Merkezlik** ise en düşük doğruluk ve geri çağırma oranları ile üç algoritma arasında en düşük performansı göstermiştir.

**Giriş**

Projemle ilgili birçok kaynağı ve bunlarda açıklanan çözümleri inceledikten sonra, kredi riski analizi sorununu çözmek için en popüler üç algoritmayı seçtim. Daha sonra, bu proje için uygun veri setine ve seçilen algoritmaların yürütülmesine seçtim.

veri setimiz çok gereksiz özellik içerdiği nedeniyle ,hedefle maksimum korelasyona sahip özellikleri seçmek gerekti. Bunu yapmak için RandomForestClassifier'ı kullandım. Bu, en iyilerini seçmek için her özelliği hedef üzerindeki etkileriyle birlikte kontrol eden bir makine öğrenme algoritmasıdır. Algoritmalar aracılığıyla ana girdiler olacak özellikleri seçtikten sonraki adım, modeli eğitmek için verileri ön işlemeye başladım.

Veri setinin modelde zayıflığa sebep olacak eksik değerler içermediğini kontrol ettim,bundan sonra veri modellerinin eğitim seviyesini yükseltmek amacıyla, verilerin ölçü birimlerini standartlaştırmak için Normalizasyon yöntemini kullandım. Bu adım, özellikler arasındaki ölçüm birimlerindeki farkı azaltarak modellerin öğrenme oranını ve performansını iyileştirmektir.

Daha sonra, veri setini üç parçaya böldüm (%85 eğitim ve %15 test) ve eğitim seti %70 eğitim ve %15 doğrulama olarak bölündü.

böylece ön işleme aşamasını bitirdim, ve veri setimiz modellerin eğitimi için hazırladım.

Projenin dayanacağı modelleri tanımladım: Gradient Boosting, Decision Tree ve Logistic Regression.

1. **Gradient Boosting**: Gradyan artırma, zayıf modelleri (örneğin küçük karar ağaçları) güçlendiren ve güçlü bir model oluşturan bir makine öğrenimi tekniğidir. Bu algoritma, bir önceki modelin hatalarını (Residual Errors) azaltarak iteratif olarak çalışır. Kayıp fonksiyonu (Loss Function) kullanılarak modeller ardışık olarak eğitilir ve performans adım adım geliştirilir.
2. **Decision Tree**: Karar ağacı, verileri anlamak ve karar almak için kullanılan bir algoritmadır. Bu algoritma, verileri belirli kurallara göre alt gruplara ayırarak bir karar ağacı oluşturur. Model, kök düğümden (Root Node) başlar ve dallar (Branches) boyunca ilerleyerek yaprak düğümlerine (Leaf Nodes) ulaşır; bu düğümler kararları içerir.
3. **Logistic Regression**: Lojistik regresyon, sınıflandırma için kullanılan bir makine öğrenimi algoritmasıdır, özellikle ikili sınıflandırma (Binary Classification) problemlerinde yaygın olarak kullanılır. Örneğin, "Evet/Hayır" veya "Kabul/Red" gibi durumlar için idealdir. Bu algoritma, çıktıları 0 ile 1 arasında olasılık değerlerine dönüştürmek için lojistik fonksiyon (Sigmoid Fonksiyonu) kullanır. Ana hedef, bağımsız değişkenler (Features) ile bağımlı değişken (Target) arasında doğrusal bir ilişki bulmaktır.
4. **Malatya Centralty:** Malatya Merkezilik: Vertex cover problemini çözmek için polinomsal bir yaklaşım sunan bir algoritmadır. Düğümlerin merkezilik değerlerini, düğüm derecesinin bitişik düğümlerin derecelerine oranlarının toplamı ile hesaplar ve veri gruplarını sınıflandırmada kullanılır.

Eğitim seti, eğitimi başlatmak için kullanılır. Daha sonra, modelin kararlılığını değerlendirmek için ortalama doğruluk ve standart sapma hesaplanır; ayrıca doğruluk ve hassasiyet gibi metrikler de hesaplanır. Uygulanan modellerin sonuçları karşılaştırılarak en iyi performansı gösteren model belirlenir.

**Yöntem**

**Veri seti**

* **Kaynak:** Veri setini Kaggle'dan aldım. Bu, bir Alman bankasının gerçek verilerinden oluşan bir veri setidir.
* **Boyut:** Veri seti **1 milyon** gözlemdenoluşmaktadır.

modellerde girdi olarak kullanılacak özellikler belirlenmesi, model tahmin gücünü doğrudan etkileyen faktördür. Aynı zamanda girdi olarak kullanılacak gözlemler içinde bu durum geçerlidir.

RandomForestClassifier algoritması kullanılarak hedef değişkenle en yakından ilişkili en önemli on özellik belirlendi ve seçildi. (RandomForestClassifier, rastgele seçilmiş veri örnekleri ve özellik alt kümeleriyle birden fazla karar ağacı eğitip bunların tahminlerini birleştirerek sınıflandırma yapan ve yüksek doğruluk sağlayan bir makine öğrenimi algoritmasıdır.)

Seçilen özellikler bunlardır**:**

* **Kredi Puanı**
* **Yıllık Gelir**
* **Yıllık Giderler**
* **Kredi Tutarı**
* **Borç-Gelir Oranı**
* **İflas Geçmişi**
* **Önceki Kredi Temerrütleri**
* **Ödeme Geçmişi**
* **Net Varlık Değeri**
* **Faiz Oranı.**
* **Hedef Değişken:** **Kredi Onayı** (kabul/red)

A graph with blue squares

Description automatically generated

RandomForestClassifier algoritmasi veri setimiz üzerinde sonuçları

Veri seti hakkında bilgiler

|  |  |  |  |
| --- | --- | --- | --- |
| **özellik** | **Her sütundaki en sık tekrar eden değer (mode):** | **Her sütundaki ortalama değer (mean):** | **Her sütundaki benzersiz değer sayısı:** |
| Kredi Puanı | 508 | 574.559628 | 550 |
| Yıllık Gelir | 53002 | 59971.2273 | 82425 |
| Yıllık Giderler | 30128 | 29990.07072 | 57746 |
| Kredi Tutarı | 19143 | 20001.6435 | 31209 |
| Borç-Gelir Oranı | 0.000061 | 0.500026 | 999509 |
| İflas Geçmişi | 0 | 0.049872 | 2 |
| Önceki Kredi Temerrütleri | 0 | 0.100256 | 2 |
| Ödeme Geçmişi | 11 | 14.503627 | 30 |
| Net Değer | 45409 | 50035.89736 | 147941 |
| Faiz Oranı | 0.010084 | 0.155151 | 998240 |
| Kredi Onayı | 0 | 0.373169 | 2 |

**Veri Ön İşleme**

* **Eksik Değerlerin Kontrolü:** Veri setinde eksik değerlerin olup olmadığını kontrol ettim.
* **Veri Standardizasyonu:** Özellikleri standartlaştırmak için StandardScaler kullandım. Bu, modellerin daha iyi performans göstermesine yardımcı oldu.( Standardizasyon (Standartlaştırma), verilerin ölçeklerini birbiriyle uyumlu hale getirmek için kullanılan bir yöntemdir. Bu yöntemle, veriler ortalaması 0 ve standart sapması 1 olacak şekilde dönüştürülür. Formül şu şekildedir:

x: Veri değeri

μ: Verilerin ortalaması

σ: Verilerin standart sapması

Bu işlem, özellikle farklı ölçeklere sahip değişkenlerin bir arada kullanıldığı makine öğrenimi algoritmalarında modelin doğruluğunu artırmak için önemlidir.)

* **Veri Seti Bölme:** Veri seti, modelin eğitimi için %70 eğitim, modelin ayarlanması için %15 doğrulama ve modelin performansını değerlendirmek için %15 test olarak bölündü.

**Çalışmada Kullanılan Sınıflandırma Algoritmaları**

**1. Lojistik Regresyon:**

* **Genel Açıklama:**
  + Lojistik regresyon, **denetimli öğrenme** algoritmalarından biridir ve genellikle **ikili sınıflandırma** (örneğin, evet/hayır, onaylandı/onaylanmadı) problemlerini çözmek için kullanılır.
  + Bağımsız değişkenler (özellikler) ile bağımlı değişken (hedef) arasındaki ilişkiyi lojistik fonksiyon kullanarak modellemektedir.
  + Algoritma, **logistik fonksiyonunu** (**sigmoid fonksiyonu**) kullanarak sürekli değerleri 0 ile 1 arasında olasılıklara dönüştürür.
  + **Lojistik Fonksiyon (Sigmoid Fonksiyonu):** 
    - ho tahmin edilen olasılık
    - b katsayılar (model parametreleri)
    - x bağımsız değişkenler
* **Neden Kullandım:** Basitlik ve Yorumlanabilirlik, ve Hız ve Verimlilik

**2. Karar Ağacı (Decision Tree):**

* **Genel Açıklama:**
  + Karar ağacı, özelliklere dayalı olarak kararlar veren ve bu kararları ağaç yapısı şeklinde gösteren bir makine öğrenma algoritmasıdır.
  + Her düğümde en iyi bölünme noktasını belirlemek için belirli bir özelliği kullanır ve bu sayede veriyi sınıflandırır.
  + Verileri, özelliklerin belirli eşik değerlerine göre dallandırarak son karara (yapraklara) ulaşır.
  + Entropy förmülü :
    - Entropy =
    - p(x) bir sınıfın olasılığı.
* **Neden Kullandım:** Anlaşılabilirlik ve Görselleştirme, Çeşitli Veri Tipleriyle Uyumluluk.

**3. Gradient Boosting:**

* **Genel Açıklama:**
  + Gradient Boosting, **makine öğrenimi** algoritmalarından biridir ve genellikle **sınıflandırma** ve **regresyon** problemlerini çözmek için kullanılır. Bu yöntem, bir dizi **zayıf öğreniciyi** ardışık olarak birleştirerek güçlü bir model oluşturur. Gradient Boosting(Gradyan arrtırma)'deki zayıf öğreniciler genellikle **karar ağaçlarıdır yada linner regresiyon.**
  + Her bir yeni model, önceki modelin hatalarını düzeltmeye çalışır ve böylece modelin performansı kademeli olarak iyileşir.
  + **Nasıl Çalışır?**

Gradient Boosting'in temel çalışma prensibi şu şekildedir:

1. **Başlangıç Modeli Oluşturma:**

İlk olarak, basit bir model (örneğin, hedef değişkenin ortalaması) ile başlar.

1. **Hataların Hesaplanması:**

Mevcut modelin tahmin hataları (rezidüeller) hesaplanır.

1. **Yeni Model Eğitimi:**

Yeni bir model, bu hataları tahmin etmek için eğitilir.

1. **Modelin Güncellenmesi:**

Yeni model, önceki modele eklenir. Bu işlem genellikle bir öğrenme oranı (ν) ile ağırlıklandırılır.

1. **Tekrar:** Adımlar belirli bir iterasyon sayısı boyunca veya hata istenilen seviyeye düşene kadar tekrarlanır.

Her iterasyonda, model hataları azaltmak için gradyan inişi kullanır. Bu sayede, model sürekli olarak iyileşir ve daha doğru tahminler yapar.

* **Neden Kullandım:** Yüksek Tahmin Gücü, Esneklik ve Genel Başarı.

**4. Malaya Merkezlik algoritması.**

**Nasıl uygulandı**

1. **verileri yüklenir ve ön işlemler yapılır.**

* Veri seti yüklenir ve ayrılır(Değişkenler ve hedef).

1. **Graf laroluşturulur**
   * Graflar oluşurmak için her iki sınıftan **"KNN"** algoritması kullanarak en iyi 10 düğüm seçilir.
   * Her veri grubu için düğümler arasındaki benzerliklere dayalı grafikler oluşturulur.
   * Düğümler arasındaki bağlayan kenarlar ise benzerlikleri temsil eder.
2. **Düğümlerin Ağırlıklarının Hesaplanması**:

* Her düğüm için merkezilik (centrality) değeri, komşu düğümlerle olan ilişkisine göre hesaplanır.

1. **Yeni Bir Örneğin Sınıflandırılması**:

* Yeni bir düğüm eklendiğinde, bu örnek iki grafiğe de dahil edilir.
* Her grafikteki merkezilik değeri hesaplanır.

1. **Grubun Belirlenmesi**:

* Hangi grafta merkzlik diğeri daha büyükse ona ayıttır.

Malatya Algoritması düğüm (v) değerlerinin hesaplanmasında kullanılan algoritma aşağıda verilmiştir.  Bu denklemde, n sırasıyla grafikteki tüm düğümleri temsil eder. Bir düğüm vi için, bitişik düğümlerin kümesi N(vi) olarak temsil edilir.

A graph diagram of a graph

Description automatically generated with medium confidenceψ(𝑣𝑖)=∑ 𝑑(𝑣𝑖) 𝑑(𝑣𝑗)

Malatya Merkezlik algoritmasi çalışma makanizması

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| düğüm | Kredi Puanı | Yıllık Gelir | Yıllık Harcamalar | Kredi Tutarı | Borç/Gelir Oranı | İflas Geçmişi | Önceki Kredi Temerrütleri | Ödeme Geçmişi | Net Varlık | Faiz Oranı |
| 1 | 664 | 52998 | 30735 | 19339 | 0.852186 | 0 | 0 | 28 | 62430 | 0.234736 |
| 2 | 628 | 52730 | 31857 | 19814 | 0.985587 | 0 | 0 | 22 | 62743 | 0.038005 |
| 3 | 590 | 68108 | 39469 | 19556 | 0.327175 | 0 | 0 | 24 | 61754 | 0.199755 |
| 4 | 822 | 69290 | 38848 | 19396 | 0.245932 | 0 | 1 | 28 | 61850 | 0.131495 |
| 5 | 643 | 59501 | 23220 | 14649 | 0.332150 | 0 | 0 | 19 | 33189 | 0.207431 |
| 6 | 738 | 59590 | 22997 | 14574 | 0.855585 | 0 | 0 | 0 | 34607 | 0.107384 |
| 7 | 771 | 53737 | 28001 | 19185 | 0.086573 | 0 | 0 | 29 | 68362 | 0.245849 |
| 8 | 647 | 54717 | 27736 | 19451 | 0.132327 | 0 | 0 | 7 | 66922 | 0.264410 |
| 9 | 753 | 44888 | 32275 | 15505 | 0.022508 | 0 | 0 | 0 | 23716 | 0.239546 |
| 10 | 521 | 45733 | 32004 | 17052 | 0.386387 | 0 | 0 | 16 | 23453 | 0.160532 |

En iyi kabul düğümlerin özellikleri

En iyi reddedilmiş düğümlerin özellikleri

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| düğüm | Kredi Puanı | Yıllık Gelir | Yıllık Harcamalar | Kredi Tutarı | Borç/Gelir Oranı | İflas Geçmişi | Önceki Kredi Temerrütleri | Ödeme Geçmişi | Net Varlık | Faiz Oranı |
| 1 | 425 | 37838 | 18525 | 18509 | 0.721552 | 0 | 1 | 16 | 38842 | 0.231667 |
| 2 | 710 | 38117 | 18506 | 18729 | 0.636813 | 0 | 0 | 24 | 39347 | 0.044492 |
| 3 | 561 | 61488 | 31787 | 16748 | 0.343825 | 0 | 0 | 17 | 53600 | 0.043157 |
| 4 | 723 | 61269 | 31221 | 17240 | 0.404977 | 0 | 0 | 25 | 54165 | 0.257034 |
| 5 | 369 | 65361 | 40264 | 17221 | 0.601096 | 0 | 0 | 11 | 75812 | 0.199792 |
| 6 | 366 | 65831 | 40746 | 18131 | 0.148004 | 0 | 0 | 28 | 75536 | 0.091624 |
| 7 | 343 | 58325 | 36817 | 20147 | 0.069956 | 0 | 0 | 20 | 61247 | 0.220284 |
| 8 | 399 | 58734 | 37693 | 20736 | 0.476936 | 0 | 0 | 7 | 61627 | 0.186487 |
| 9 | 398 | 67364 | 28075 | 21019 | 0.656422 | 0 | 0 | 0 | 30597 | 0.131936 |
| 10 | 690 | 67578 | 28971 | 21239 | 0.244458 | 0 | 0 | 18 | 29429 | 0.195977 |

En iyi düğümlerin malatya merkezlilik değeri

|  |  |  |
| --- | --- | --- |
| Düğüm numarası | Kabul düğümlerin Centrality Diğeri | Redd düğümlerin Centrality Diğeri |
| 1. | 1.90 | 10.90 |
| 2. | 2.90 | 0.526316 |
| 3 | 3.69473 | 2.90 |
| 4. | 2.60818 | 3.90 |
| 5. | 6.49473 | 4.90 |
| 6. | 7.91111 | 5.90 |
| 7. | 8.44210 | 6.90 |
| 8. | 9.44210 | 7.90 |
| 9. | 11.0222 | 8.90 |
| 10. | 11.3894 | 9.90 |

Modelin işleyişini iki örnekle açıklayalım

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Yeni  düğüm | Kredi Puanı | Yıllık Gelir | Yıllık Harcamalar | Kredi Tutarı | Borç/Gelir Oranı | İflas Geçmişi | Önceki Kredi Temerrütleri | Ödeme Geçmişi | Net Varlık | Faiz Oranı |
| A | 359 | 63229 | 32186 | 17614 | 0.798218 | 1 | 0 | 22 | 38017 | 0.08759 |
| B | 529 | 54015 | 22321 | 17714 | 0.324165 | 0 | 0 | 26 | 47364 | 0.08524 |

iki düğümün kabul ve redd graflarda malatya merkezlilik değerleri

|  |  |  |  |
| --- | --- | --- | --- |
| düğüm | kabul Centrality Diğeri | redd Centrality Diğeri | sonuc |
| A | 10.90 | 10.5772 | **kabul** |
| B | 10.57 | 10.9 | **redd** |

**A** düğüm kabul grafta merkezlilik değeri daha düyük oldugu için sonucu **kabul edilir.  
B** düğüm redd grafta merkezlilik değeri daha düyük oldugu için sonucu **reddedilir.**

**Model Ayarlaması ve Değerlendirme**

* **Değerlendirme Metrikleri:** Modelleri değerlendirmek için aşağıdaki metrikleri kullandık:
  + A table with a list of different languages

    Description automatically generated with medium confidence**Karışıklık matrisi:** bir modelin performansını değerlendirmek için kullanılan ve görsel olarak bir tablo şeklinde temsil edilen bir araçtır.
  + **Doğruluk (Accuracy):** Tüm doğru tahminlerin toplam tahmin sayısına oranıdır.

* + **Duyarlılık (Recall):** Gerçek pozitiflerin doğru şekilde tahmin edilen oranıdır.

* + **Kesinlik (Precision):** Pozitif tahminlerin ne kadarının doğru olduğunu **gösterir.**

* + **F1 Skoru (F1 Score):** Kesinlik ve duyarlılığın harmonik ortalamasıdır.

**En İyi Performans Gösteren Model:** **Gradient Boosting** modeli, diğer modellere göre en iyi performansı gösterdi.

**Sonuçlar**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model Performans Karşılaştırması:** | |  |  |  |  | |
| **Model** | **Doğruluk** | **Duyarlılık / Tespit Oranı** | **Kesinlik** | **F1 Score** | | **Karmasiklik matris** |
| Gradient Boosting | 75.13 | 59.98 | 74.98 | 75.13 | [[37.85, 24.875], [0.0036, 37.3]] | |
| Decision Tree | 70.05 | 60.11 | 59.55 | 70.05 | [[48.075, 14.64], [15.24, 22.06]] | |
| Logistic Regression | 68.69 | 59.05 | 55.92 | 68.69 | [[97951, 27467], [34978, 39604]] | |
| Maltya merkezlik | 62.00 | Recall: 28.95 | 50.00 | 36.67 | [[51 , 11]  [27, 11]] | |

* **Sonuçlar:**

**1. Gradient Boosting**

**1. Gradient Boosting (Doğruluk: %75.13)**

* **Doğruluk (%75.13):** Gradient Boosting, genel doğruluk açısından en iyi modeldir.
* **Duyarlılık (%59.98):** Model, pozitif durumları tespit etmede makul bir performans göstermektedir.
* **Kesinlik (%74.98):** Yanlış alarmları azaltma konusunda oldukça başarılıdır.
* **F1-Skoru (%75.13):** Doğruluk ve hatırlama arasında iyi bir denge sağlamaktadır.
* **Karmaşıklık Matrisi:** Pozitif sınıfların doğru tespit oranı yüksek (74,575), yanlış pozitif ve negatiflerin sayısı ise genel performans göz önüne alındığında kabul edilebilir düzeydedir (49,750).
* Sonuç: Genel olarak, Gradient Boosting modeli, diğer modellere göre daha dengeli ve güçlü bir performans göstermiştir.

**2. Karar Ağacı (Decision Tree) (Doğruluk: %70.05)**

* **Doğruluk (%70.05):** Gradient Boosting'e kıyasla daha düşük olsa da makul bir doğruluk oranına sahiptir.
* **Recall (%60.11):** Gradient Boosting'den biraz daha iyi bir hatırlama oranı sunar, pozitif sınıfları tespit etme konusunda daha başarılıdır.
* **Precision (%59.55):** Gradient Boosting'e kıyasla daha düşük olup, daha fazla yanlış pozitif sınıf tespit etmektedir.
* **F1-Skoru (%70.05):** Göreceli olarak dengeli fakat Gradient Boosting'in gerisindedir.
* **Karmaşıklık Matrisi:** Yanlış pozitif sınıfların daha yüksek olması (30,575) doğruluk oranındaki düşüşü açıklamaktadır.

**Sonuç:** Bu model, doğruluk açısından iyi performans gösterse de, pozitif tahminlerde Gradient Boosting kadar güçlü değildir.

**3. Lojistik Regresyon (Doğruluk: %68.69)**

* **Doğruluk (%68.69):** Karar Ağacı ve Gradient Boosting'e kıyasla daha düşük doğruluğa sahiptir.
* **Recall (%59.05):** Pozitif sınıfları tespit etme oranı düşüktür ve daha fazla pozitif sınıfı kaçırma eğilimindedir.
* **Precision (%55.92):** Diğer modellere göre daha düşük, yanlış pozitif oranı daha yüksektir.
* **F1-Skoru (%68.69):** Gradient Boosting ve Karar Ağacı'nın gerisindedir.
* **Karmaşıklık Matrisi:** Yanlış pozitif sınıfların artması (34,978) ve doğru pozitif sınıfların azalması modelin performans düşüklüğünü açıklamaktadır.

**Sonuç:** Logistic Regression modeli, diğer modellere göre daha zayıf bir performans göstermiştir.

**4. Malatya Merkezlilik (Doğruluk: %64.5)**

* **Doğruluk (%64.5):** En düşük doğruluk oranına sahiptir.
* **Recall (%12):** Pozitif durumları tespit etme konusunda oldukça zayıf bir performans göstermektedir.
* **Precision (%28):** Yanlış alarmlar oldukça fazladır ve doğruluk oranı düşüktür.
* **F1-Skoru (%16.60):** Doğruluk ve hatırlama arasında dengesiz bir performans göstermiştir.
* **Karmaşıklık Matrisi:** Yanlış negatiflerin (57,733) ve yanlış pozitiflerin (15,624) yüksek sayısı modelin düşük performansını açıklamaktadır**.**

**Sonuç:** Bu modelin performansı, diğer modellerle kıyaslandığında düşüktür.

**Genel Karşılaştırma:**

Gradient Boosting: Genel performans (doğruluk, hatırlama, doğruluk, F1-Skoru) açısından en iyi modeldir ve en güvenilir sonuçları sunar.

Karar Ağacı (Decision Tree): Kabul edilebilir bir performans sunar ancak Gradient Boosting'in gerisindedir.

Lojistik Regresyon: Daha basit bir seçenek olarak makul bir performansa sahiptir, ancak Karar Ağacı ve Gradient Boosting'e kıyasla daha zayıf kalmaktadır.

Malatya Merkezlilik: Diğer modellere kıyasla oldukça zayıf bir performansa sahiptir ve geliştirilmesi gerekmektedir.

Sonuç olarak, Gradient Boosting, kredi risk analizi gibi projelerde tercih edilmesi gereken en iyi modeldir.

**Deney Sonuçları**

Yapılan çalışma kapsamında, ([**Financial Risk Data - Large**](https://www.kaggle.com/datasets/deboleenamukherjee/financial-risk-data-large)) kaggle’den ait kredi veri seti bilgilerini içeren veriler kullanılarak, veri kalitesi ve model ön işleme çalışmaları gerçekleştirilmiştir. Yeni başvuru sahiplerinin veya aktif kredilerin temerrüt risk oranını değerlendirmek amacıyla, ilgili istatistiksel ve makine öğrenmesi algoritmaları uygulanmış olup, algoritmaların iyi ve kötü müşterileri ayrıştırıcı gücünün belirlenmesi performans ölçüleriyle sağlanmıştır.

Öznitelik seçimi kapsamında istatistiksel hatalardan arınma ve sınıflandırma gücü yüksek olan değişkenlerin belirlenmesi için çeşitli değişken indirgeme tekniklerinden faydalanılmıştır.

İlgili algoritmalarda girdi değişkeni olarak kullanılacak nihai özniteliklerin belirlenmesi için RandomForestClassifier değişken indirgeme tekniği olarak kullanılmış,.

Ham veri kümesinin %70’ı eğitim, %15’si doğrulama ve %15’si test veri seti olacak şekilde ayrılarak, makine öğrenmesi teknikleri bu veri setleri üzerinden gerçekleştirilmiştir.

karmaşıklık matrisinden elde edilen doğruluk, hassasiyet, özgünlük, kesinlik ve F1 skor ölçüleri incelenerek, dört algoritmanın makine öğrenmesi yetenekleriyle gerçekleştirdiği performansları değerlendirilmiştir.

müşterilerinin temerrüt riskini değerlendirmek için en iyi sınıflandırma başarısını her bir veri setinde sağlayan modelin, Gradyan Artırma algoritması olduğu gözlemlenmiştir.

**Kaynaklar**

1. Bahçe Çizer, E. (2017, Aralık). *Makine Öğrenmesi Teknikleriyle Kredi Risk Analizi*.
2. Gursakal, S. (2023, Mart). *Kredi Temerrüt Riskini Tahmin Etmede Makine Öğrenme Algoritmalarının Karşılaştırılması*.
3. Karcı, A., Yakut, S., & Öztemiz, F. (2022, Kasım 27). *Minimum Tepe Örtüsü Probleminin Çözümünde Merkeziyet Değerine Dayalı Yeni Bir Yaklaşım: Malatya Merkeziyet Algoritması*.
4. Keskenler, M. F., Dal, D., & Aydın, T. (2021, Mayıs 2). *Yapay Zeka Destekli ÇOKS Yöntemi ile Kredi Kartı Sahtekarlığının Tespiti*.
5. Kurt, R. (2020, Eylül 18). *BASEL Düzenlemelerine Uygun Olarak Kredi Risk Analizi*.
6. Chow, J. C. K. (2017, April). ***Analysis of Financial Credit Risk Using Machine Learning.***
7. Sakal, M. (2020, Haziran 20). *Makine Öğrenmesi Algoritmaları Kısa Açıklamaları****.***
8. Northern Trust. (2023). *Machine Learning Algorithms for Credit Risk Assessment: An Economic and Financial Analysis*.
9. *Gradient boosting machines, a tutorial( Dec, 2013) (Alexey Natekin)*

**Ekler**

Ek1 – kullanılan ana kütüphaneler :

* from sklearn.linear\_model import LogisticRegression
* from sklearn.ensemble import GradientBoostingClassifier
* from sklearn. tree import DecisionTreeClassifier
* from sklearn.metrics import accuracy\_score, precision\_score, recall\_score, f1\_score, confusion\_matrix

Ek2 – Modellerin tanımı:

models = {  
 "Gradient Boosting": GradientBoostingClassifier(),  
 "Decision Tree": DecisionTreeClassifier(),  
 "Logistic Regression": LogisticRegression()  
}

Ek3– Model eğtim:

model.fit(X\_train, y\_train)

Ek4 – Performans değerlendirme:

accuracy = accuracy\_score(self.y\_test, y\_pred)  
precision = precision\_score(self.y\_test, y\_pred)  
recall = recall\_score(self.y\_test, y\_pred)  
f1 = f1\_score(self.y\_test, y\_pred)  
conf\_matrix = confusion\_matrix(self.y\_test, y\_pred)  
conf\_matrix\_percentage = (conf\_matrix / conf\_matrix.sum())

Ek5 – Knn algoritması

from sklearn.neighbors import NearestNeighbors

import numpy as np

def select\_top\_knn(data\_subset, features, n\_neighbors=10):

subset\_features = data\_subset[features].values

nbrs = NearestNeighbors(n\_neighbors=n\_neighbors + 1, metric='euclidean') # +1 to account for self-neighbor

nbrs.fit(subset\_features)

distances, indices = nbrs.kneighbors(subset\_features)

all\_indices = indices[:, 1:].flatten() # Exclude self-neighbor (first column)

all\_distances = distances[:, 1:].flatten()

sorted\_indices = np.argsort(all\_distances)[:n\_neighbors]

top\_indices = all\_indices[sorted\_indices]

return data\_subset.iloc[top\_indices]

Ek6 – : Malatya Merkezlik Kabul ve red için benzerlik grafiği oluşturma

import networkx as nx  
import matplotlib.pyplot as plt  
  
  
def plot\_graph(data):  
 c1, g1 = malatya\_centrality(data, top\_accepted)  
 c2, g2 = malatya\_centrality(data, top\_rejected)  
  
 if c1 > c2:  
 belong\_graph = {'sınıf': 'accept', 'graph': g1}  
 non\_belong\_graph = {'sınıf': 'reject', 'graph': g2}  
 else:  
 belong\_graph = {'sınıf': 'reject', 'graph': g2}  
 non\_belong\_graph = {'sınıf': 'accept', 'graph': g1}  
  
 graphs = [g1, g2]  
  
 for graph in graphs:  
 G = nx.Graph()  
  
 node\_list = [node for node in graph]  
  
 G.add\_nodes\_from([node['index'] for node in node\_list])  
  
 edges = []  
  
 for node in node\_list:  
 for link in node['links']:  
 if link['similarities'] != 0:  
 edges.append((node['index'], link['index'], link['similarities']))  
  
 G.add\_edges\_from([(u, v, {"label": label}) for u, v, label in edges])  
  
 node\_colors = ['yellow' if node == graph[0]['index'] else 'skyblue' for node in G.nodes]  
  
 pos = nx.spring\_layout(G)  
 nx.draw(G, pos, with\_labels=True, node\_color=node\_colors, node\_size=800)  
  
 edge\_labels = nx.get\_edge\_attributes(G, "label")  
 nx.draw\_networkx\_edge\_labels(G, pos, edge\_labels=edge\_labels, font\_color="red")  
  
 plt.show()  
 return belong\_graph, non\_belong\_graph

Ek7 – : Malatya Merkeziliği Hesaplama

def malatya\_centrality(data, dataset):  
 graph = [{'index': 0, 'weight': 0, 'links': []} for \_ in range(11)]  
  
 current\_data = data  
  
 for i in range(len(dataset)):  
 compare\_by\_data = dataset.iloc[i]  
 similarities = (current\_data == compare\_by\_data).sum()  
 graph[0]['links'].append({'index': dataset.index[i], 'similarities': similarities})  
 graph[0]['weight'] += similarities  
  
 for i in range(len(dataset)):  
 current\_data = dataset.iloc[i]  
 graph[i + 1]['index'] = dataset.index[i]  
  
 for j in range(i + 1, len(dataset)):  
 compare\_by\_data = dataset.iloc[j]  
 similarities = (current\_data == compare\_by\_data).sum()  
 graph[i + 1]['links'].append({'index': dataset.index[j], 'similarities': similarities})  
 graph[i + 1]['weight'] += similarities  
  
 for i in range(len(graph) - 1, -1, -1):  
 top = 0  
 for j in range(i - 1, -1, -1):  
 top += [link for link in graph[j]['links'] if link['index'] == graph[i]['index']][0]['similarities']  
 graph[i]['weight'] += top  
  
 for i in range(len(graph) - 1, -1, -1):  
 centrality = 0  
 if i == 0:  
 for link in graph[i]['links']:  
 centrality += (  
 graph[i]['weight'] / next((node for node in graph if node['index'] == link['index']), None)[  
 'weight'])  
 graph[i]['centrality'] = centrality  
 for j in range(i - 1, -1, -1):  
 for link in graph[j]['links']:  
 if link['index'] == graph[i]['index']:  
 centrality += (graph[i]['weight'] / graph[j]['weight'])  
 graph[i]['centrality'] = centrality  
  
 return graph[0]['centrality'], graph

Ek8 – :Merkeziyet ve benzerlik kullanarak sınıflandırma

def class\_predict(data):  
 c1, \_ = malatya\_centrality(data, top\_accepted)  
 c2, \_ = malatya\_centrality(data, top\_rejected)  
  
 if c1 > c2:  
 return 1  
 else:  
 return 0

Ek9 – projenin kodu :

[malek721/KrediRiskML](https://github.com/malek721/KrediRiskML)

Ek10– kullanılan veri seti :

[Financial Risk Data - Large](https://www.kaggle.com/datasets/deboleenamukherjee/financial-risk-data-large)

AdSoyad: MUHAMMED MALİK ELHIDIR  
Öğrenci No: 02210224068  