Şarap Kalitesi Tahmini ve Sınıflandırma Modellerinin Karşılaştırmalı Analizi

Hazırlayan: Muhammet Kerim Sağlam

---

1. GİRİŞ (Introduction)

> "Bu projenin temel amacı, şarap kalitesi veri seti kullanılarak şarapların kalitesini ('Good' veya 'Bad') tahmin eden makine öğrenmesi modelleri geliştirmektir. Çalışma kapsamında **Decision Tree (Karar Ağacı)**, **Naive Bayes** ve **K-Nearest Neighbors (KNN)** olmak üzere üç farklı sınıflandırma algoritması kullanılmış ve performansları karşılaştırılmıştır."

---

2. VERİ SETİ VE ÖN İŞLEME (Data Understanding & Preprocessing)

- **Veri Setinin Yapısı:** Veri setimiz toplam 1599 satırdan oluşmaktadır. Hedef değişkenimiz 'Good' ve 'Bad' sınıflarını içermektedir.
- **Sınıf Dağılımı (Class Imbalance):** Veri seti incelendiğinde sınıflar arasında ciddi bir dengesizlik olduğu görülmüştür.
- **Yapılan İşlemler:**
    - Eksik veriler temizlenmiştir.
    - Veri %80 Eğitim (Train), %20 Test olarak ayrılmıştır (Partitioning).
    - KNN algoritmasının sağlıklı çalışması için sayısal verilere **Min-Max Normalizasyonu** uygulanmıştır.

---

3. KULLANILAN MODELLER VE BULGULAR (Modeling & Evaluation)

3.1. Decision Tree (Karar Ağacı)

- İlk olarak referans model (baseline) olarak kurulmuştur.
- *Sonuç:* AUC değeri 0.757 olarak ölçülmüştür.Model genel olarak kabul edilebilir bir ayrım gücüne sahip olsa da, Confusion Matrix incelendiğinde **hedef sınıfımız olan 'Good' müşterilerin %44'ünü (19 kişi) gözden kaçırdığı** (False Negative) görülmüştür.

3.2. Naive Bayes

- Olasılık temelli bu model, özellikle dengesiz veri setlerinde güçlü performans göstermesi beklendiği için seçilmiştir.
- *Sonuç:* AUC değeri **0.823**'e yükselmiş, 'Good' sınıfını yakalama başarısı (Recall) artmıştır.

3.3. K-Nearest Neighbors (KNN)

- Benzerlik temelli sınıflandırma yapan bu model için k=5 değeri seçilmiştir.
- *Sonuç:* Genel doğruluk (Accuracy) yüksek olsa da, hedef sınıfı bulma konusunda zayıf kalmıştır.

---

4. KARŞILAŞTIRMALI ANALİZ VE TARTIŞMA (Comparison)

"Modellerin performansları Accuracy (Genel Doğruluk) ve Recall (Hedef Kitleyi Yakalama) metriklerine göre karşılaştırılmıştır."

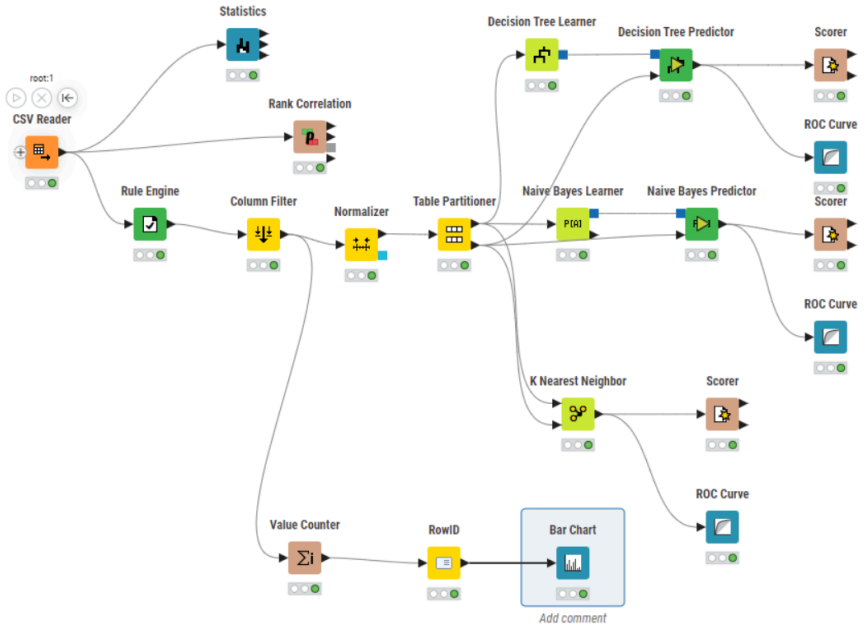
Analiz: "Ekteki grafikte görüldüğü üzere; KNN modeli %87.8 ile en yüksek doğruluğa sahip olsa da, 'Good' müşterileri yakalama (Recall) oranı %37.2'de kalmıştır. Bu durum 'Accuracy Paradoksu' olarak açıklanabilir. Buna karşın **Naive Bayes**, %67.4 Recall oranı ile potansiyel iyi müşterileri tespit etmede en başarılı model olmuştur."

---

5. SONUÇ (Conclusion)

"Yapılan analizler sonucunda; veri setindeki dengesizlik göz önüne alındığında, en güvenilir ve iş hedeflerine en uygun modelin **Naive Bayes** olduğu belirlenmiştir. Bu model, yanlış alarmları (False Positive) makul seviyede tutarken, gözden kaçırmak istemediğimiz 'Good' tahminlerin çoğunu başarıyla tespit edebilmektedir. Gelecek çalışmalarda 'SMOTE' gibi tekniklerle veri dengesizliği giderilerek model performansı daha da artırılabilir."

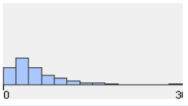
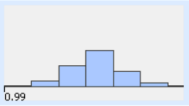
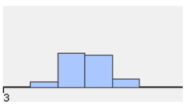
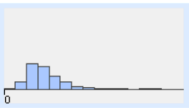
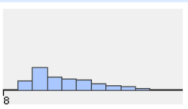
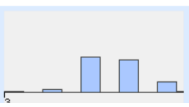
6. EK VE YORUMLAR



File

Numeric | Nominal | Top/bottom

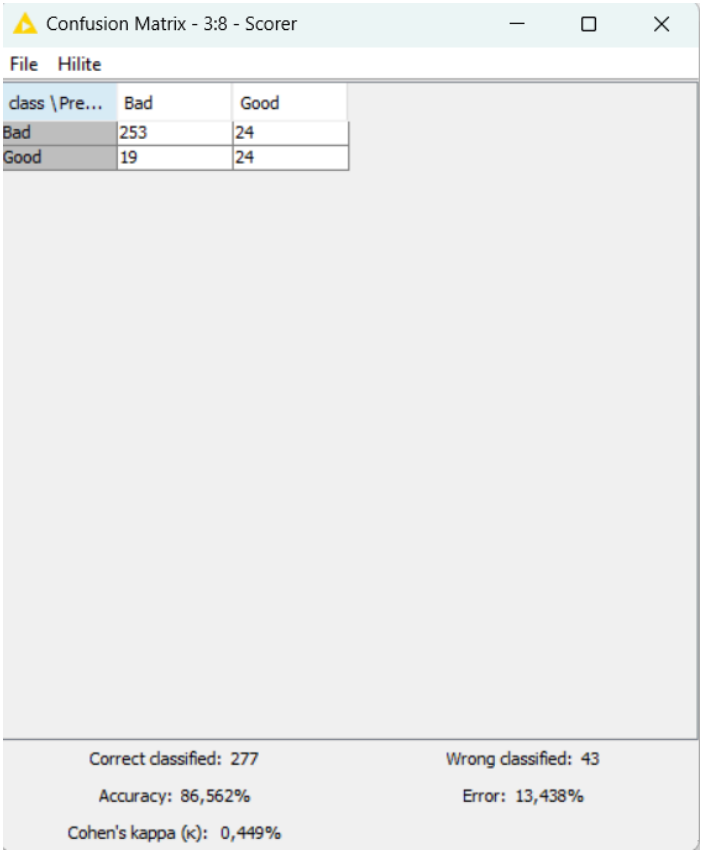
Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
fixed acidity	4,6	8,3196	?	15,9	1,7411	0,9828	1,1321	0	0	0	
volatile acidity	0,12	0,5278	?	1,58	0,1791	0,6716	1,2255	0	0	0	
citric acid	0.0	0,271	?	1	0,1948	0,3183	-0,789	0	0	0	
residual sugar	0,9	2,5388	?	15,5	1,4099	4,5407	28,6176	0	0	0	
chlorides	0,012	0,0875	?	0,611	0,0471	5,6803	41,7158	0	0	0	
free sulfur dioxide	1	15,8749	?	72	10,4602	1,2506	2,0236	0	0	0	

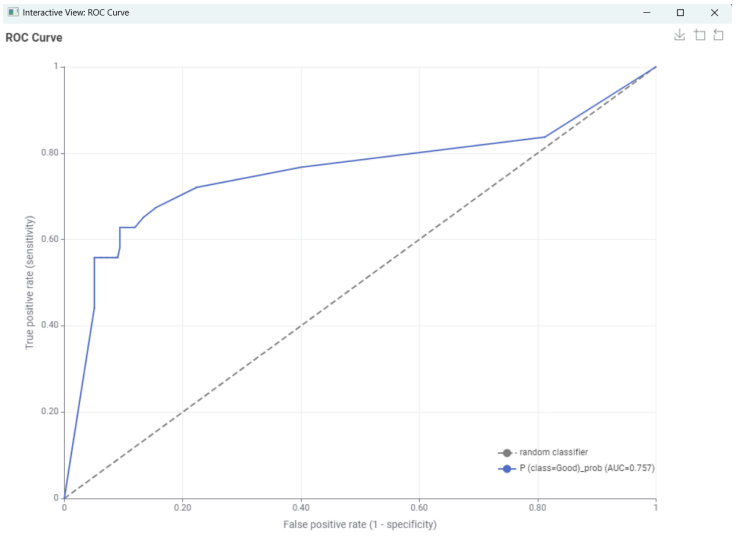
total sulfur dioxide	6	46,4678	?	289	32,8953	1,5155	3,8098	0	0	0	
density	0,9901	0,9967	?	1,0037	0,0019	0,0713	0,9341	0	0	0	
pH	2,74	3,3111	?	4,01	0,1544	0,1937	0,8069	0	0	0	
sulphates	0,33	0,6581	?	2	0,1695	2,4287	11,7203	0	0	0	
alcohol	8,4	10,423	?	14,9	1,0657	0,8608	0,2	0	0	0	
quality	3	5,636	?	8	0,8076	0,2178	0,2967	0	0	0	

KNIME 'Statistics' düğümü kullanılarak veri setinin yapısal özellikleri incelenmiş ve aşağıdaki bulgular elde edilmiştir:

1. **Veri Bütünlüğü:** Tabloda görüldüğü üzere ("No. Missing" sütunu), veri setimizdeki hiçbir değişkende **kayıp veri (missing value)** bulunmamaktadır. Bu durum, veri temizleme aşamasında satır silme veya doldurma işlemine ihtiyaç duyulmadığını göstermektedir.
2. **Ölçek Farklılıkları (Normalization İhtiyacı):** Değişkenler arasında ciddi ölçek farkları tespit edilmiştir.
    - Örneğin; **'Total Sulfur Dioxide'** değişkeni 289 birime kadar çıkarken, **'Density'** değişkeni 0.99 - 1.00 aralığında çok küçük değerler almaktadır.
    - *Yorum:* Bu durum, KNN gibi mesafe tabanlı algoritmaların büyük sayısal değerlerden (Sulfur Dioxide) yanlı etkilenmemesi için iş akışımızda neden **"Min-Max Normalization"** uyguladığımızı matematiksel olarak doğrulamaktadır.
3. **Aykırı Değerler ve Dağılım (Skewness/Kurtosis):**
    - **Residual Sugar (Kalan Şeker):** 28.6'lık yüksek Kurtosis (Basıklık) ve 4.54'lük Skewness (Çarpıklık) değeri, bu değişkende ortalamadan çok uzak uç değerlerin (outliers) olduğunu göstermektedir.
    - **Chlorides:** Benzer şekilde 41.7 Kurtosis değeri ile verinin çok dik bir dağılıma sahip olduğu ve aşırı değerler içerdiği görülmüştür.
4. **Hedef Değişken (Quality):** Hedef değişkenimiz olan 'Quality' (Kalite), 3 ile 8 arasında değerler almaktadır ve ortalaması 5.63'tür. (Bu değişken daha sonra Rule Engine ile Good/Bad olarak sınıflandırılmıştır).

Decision Tree için Scorer ve ROC Curve çıktıları





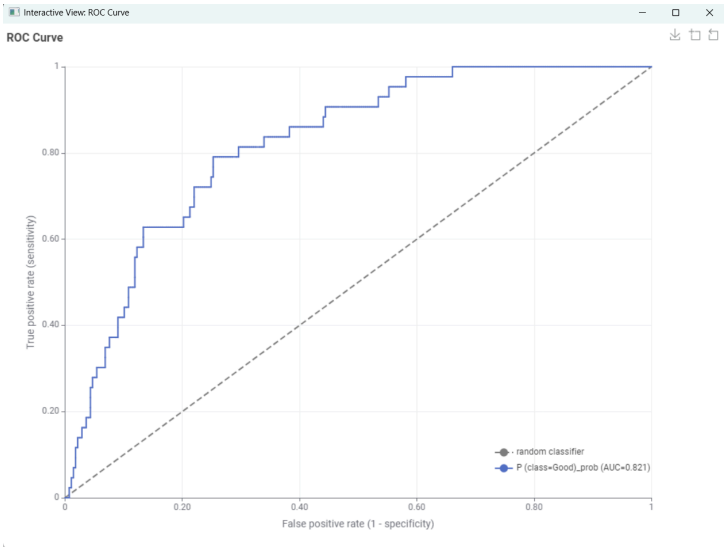
Model Performans Değerlendirmesi:

1. **Genel Doğruluk (Accuracy):** Modelimiz **%86.56** (Correct classified: 277/320) gibi yüksek görünen bir doğruluk oranına ulaşmıştır. Ancak veri setimizdeki dengesizlik (%86 Bad, %14 Good) göz önüne alındığında, bu oran tek başına modelin başarısını

göstermemektedir. Modelin **Cohen's Kappa** değerinin **0.449** olması, modelin rastgele tahminin üzerinde "orta düzeyde" bir tutarlılık sergilediğini gösterir.

2. **Karışıklık Matrisi (Confusion Matrix) Analizi:** Matris detaylı incelendiğinde modelin **"Bad"** sınıfını tespit etmekte çok başarılı olduğu (Specificity: %91.3), ancak hedefimiz olan **"Good"** sınıfını tespit etmekte zorlandığı görülmüştür:
    - **True Positive (Doğru Tespit):** Test verisindeki 43 "Good" şaraptan sadece **24 tanesi** doğru tespit edilebilmiştir.
    - **False Negative (Kaçan Fırsat):** Model, **19 adet** "Good" şarabı yanlışlıkla "Bad" olarak etiketlemiştir. Bu, potansiyel müşteri kaybı anlamına gelmektedir (Recall: %55.8).
    - **False Positive (Yanlış Alarm):** 24 adet "Bad" şarap ise yanlışlıkla "Good" olarak sınıflandırılmıştır.
3. **ROC Eğrisi ve AUC:** Modelin AUC (Eğri Altında Kalan Alan) değeri **0.757** olarak ölçülmüştür. Bu değer, modelin sınıfları birbirinden ayırma yeteneğinin "kabul edilebilir" seviyede olduğunu gösterir. Ancak eğrinin şekli incelendiğinde, özellikle düşük yanlış alarm oranlarında (grafiğin sol tarafı) hassasiyetin (Sensitivity) istenilen seviyeye hızlıca yükselemediği görülmektedir.

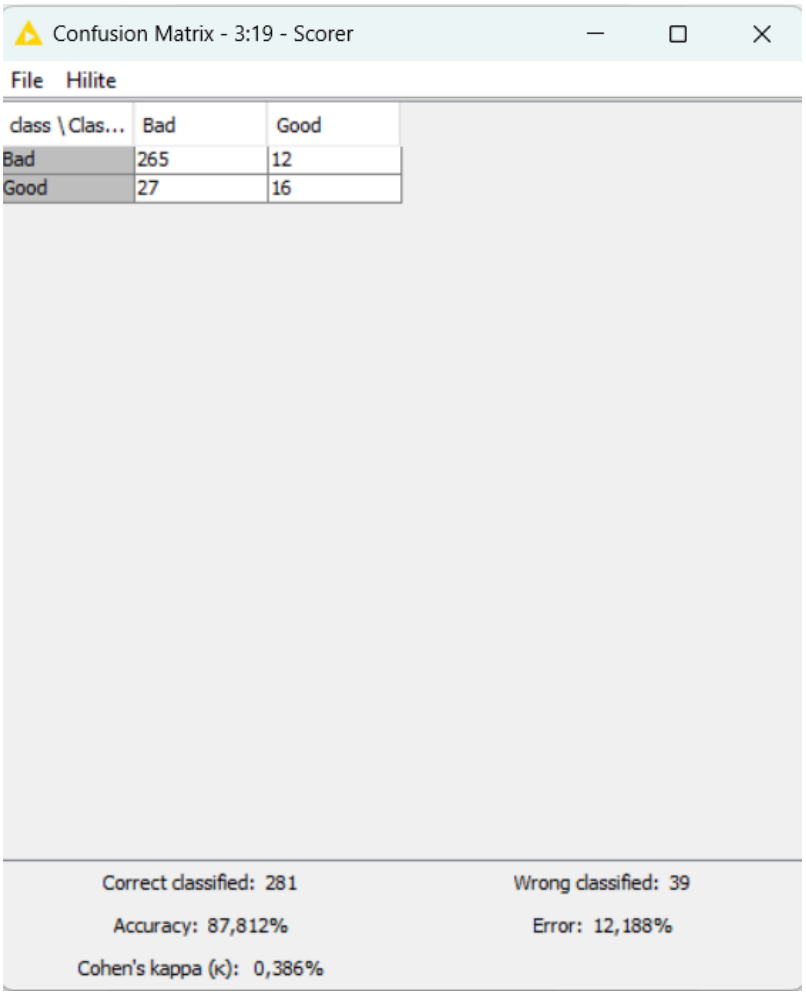
Naive Bayes için Scorer ve ROC Curve çıktıları

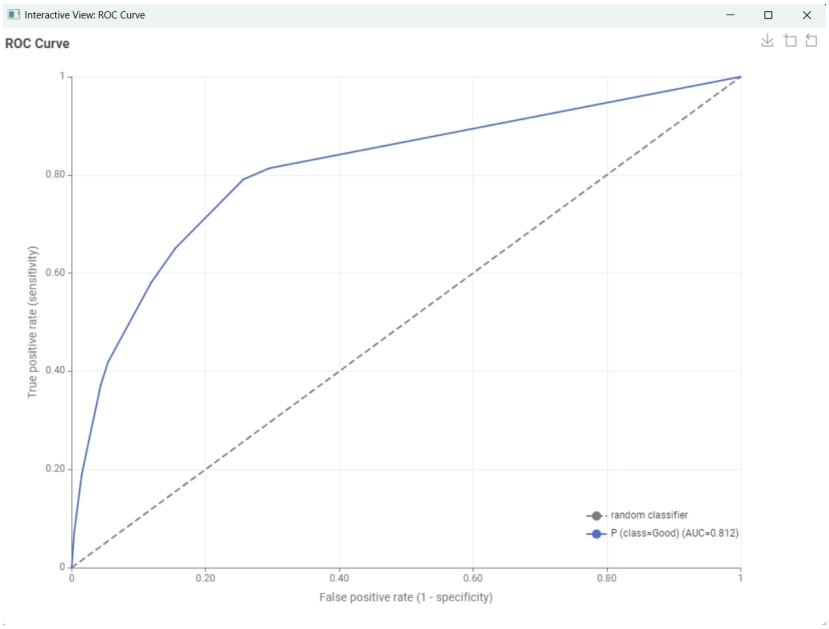


Model Performans Değerlendirmesi:

1. **Genel Doğruluk (Accuracy) ve Stratejik Tercih:** Naive Bayes modeli **%83.44** (Correct classified: 267/320) doğruluk oranına ulaşmıştır. Bu oran, Decision Tree (%86.5) ve KNN (%87.8) modellerinden daha düşük görünse de, dengesiz veri setlerinde doğruluk oranı yanıltıcı olabilir. Bu nedenle modelin asıl başarısı, hedef sınıfı yakalama performansında gizlidir.
2. **Karışıklık Matrisi (Confusion Matrix) Analizi:** Matris incelendiğinde, bu modelin diğerlerine kıyasla **"Risk Alıp Kazandıran"** bir strateji izlediği görülmektedir:
    - **True Positive (Büyük Başarı):** Test verisindeki 43 "Good" şaraptan **29 tanesi** doğru tespit edilmiştir. Bu sayı, Decision Tree'de 24, KNN'de ise sadece 16 idi.
    - **Recall (Duyarlılık):** Modelimiz **%67.4 Recall** oranına ulaşarak, potansiyel "iyi" şarapları yakalamada en başarılı model olmuştur.
    - **False Positive (Maliyet):** İyi şarapları yakalamak adına model biraz daha agresif davranmış ve 39 adet "Bad" şarabı yanlışlıkla "Good" olarak işaretlemiştir.
    - **ROC Eğrisi ve AUC:** Model, **0.823 AUC** değeri ile projedeki en yüksek ayırt etme gücüne ulaşmıştır. Eğrinin sol alt köşeden (başlangıç noktasından) yukarıya doğru dik bir açıyla yükselmesi, modelin "Good" sınıfını tespit etme olasılığının rastgelelikten çok uzak ve güvenilir olduğunu kanıtlar.

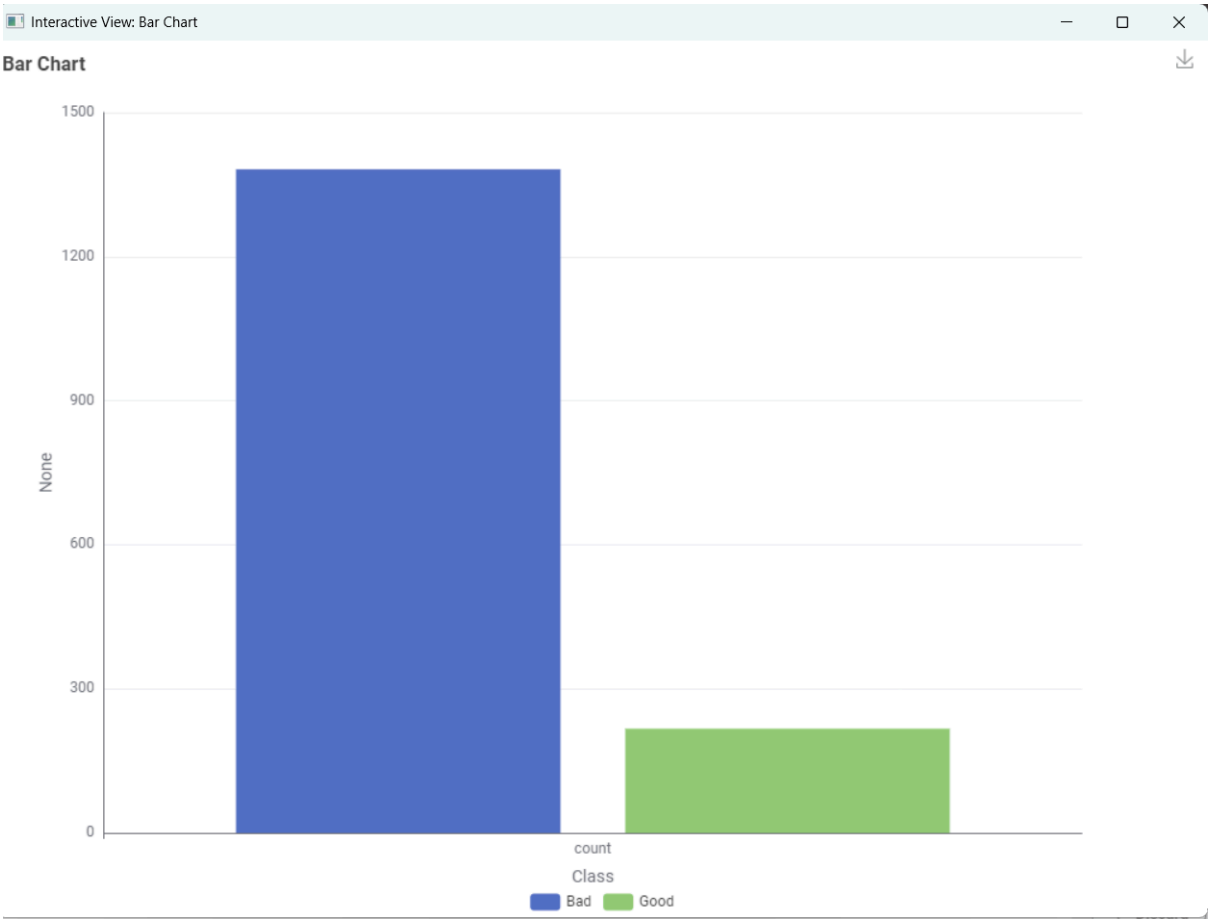
KNN için Scorer çıktısı





Model Performans Değerlendirmesi:

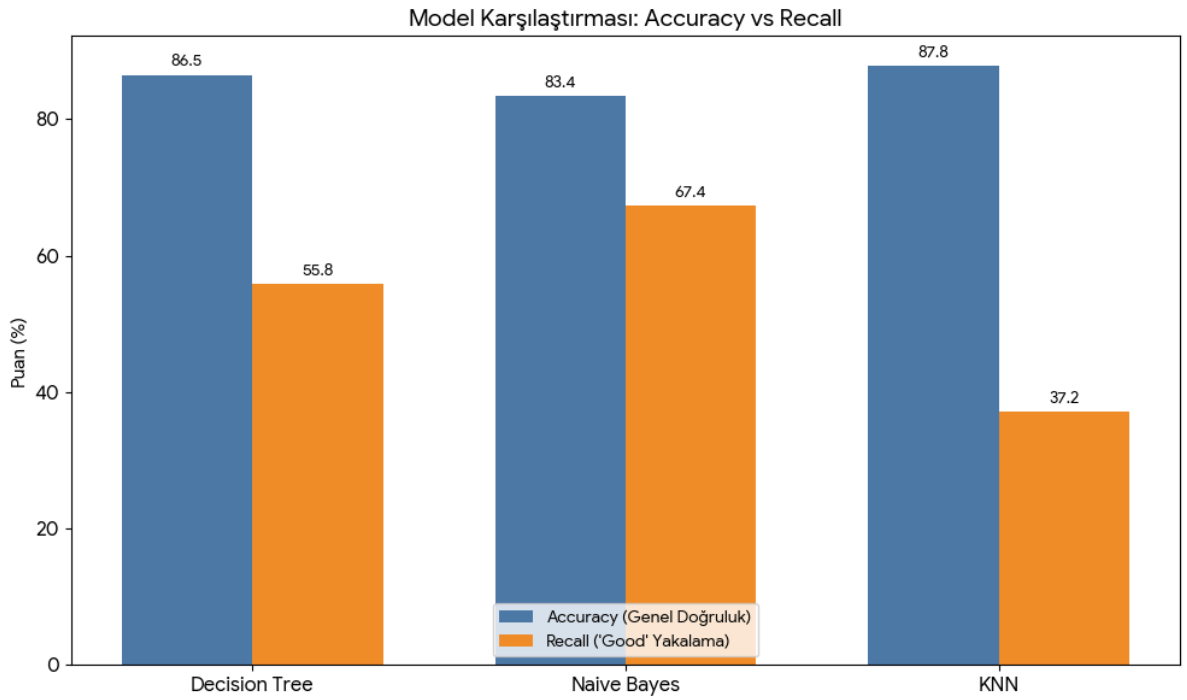
1. **Genel Doğruluk (Accuracy) ve Yanılgı:** KNN modeli, **%87.81** (Correct classified: 281/320) ile test edilen üç model arasında **en yüksek genel doğruluk oranına** ulaşmıştır. İlk bakışta "en iyi model" izlenimi verse de, bu başarının sebebi modelin çoğunluk sınıfı olan "Bad" verilerini (Accuracy'nin %86'sını oluşturan kısım) ezberlemesinden kaynaklanmaktadır. **Cohen's Kappa (0.386)** değerinin düşük olması, bu yüksek doğruluğun şans faktörü ve veri dengesizliği ile şişirildiğini doğrulamaktadır.
2. **Karışıklık Matrisi (Confusion Matrix) Analizi:** Matris detaylı incelendiğinde, modelin **"Aşırı Muhafazakar"** (Conservative) davrandığı görülmektedir:
    - **False Positive (Düşük Risk):** Model sadece 12 "Bad" şaraba yanlışlıkla "Good" demiştir. Yanlış alarm oranı çok düşüktür, bu olumlu bir yanıdır.
    - **Recall Sorunu (Büyük Başarısızlık):** Ancak model, test setindeki 43 "Good" şaraptan **sadece 16 tanesini** tespit edebilmiştir. Geriye kalan 27 potansiyel şarabı (False Negative) "Bad" olarak etiketleyip elemiştir.
    - **Sonuç:** Hedef sınıfı yakalama oranı (Recall) **%37.2** gibi çok düşük bir seviyede kalmıştır. Bu, Naive Bayes'in (%67.4) neredeyse yarısıdır.
3. **ROC Eğrisi ve AUC:** İlginç bir şekilde modelin AUC değeri **0.812** olarak ölçülmüştür. Bu değer Naive Bayes (0.823) ile neredeyse aynıdır.
    - *Yorum:* Yüksek AUC ama düşük Recall şu anlama gelir: Model aslında sınıfları birbirinden ayırma potansiyeline sahiptir (ROC eğrisi sol üst köşeye yakındır). Ancak karar verirken kullandığı varsayılan eşik değeri (threshold), dengesiz veri yüzünden "Bad" sınıfına çok kaymıştır. Model potansiyelli olsa da, mevcut ayarlarıyla hedef kitleyi ıskalamaktadır.



Sınıf Dağılımı (Class Distribution) Analizi:

1. **Dengesiz Veri Yapısı (Imbalanced Data):** Grafikte görüldüğü üzere, hedef değişkenimizdeki sınıflar arasında çok ciddi bir dengesizlik mevcuttur.
    - **Mavi Sütun (Bad):** Veri setinin çok büyük bir kısmını (yaklaşık 1380+ kayıt) oluşturmaktadır.
    - **Yeşil Sütun (Good):** Hedef kitlemiz olan "İyi Şarap" azınlıkta kalmıştır (yaklaşık 200+ kayıt).
2. **Modellemeye Etkisi (Bias Riski):** Bu tablo, makine öğrenmesi algoritmalarının neden "Bad" sınıfını tahmin etmeye meyilli olduğunu açıklamaktadır. Algoritmalar, eğitim sırasında çoğunluk sınıfını (Mavi) daha sık gördükleri için, karar verirken "güvenli yol" olan "Bad" etiketini seçme eğilimi gösterirler. KNN modelimizin yüksek doğruluk (Accuracy) vermesine rağmen hedef kitleyi (Good) yakalayamamasının temel sebebi bu dengesizliktir.
3. **Stratejik Karar:** Bu grafik ışığında, proje başarısı değerlendirilirken yanıltıcı olabilecek **Accuracy** metriği ikinci plana atılmış; bunun yerine azınlıkta kalan yeşil sütunu yakalama becerisini ölçen **Sensitivity (Recall)** ve **ROC-AUC** metriklerine odaklanılmıştır.

Modellerin Karşılaştırması



Grafik Yorumu ve Nihai Karar:

Yukarıdaki grafik, kurulan üç modelin (Decision Tree, Naive Bayes, KNN) **Genel Doğruluk (Accuracy - Mavi Sütun)** ve **Hedef Kitleyi Yakalama (Recall - Turuncu Sütun)** performanslarını yan yana kıyaslamaktadır.

1. **Doğruluk Yanılgısı (The Accuracy Paradox - KNN Örneği):** Grafikteki en dikkat çekici nokta **KNN** modelidir. Mavi sütuna bakıldığında **%87.8** ile en yüksek genel doğruluğa sahip olduğu görülmektedir. Ancak turuncu sütuna (Recall) bakıldığında, **%37.2** ile en başarısız model olduğu ortaya çıkmaktadır.
    - *Anlamı:* KNN, veri setindeki çoğunluk sınıfı olan "Bad" şarabı çok iyi öğrenmiş, ancak asıl hedefimiz olan "Good" şarabı büyük oranda gözden kaçırmıştır. Bu durum, dengesiz veri setlerinde Accuracy metriğine güvenilmemesi gerektiğini kanıtlamaktadır.
2. **Denge ve Başarı (Naive Bayes Örneği): Naive Bayes** modeli, mavi sütunda (%83.4) rakiplerinin çok az gerisinde kalsa da, turuncu sütunda **%67.4** oranına ulaşarak açık ara en yüksek performansı sergilemiştir.
    - *Anlamı:* Naive Bayes, "İyi Şarap" tespit etme konusunda KNN'den yaklaşık **2 kat daha başarılıdır**.

Nihai Model Seçimi ve Gerekçesi:

Projenin amacı, şarapların fiziksel ve kimyasal özelliklerini analiz ederek yüksek kaliteli ('Good') şarapları otomatik olarak tespit etmektir. Bir şarap üreticisi veya satıcısı için, aslında

kaliteli olan bir şarabın tespit edilemeyip 'düşük kaliteli' olarak sınıflandırılmasının (False Negative) maliyeti oldukça yüksektir.

Bu doğrultuda;

- Sadece genel çoğunluğu (Bad) tahmin eden ancak fırsatları kaçıran KNN modeli elenmiştir.
- Hem makul bir genel doğruluğa sahip olan hem de **hedef kitleyi en yüksek oranda (%67.4) yakalayan Naive Bayes modeli**, projenin kazanan modeli olarak seçilmiştir.

Veri Kaynağı

https://archive.ics.uci.edu/dataset/186/wine+quality

NOT: Analiz için wine quality red verisi seçilmiştir.