logo, yazı tipi, simge, sembol, grafik içeren bir resim

Açıklama otomatik olarak oluşturuldu

**YALOVA ÜNİVERSİTESİ MÜHENDİSLİK FAKÜLTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ**

**MAKİNE ÖĞRENİMİ**

**OTİZM VERİ SETİ İLE MAKİNE ÖĞRENİMİ**

**190101090- Lütfi Yiğit SAĞLAM**

**YALOVA, 2024**

**PROJE ÖZETİ**

Bu proje, küçük çocuklarda otistik spektrum bozukluğu tarama verileri ile çocukların otistik spektrum bozukluğuna sahip olup olmamasıyla ilgili sınıflandırma yöntemi çalışmasını içermektedir. Bu veri setinde, ASD vakalarını davranış bilimlerinde kontrol grubundan ayırt etmede etkili olduğu kanıtlanmış on davranışsal özellik (Q-Chat-10) ve diğer bireylerin özellikleri kaydedilmiştir.

Toplam veri sayısı: 3679

Öznitelik sayısı: 18

Öznitelikler:

A1-A10: Q-Chat-10 içindeki öğeler, soruların olası cevapları: "Her zaman, Genellikle, Bazen, Nadiren ve Hiç" öğelerinin değerleri veri setinde "1" veya "0" olarak eşlenmiştir. Q-chat-10'daki 1-9 sorular için (A1-A9) cevap Bazen / Nadiren / Hiç ise, o zaman soruya (A1-A9) "1" atanır. Ancak, 10. soru için (A10), cevap Her Zaman / Genellikle / Bazen ise, o zaman o soruya "1" atanır. Kullanıcı 3'ten fazla puan alırsa, Q-chat-10 puanı) o zaman potansiyel ASD özellikleri gözlemlenir, aksi halde ASD özellikleri gözlenmez. Veri setindeki geri kalan özellikler, ASDTests tarama uygulamasını kullanırken kullanıcı tarafından elde edilen skora dayalı olarak otomatik olarak atanmıştır.

Modelin performansını değerlendirmek için 10 Fold Cross Validation uygulanarak, Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, XGBoostClassifier, Ada Boost Classifier, SVC Classifier algoritmaları kullanılmıştır. Accuracy, Precision, Recall ve F1 Score değerleri görselleştirilmiştir. Feature Selection yöntemlerinden PCA ve Variance Thresholding ile de modelin performans sonuçları karşılaştırılmıştır.

**VERİ SETİ ANALİZİ**

Veri setindeki otizm dağılım yüzdesi aşağıdaki grafikle görselleştirilmiştir;

metin, ekran görüntüsü, logo, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

Veri setindeki yaşlara göre otizm dağılım grafiği bar grafiğinde gösterilmiştir;

metin, ekran görüntüsü, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

Cinsiyete göre ASD Vaka sayısı grafiği aşağıdaki gibi görselleştirilmiştir;

metin, ekran görüntüsü, dikdörtgen, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

daire, diyagram, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Veri setindeki her bir özelliğin dağılımı aşağıdaki gibi görselleştirilmiştir;

ekran görüntüsü, tasarım içeren bir resim

Açıklama otomatik olarak oluşturuldu

ekran görüntüsü, metin, tasarım içeren bir resim

Açıklama otomatik olarak oluşturuldu

**KULLANILAN MODELLER**

1. **Logistic Regression**:
   * Lojistik regresyon, sınıflandırma problemlerinde kullanılan temel bir lineer modeldir.
   * İkili sınıflandırma yaparken sınıf olasılıklarını hesaplar ve genellikle sigmoidal bir fonksiyonla sonuçları tahmin eder.
2. **Decision Tree Classifier**:
   * Karar ağaçları, veriyi çeşitli dallara ayırarak karar verme sürecini bir ağaç yapısı şeklinde modelleyen açıklanabilir algoritmalardır.
   * Hedef değişkeni tahmin etmek için dalların sonundaki yaprakları kullanarak veri setini parçalara böler.
3. **Random Forest Classifier**:
   * Rastgele ormanlar, birçok karar ağacını birleştirerek çalışan bir topluluk yöntemidir ve genellikle daha yüksek doğruluk sağlar.
   * Bu model, her bir ağacın tahminlerini birleştirerek aşırı uyum riskini azaltır ve genellikle daha iyi genelleme performansı sergiler.
4. **Gradient Boosting Classifier**:
   * Gradient boosting, hataları azaltmak için zayıf tahmincilerin (genellikle karar ağaçları) ardışık olarak birleştirilmesiyle çalışan güçlü bir topluluk yöntemidir.
   * Her yeni model, önceki modelin hatalarını düzeltmeye çalışarak adım adım iyileştirme yapar.
5. **XGBoost (XGBClassifier)**:
   * XGBoost, gradient boosting algoritmasının optimize edilmiş bir versiyonudur ve özellikle hız ve performans açısından etkilidir.
   * Düzenlemeler ve çeşitli optimizasyon teknikleri kullanarak daha hızlı ve daha doğru sonuçlar elde etmeyi amaçlar.
6. **AdaBoost Classifier**:
   * AdaBoost, adaptif boosting algoritması olarak bilinir ve zayıf tahmincilerin bir topluluğunu kullanarak doğruluğu artırır.
   * Her bir ardışık model, önceki modelin yanlış sınıflandırdığı örneklere daha fazla ağırlık vererek hataları azaltmaya çalışır.
7. **Support Vector Classifier (SVC)**:
   * SVC, destek vektör makinelerinin sınıflandırma versiyonudur ve iki sınıf arasındaki en iyi ayırma hiperdüzlemini bulmayı amaçlar.
   * Karar sınırını belirlerken maksimum marjini kullandığı için genellikle yüksek doğruluğa sahip ve aşırı uyuma karşı dayanıklıdır.

**PROJE KODLAMASI ve ÇIKTISI**

Modelin eğitim kısmında 10 katlı çapraz doğrulama nesnesi oluşturulup, modelin performansı değerlendirildi.

metin, ekran görüntüsü, doküman, belge, yazı tipi içeren bir resim

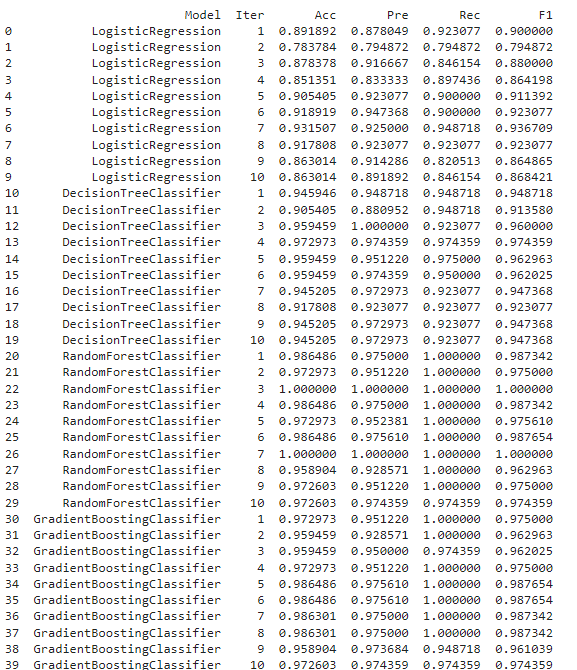
Açıklama otomatik olarak oluşturuldu

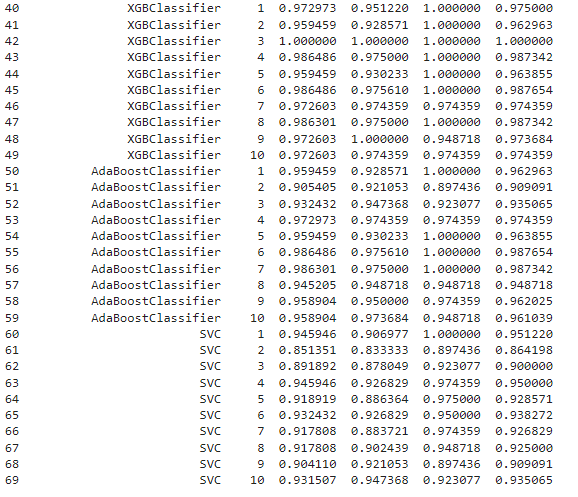
Her model için test verilerinden elde edilen confusion matrix’leri aşağıdaki grafiklerle görselleştirilmiştir;

ekran görüntüsü, kare, dikdörtgen, meneviş mavisi içeren bir resim

Açıklama otomatik olarak oluşturuldu

* **AdaBoostClassifier**: 307 doğru negatif, 350 doğru pozitif tahmini ve 37 yanlış negatif, 42 yanlış pozitif tahmini vardır. Yanlış pozitifler ve yanlış negatifler dengeli değil, ancak modelin genel olarak iyi performans gösterdiğini gösteriyor.
* **DecisionTreeClassifier**: Yanlış pozitifler ve yanlış negatifler minimumdur, bu da modelin iyi bir performans sergilediğini gösterir.
* **GradientBoostingClassifier**: Çok az yanlış pozitif ve yanlış negatif vardır, bu da modelin yüksek doğruluk ve hatırlama sağladığını gösterir.
* **LogisticRegression**: Yanlış pozitifler ve yanlış negatifler minimumdur, bu da modelin iyi bir performans sergilediğini gösterir.
* **RandomForestClassifier**: Çok az yanlış pozitif ve yanlış negatif vardır, bu da modelin yüksek doğruluk ve hatırlama sağladığını gösterir.
* **SVC**: Yanlış pozitifler ve yanlış negatifler dengeli değil, ancak modelin genel olarak iyi performans gösterdiğini gösteriyor.
* **XGBClassifier**: Yanlış pozitifler ve yanlış negatifler dengeli değil, ancak modelin genel olarak iyi performans gösterdiğini gösteriyor.

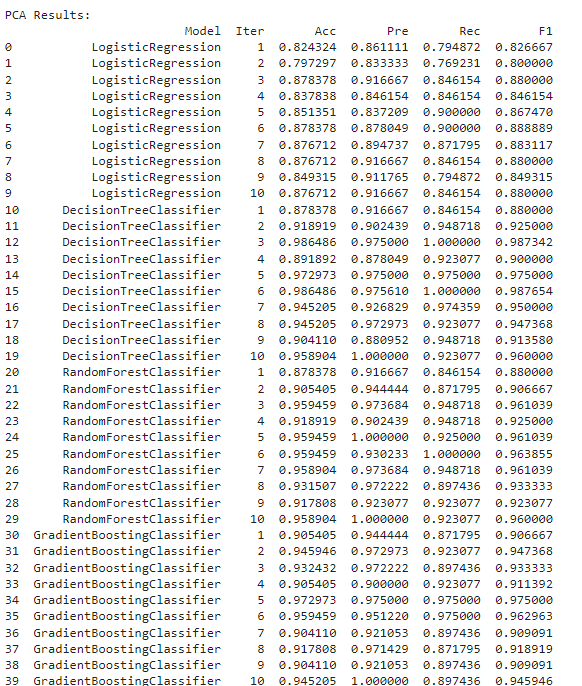


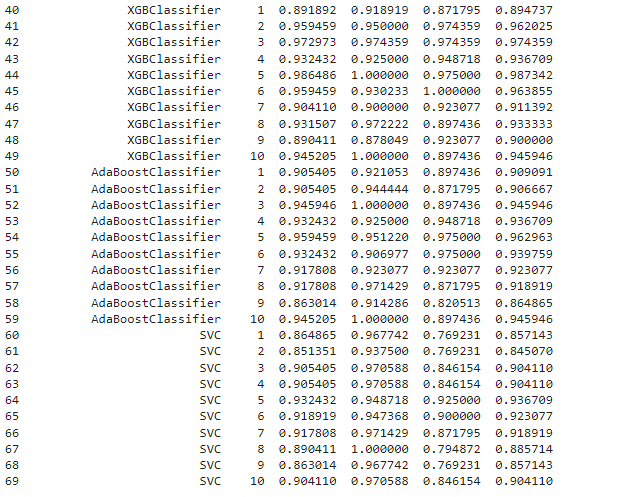


1. **Logistic Regression (Lojistik Regresyon):** Ortalama doğruluk oranı yaklaşık %88 civarında ve diğer performans metrikleri de genellikle dengeli.
2. **Decision Tree Classifier (Karar Ağacı Sınıflandırıcısı):** Çoğu iterasyonda yüksek doğruluk elde edilmiş, ancak bazı iterasyonlarda (örneğin, iterasyon 2) doğruluk oranı düşük olmuş.
3. **Random Forest Classifier (Rastgele Orman Sınıflandırıcısı):** Tüm iterasyonlarda yüksek doğruluk oranları elde edilmiş, hatta bazı iterasyonlarda %100 doğruluk sağlanmış.
4. **Gradient Boosting Classifier (Gradyan Artırma Sınıflandırıcısı):** Genellikle yüksek doğruluk elde edilmiş, ancak bazı iterasyonlarda (örneğin, iterasyon 9) doğruluk biraz düşük olmuş.
5. **XGB Classifier (XGBoost Sınıflandırıcısı):** Diğer gradyan artırma modelleri gibi genellikle yüksek doğruluk sağlamış.
6. **AdaBoost Classifier (AdaBoost Sınıflandırıcısı):** Genellikle yüksek doğruluk sağlamış, ancak bazı iterasyonlarda doğruluk düşük olmuş.
7. **SVC (Support Vector Classifier):** Diğer modellere kıyasla daha düşük doğruluk oranları elde edilmiş, ancak bazı iterasyonlarda yine de yüksek doğruluk sağlanmış.

Sonuç olarak, XGB genellikle en yüksek doğruluk oranlarına sahipken, diğer modeller de genellikle iyi performans göstermiştir.

Veri setimize Feature Selection işlemi olarak ilk önce PCA (Principal Component Analysis (component = 2) sonrasında ise Variance Thresholding (threshold=0.5) uygulandığında ise sonuçlar aşağıdaki gibi gözlemlenmiştir.





1. **Logistic Regression:**
   * Genel olarak iyi bir performans gösterdi, ancak bazı iterasyonlarda diğer modellere kıyasla düşük duyarlılık gözlendi.
2. **Decision Tree Classifier:**
   * Çoğu iterasyonda yüksek performans gösterdi ve özellikle bazı iterasyonlarda mükemmel sonuçlar elde etti.
3. **Random Forest Classifier:**
   * Yüksek doğruluk ve F1 puanlarına sahipti, ancak bazı iterasyonlarda diğer modellere göre biraz düşük hassasiyet gözlendi.
4. **Gradient Boosting Classifier:**
   * Genellikle iyi bir performans gösterdi, ancak bazı iterasyonlarda düşük hassasiyet ve duyarlılık gözlendi.
5. **XGBoost Classifier:**
   * Yüksek doğruluk ve F1 puanlarına sahipti, ancak bazı iterasyonlarda düşük duyarlılık dikkat çekti.
6. **AdaBoost Classifier:**
   * Hassasiyet ve F1 puanlarında genellikle yüksek performans gösterdi, ancak bazı iterasyonlarda düşük doğruluk ve duyarlılık gözlendi.
7. **Support Vector Classifier (SVC):**
   * Hassasiyet ve F1 puanlarında genellikle iyi bir performans gösterdi, ancak bazı iterasyonlarda düşük doğruluk ve duyarlılık gözlendi.

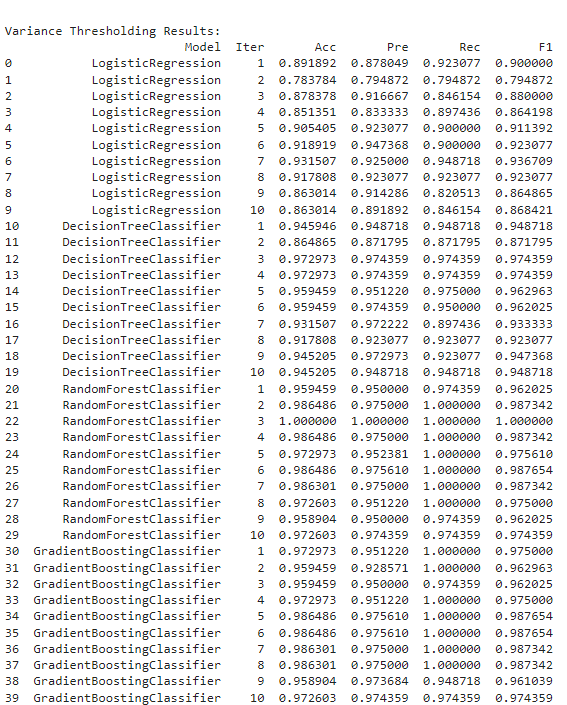
Genel olarak, karşılaştırma yapıldığında Decision Tree Classifier ve Random Forest Classifier modellerinin diğerlerine kıyasla daha istikrarlı bir performans sergilediği görüldü.

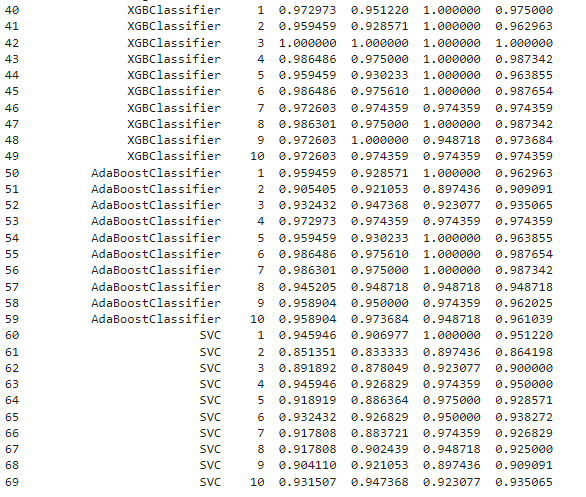
Component 1'de en yüksek ağırlığa sahip özellikler:

* A1
* A2
* A3
* A4
* A5
* A6
* A7
* A8
* A9
* A10\_Autism\_Spectrum\_Quotient
* Qchat\_10\_Score

Component 2'de en yüksek ağırlığa sahip özellikler:

* Jaundice
* Family\_mem\_with\_ASD
* Ethnicity\_others
* Who\_completed\_the\_test\_Health Care Professional
* Who\_completed\_the\_test\_Others
* Who\_completed\_the\_test\_Parent
* Who\_completed\_the\_test\_Relative
* Who\_completed\_the\_test\_School and NGO
* Who\_completed\_the\_test\_Self





1. **Logistic Regression:**
   * Genel olarak makul bir performans sergiledi, ancak bazı iterasyonlarda doğruluk ve F1 puanlarında düşüşler gözlendi.
2. **Decision Tree Classifier:**
   * Yüksek doğruluk ve F1 puanlarına sahipti, ancak modelin istikrarı bazı iterasyonlarda değişkenlik gösterdi.
3. **Random Forest Classifier:**
   * Yüksek performans gösterdi ve varyans eşiği belirleme işleminden az etkilendi.
4. **Gradient Boosting Classifier:**
   * Yüksek performans gösterdi, ancak bazı iterasyonlarda düşük doğruluk ve hassasiyet dikkat çekti.
5. **XGBoost Classifier:**
   * Genel olarak yüksek performans sergiledi ve varyans eşiği belirleme işleminden az etkilendi.
6. **AdaBoost Classifier:**
   * İyi bir performans gösterdi, ancak modelin istikrarı bazı iterasyonlarda değişkenlik gösterdi.
7. **Support Vector Classifier (SVC):**
   * İyi bir performans sergiledi, ancak varyans eşiği belirleme işleminden daha fazla etkilendiği gözlendi. Özellikle bazı iterasyonlarda doğruluk ve F1 puanlarında düşüşler gözlendi.

Varyans değeri 0.5 olarak belirlendiğinde seçilen öznitelikler ‘Age\_Years’ ve ‘Qchat\_10\_Score’ olmuştur.

**Kaynakça**

<https://medium.com/machine-learning-t%C3%BCrkiye/makine-ogrenmesi-7cfbb3d859db>

<https://www.kaggle.com/datasets/fabdelja/autism-screening-for-toddlers>