Bursa Teknik Üniversitesi

Veri Madenciliğine Giriş

Proje Raporu

Ad: Eren

Soyad: KÖSE

Öğrenci Numarası: 22360859075

Proje Hakkında

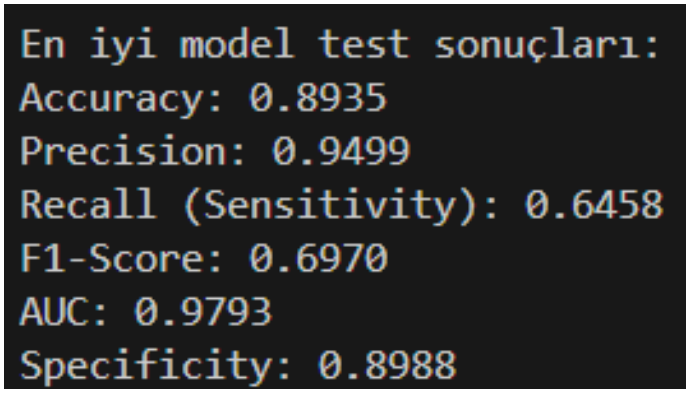
- **Açıklama:** Öğrencilerin 2 yıllık Matematik, Bilim, İngilizce dersleri için toplanan not verileri, cinsiyetleri, 2017/18 öğretim yılından hesaplanan yaşları, önerilen yıl/sınıfları, okula kabul yılları, önceki okullarında okudukları müfredatları, şu anki müfredatları, önceki okuldaki yıl/sınıfı, okula giriş sınavı matematik notu, okula giriş sınavı bilim notu, okula giriş sınavı ingilizce notu bilgileri bu veri setinde yer almaktadır. Bu veriler kullanılarak öğrencilerin "High", "Medium", "Low" şeklinde 3 farklı seviyeye ayrılması amaçlanmıştır. Bu amaç doğrultusunda öğrencilerin not ortalamalarının 85 ve üzeri olması durumu "High" seviye öğrenci, 75 ve 85 aralığı "Medium" seviye öğrenci, 75 altı ise "Low" seviye öğrenci olarak tanımlanmıştır. Bu tanımlama sonrasında ise model, gerekli öznitelikler kullanılarak eğitilmiştir.
- **Kullanılan Veri Seti:** https://www.sciencedirect.com/science/article/pii/S235234092100192X
- **Referans Alınan Makale:** https://link.springer.com/article/10.1007/s43926-022-00023-0
- **Kullanılan Yöntem:** Random Forest Classifier
- **Github Reposu:** https://github.com/erennkose/data-mining-project
- **Tanıtım Videosu:** https://youtu.be/8dG-FBh9IGY

Modelim Hakkında Bilgiler

Projemde geliştirdiğim kod öncelikle veri setini içeri aktarır. Bu veri setindeki boşluk, tırnak gibi gereksiz karakterlerin temizliği yapılır. Bu temizlik sayesinde model performansı yükselmektedir. Veri setinden çekilen verilerde kullanılan öznitelikler aşağıdaki tabloda belirtilmiştir.

Öznitelik Numarası	Öznitelik Adı
1	Gender
2	Age as of Academic Year 17/18
3	Previous Curriculum (17/18)2
4	Math20-1
5	Science20-1
6	English20-1
7	Math20-2
8	Science20-2
9	English20-2
10	Math20-3
11	Science20-3
12	English20-3

Veri setinde bulunan eksik değer içeren satırlar dropna() fonksiyonu ile kaldırılmıştır. Modelimin hedef değişkeni ise öğrencilerin kategorik seviyeleridir. Belirtilen öznitelikler kullanılarak öğrenciler belirli seviyelere atanmıştır. Tahmin için kullanılacak bağımsız değişkenler ve bu bağımsız değişkenlere bağlı olacak öğrenci seviyesi belirtilmiştir. Bunların ardından kategorik değişkenler (Gender, Previous Curriculum gibi) one-hot encoding yöntemiyle binary formata çevrilir. Ayrıca öğrencilerin not değerleri yaş gibi değerlere kıyasla değer olarak çok daha baskın olacağından bunu önleyip gerçekçi bir deneyim elde etme amaçlı öğrenci notlarına normalizasyon işlemi uygulandı. Sonrasında ise veri seti % 70 eğitim (train), % 10 doğrulama (validation), %20 test verisi olacak şekilde ayrılır. Stratify parametresi ile sınıfların orantılı dağılması sağlanır. Bu ayrılan veriler kullanılarak 4 farklı model (50, 100, 200 ve 500 ağaçla) eğitilir. Bunlar arasından genel olarak en dengeli sonuçları veren 200 ağaçlık modelin sonuçlarını paylaşacağım. Bu 200 ağaçlık modeli seçme kararını almak için gerekli Random Forest Classifier parametrelerinin olası en iyi sonuç verebilecek değerleri farklı kombinasyonlarla eğitimde denenmiştir. Bu denemeler sonucunda {'criterion': 'gini', 'max_depth': 15, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200} parametre değerlerinin bulunduğu modelin en iyi sonuçları verdiği kararlaştırılmıştır.



Şekil 1 Model Değerleri

Eğittiğim Modelin Sonuçları

Projemde yaptığım araştırmalar ve yapay zekalara danışmalarım sonucunda veri setimde Random Forest Classifier yönteminin amacıma daha uygun bir yöntem olduğunu keşfettim ve bu yöntemi seçtim. Bu yöntem sayesinde makalede bulunan değerlerin biraz üstüne çıkmayı başardım. Bunun bir diğer sebebi de gürültü yaratabilecek ve model için mantıksız olan özniteliklerin kullanılmamasıdır. Aşağıda verilen ROC eğrileri sonucunda görüldüğü üzere train verilerinin başarısı kadar test verileri de başarılı şekilde tahmin edilmiş durumda. Bu da overfittingin modelimde söz konusu olmadığını gösteriyor.

Şekil 4'te görüldüğü üzere karmaşıklık matrisinde öğrenciler gayet başarılı şekilde gerekli seviyelere ayrılıyor. Veri seti büyüklüğüne kıyasla ufak seviyede hatalarla karşılaşılıyor. En göze batan hatalı tespit kısmı gerçekte "High" seviyede olan öğrencilerin "Medium" olarak tahmin edilmesidir.

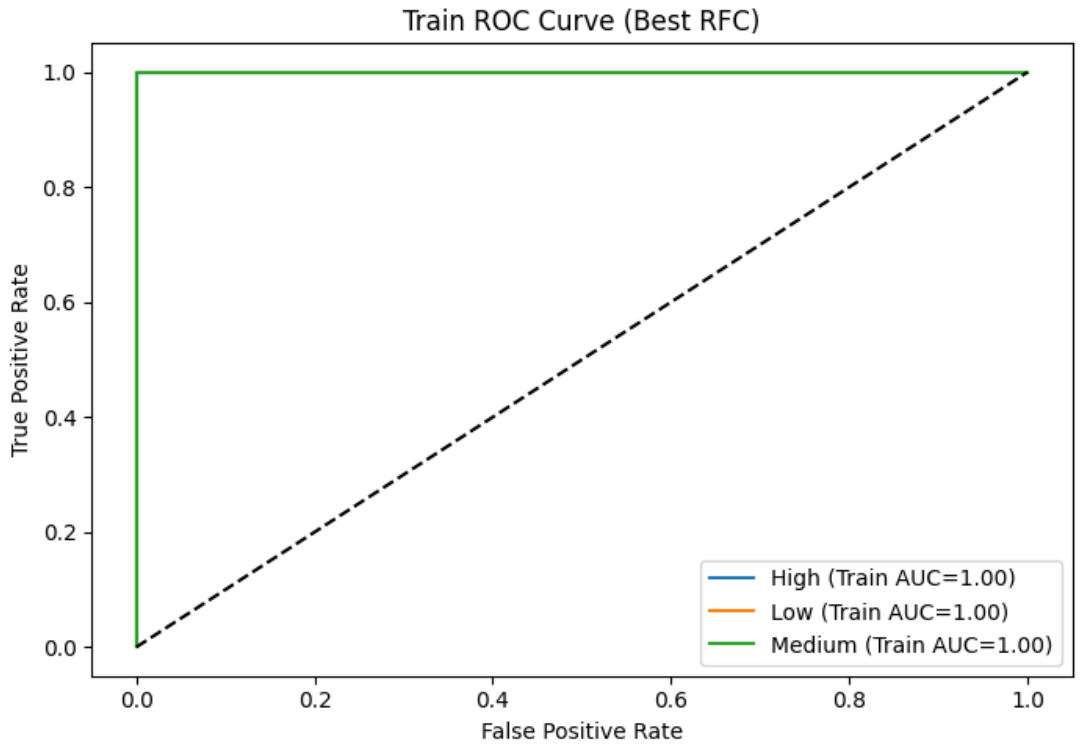
Şekil 5'e baktığımızda da modelimin performans metriklerini görebiliyoruz. Bu performans metriklerinden Accuracy değerleri modelin tahminlerinin genel olarak doğru olduğunu belirtiyor. Bu, tüm sınıflarda modelin başarılı olduğunu ortalama olarak gösteriyor. AUC değerleri ise modelin sınıfları ayırt etme kabiliyetinin çok yüksek olduğunu belirtiyor. ROC eğrisi altında kalan alan çok yüksek

olduğu için modelin sınıfları ayrıştırma gücü de o kadar yüksek oluyor. Specifity değeri de AUC ve accuracy gibi yüksek seviyede, yani hatalı sınıflandırma düşük seviyede. Precision, Recall ve F-1 Skoru değerlerinin biraz diğerlerine kıyasla düşük kalmasının sebebi ise veri setimde bulunan verilerin dengesiz dağılımından kaynaklanıyor. High, Medium ve Low seviyede olan öğrencilerin dağılımları eşit sayıda değil. Bundan dolayı Precision, Recall ve F-1 Skoru değerleri, accuracy ve AUC değerlerine kıyasla biraz düşük kalıyor fakat ona rağmen kabul edilebilir durumda.

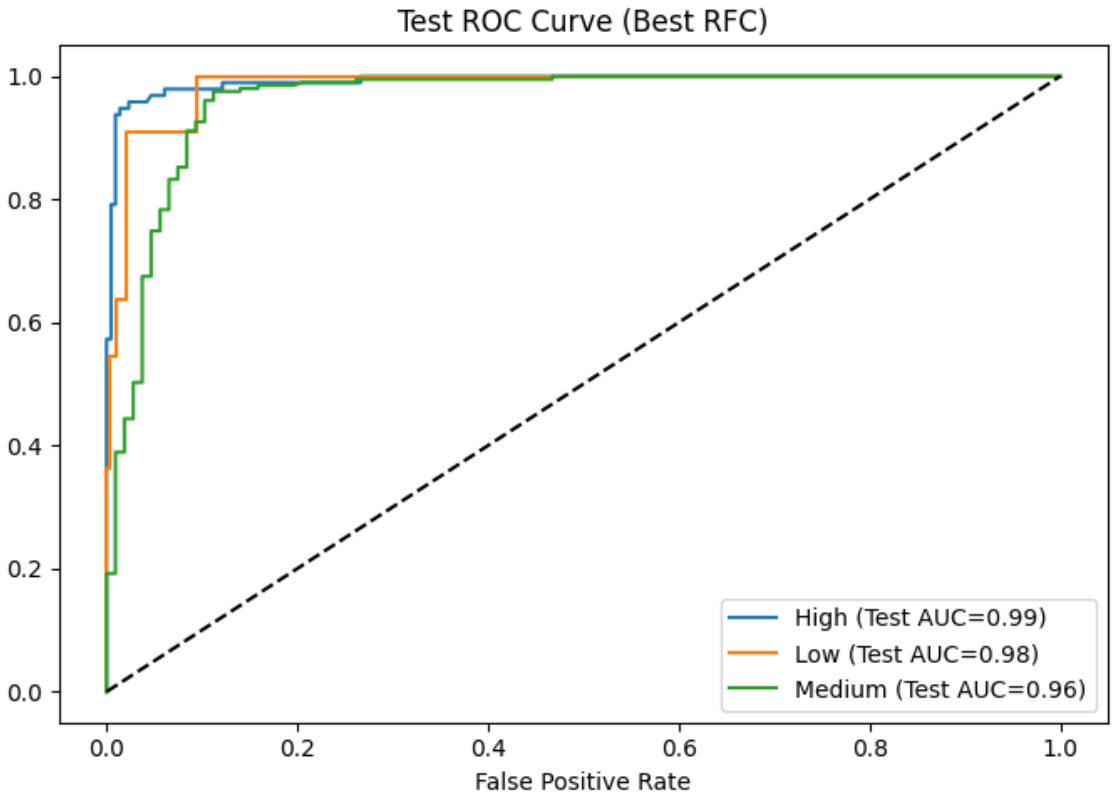
Şekil 6'ya baktığımızda önem sıralamalarını görebiliyoruz. Burada neredeyse 0 olan değerleri fark edebiliriz. Bunun göze çarpan iki sebebi vardır. Bu değerlerin 0 olmasının ilk ve en önemli nedeni veri setimin dengesiz bir dağılıma sahip olmasıdır. Mesela yaklaşık 1500 veri içerisinde yalnızca 6 adet Previous Curriculum (17/18)2 özniteliği Indian olan öğrenci var. İkinci sebebi ise benim kategorik verilerde uygulamış olduğum one-hot encoding işlemidir. Bu iki sebep birleştiğinde bu öznitelik için önem değeri neredeyse 0 oluyor.

Şekil 7'ye baktığımızda, erkek ve kadın öğrencilerin sınav ortalamaları karşılaştırılmış ve her iki grubun başarı düzeylerinin oldukça benzer olduğu görülmektedir. Kutu grafiğinde, her iki cinsiyetin medyan değerleri neredeyse eşit olup, notların dağılımı da büyük ölçüde benzerlik göstermektedir. Erkek öğrencilerde notlar biraz daha geniş bir aralıkta dağılmış olsa da, bu fark çok belirgin değildir. Her iki grupta da bazı düşük notlar aykırı değer olarak öne çıkmaktadır. Özellikle kadın grubunda 59 gibi oldukça düşük bir değer dikkat çekmektedir. Genel olarak, cinsiyete göre sınav ortalamaları arasında anlamlı bir fark bulunmadığı ve her iki grubun da benzer başarı seviyelerine sahip olduğu söylenebilir.

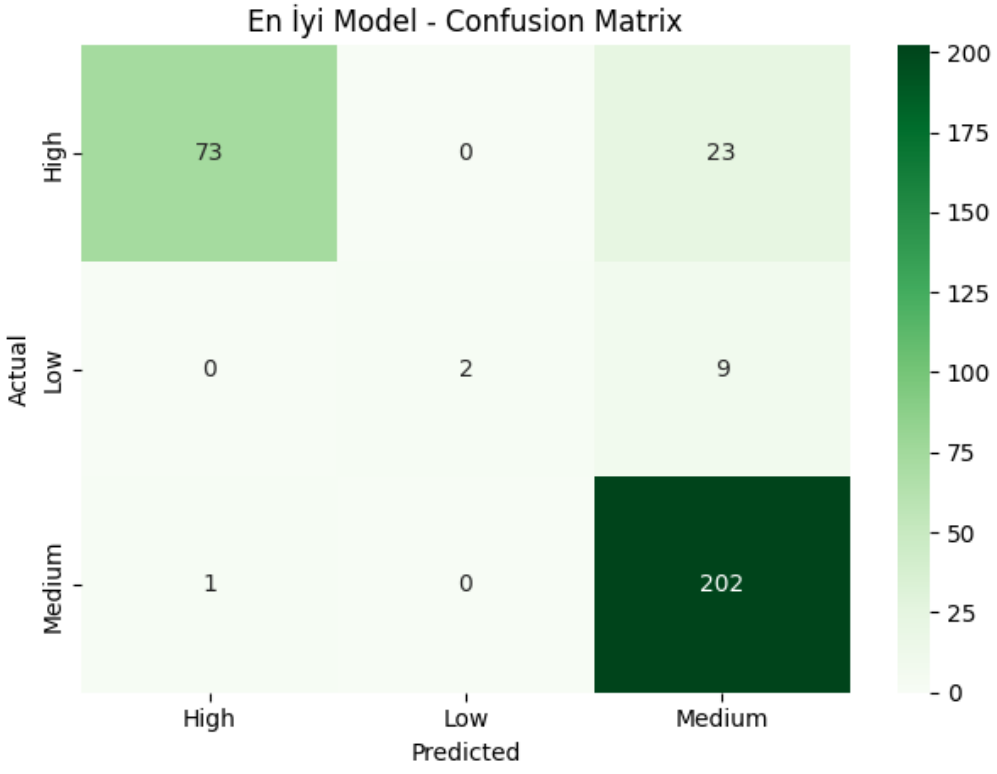
Şekil 8'e baktığımızda, öğrencilerin geçmişte aldıkları müfredata göre sınav ortalamalarını görmekteyiz. Hint, CBSE ve Avustralya müfredatına sahip öğrenciler genelde daha yüksek ortalamalara sahiptir. Amerikan ve Britanya müfredatlarında ise not dağılımları daha geniş ve değişkendir. Bu durum, müfredat türünün öğrenci başarısını etkileyebileceğini göstermektedir.



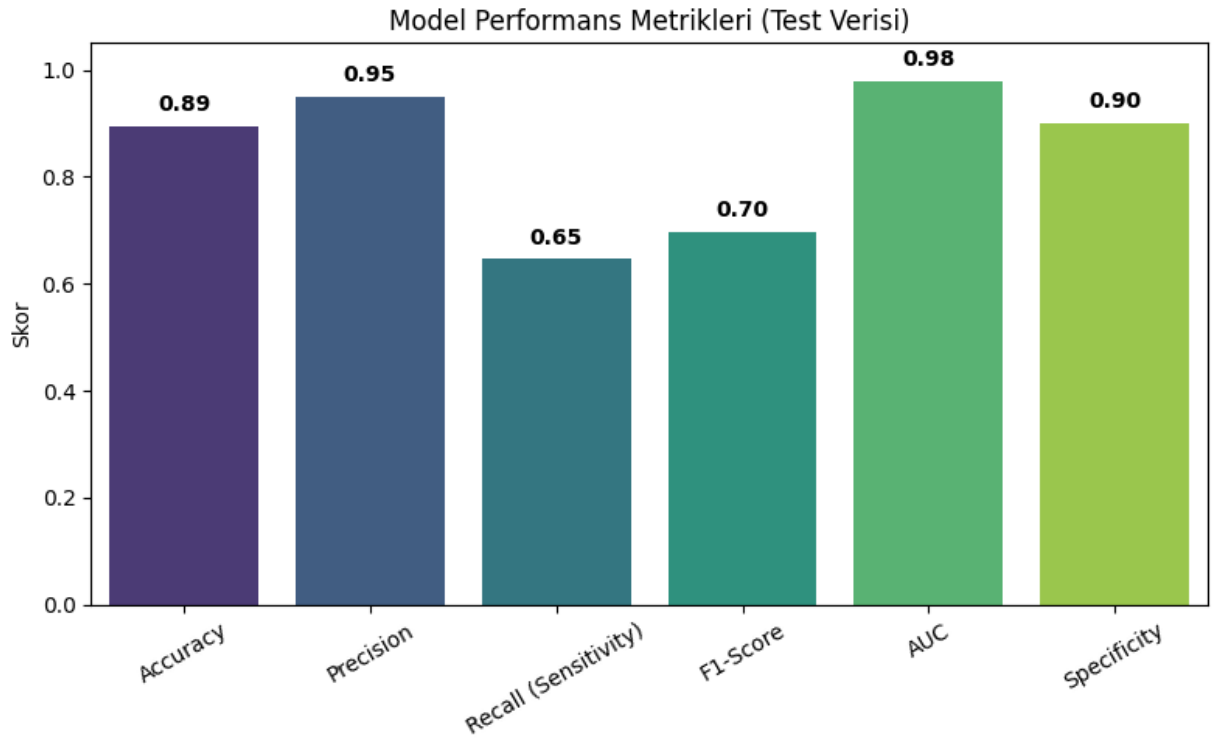
Şekil 2 ROC Eğrisi (Train)



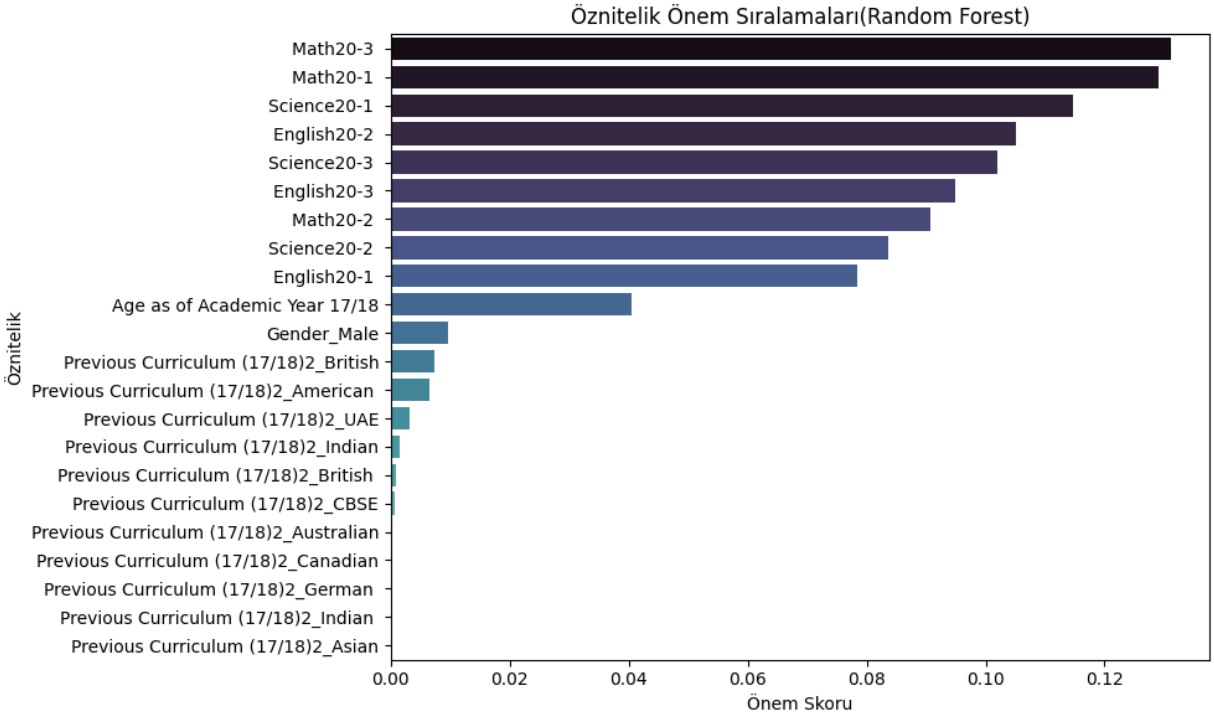
Şekil 3 ROC Eğrisi (Test)



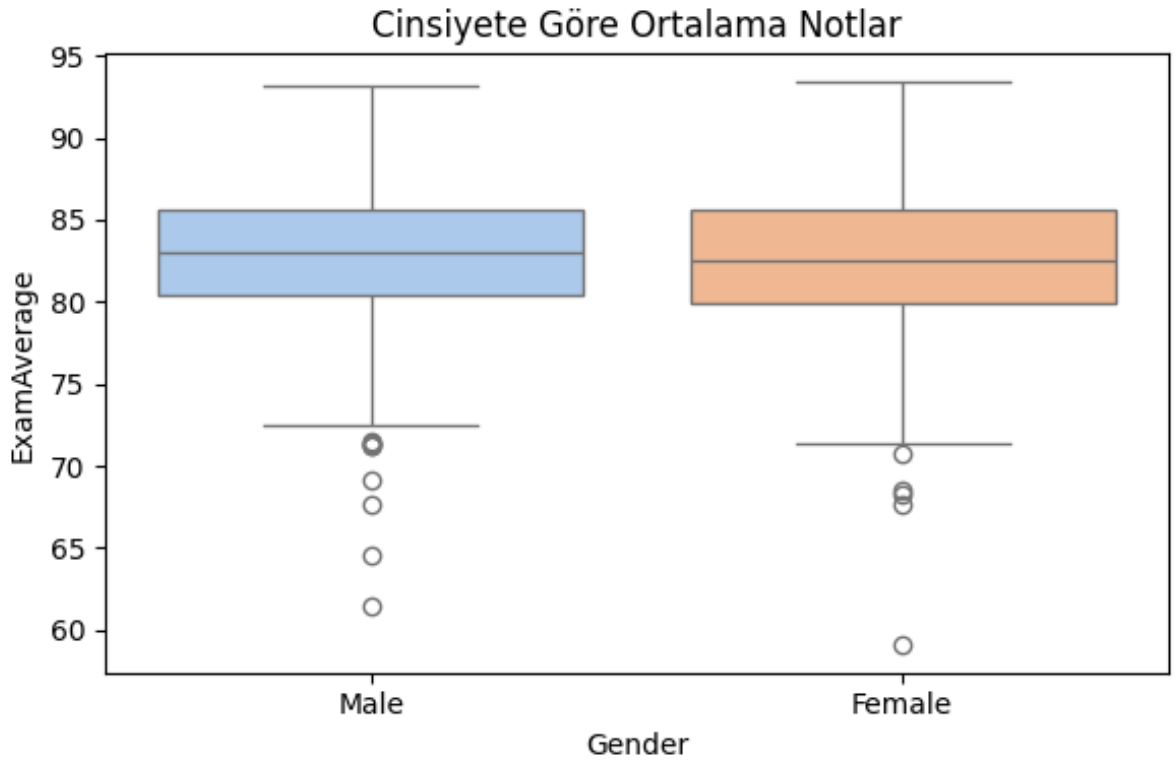
Şekil 4 Karmaşıklık Matrisi



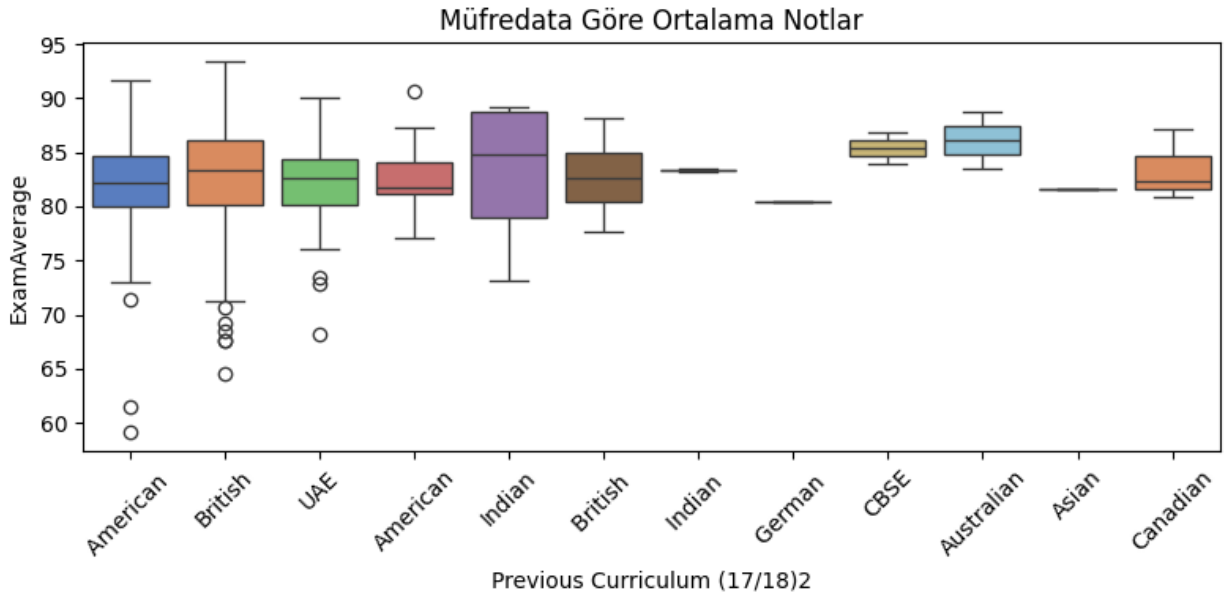
Şekil 5 Performans Metrikleri



Şekil 6 Önem Sıralamaları



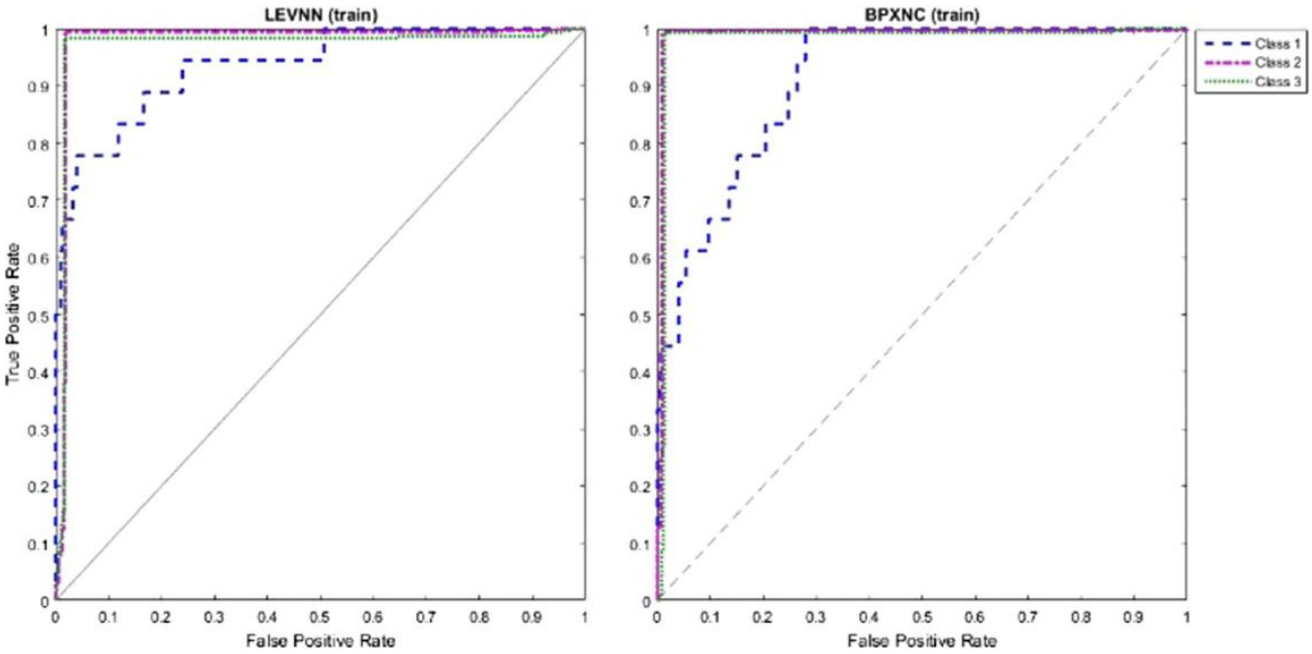
Şekil 7 Erkek - Kadın Sınav Ortalaması Karşılaştırması



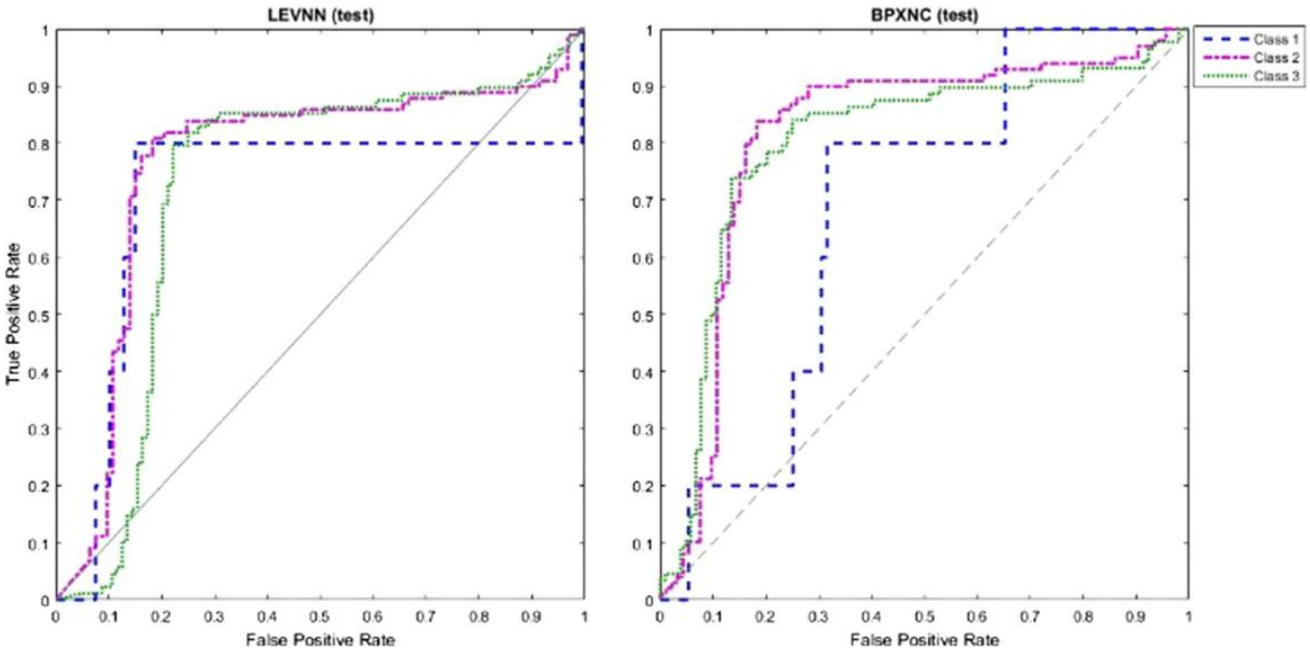
Şekil 8 Müfredatların Ortalamaya Etkisinin Karşılaştırması

Referans Aldığım Makaledeki Modelin Sonuçları

Referans aldığım makaleye baktığımızda öncelikle onun modellerinde en iyi sonucu LEVNN modelinin verdiğini gördüm. Bundan dolayı raporuma o modelin sonuçlarını ekleme kararı aldım. Ayrıca bu sonuçlara bakıldığında makalenin modelinin train verileri için başarısının, test verileri için başarısından çok daha yüksek olduğunu görebiliriz. Bu da makaledeki modelin biraz overfitting durumunu yaşıyor olabildiğini göstermektedir.



Şekil 9 Makale ROC Eğrisi (Train)



Şekil 10 Makale ROC Eğrisi (Test)

Makale ve Kendi Modelimin Karşılaştırılması

Makalede alınan değerleri ve kendi modelimin değerleri incelendiğinde;

Model Adı	Sensitivity (Recall)	Specificity	Precision	F1-Skoru	Accuracy	AUC
RFC50 (Makale)	0.632	0.674	0.466	0.473	0.660	0.720
LEVNN (Makale)	0.801	0.815	0.568	0.602	0.816	0.732
RFC200 (Benim Modelim)	0.646	0.899	0.949	0.697	0,894	0.979

Bu tablo ele alındığında;

- Sensivity'de makalenin LEVNN modeli en iyi değeri veriyor. Pozitif sınıfları en iyi şekilde yakalayan model bu modeldir.
- Specificity'de benim RFC200 modelim en iyi değeri veriyor. Negatif sınıfları en doğru şekilde tanımlayan model bu modeldir.
- Precision'da benim RFC200 modelim en iyi değeri veriyor. Doğru pozitif tahminlerde en yüksek başarıyı tanımlayan model bu modeldir.
- F-1 Skoru'nda benim RFC200 modelim en iyi değeri veriyor. Kararlılık açısından (False Positive ve False Negative dengesi) en iyi değeri veren bu modeldir.
- Accuracy'de benim RFC200 modelim en iyi değeri veriyor. Genel doğruluk açısından en iyi değeri veren model bu modeldir.
- AUC'da benim modelim en iyi değeri veriyor. Sınıflandırma açısından en iyi değeri veren model bu modeldir.

Tüm değerlere baktığımızda genel olarak benim modelim daha iyi ve doğru kararlar alabilen bir model olarak gözüküyor. Bunun birkaç nedenini şöyle sıralayabiliriz;

- Benim yazdığım kodda en iyi model GridSearch ile belirleniyor, parametreler tek tek deneniyor; makaledeki modelde ise sınırlı ayarlama var, çoğunlukla sabit değerlerle çalışılmış.
- Benim modelim için veri setinden çekilen verilerdeki eksik veriler dropna() fonksiyonu ile temizlendi. Makalede bu işlemler manuel yapılmış.
- Benim modelimde modeli karmaşıklaştıran ve önem belirtmeyen bazı öznitelikler temizlendi, bu makalenin modelinde gerçekleştirilmemiş. Bu da makale için verilerde daha fazla hata ve gürültü olma olasılığını arttırıyor.

Kaynakça

- GeeksforGeeks. (2025). *Random Forest Algorithm in Machine Learning*. Erişim tarihi: 29 Mayıs 2025, https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/
- Analytics Vidhya. (2025). *Decision Tree vs Random Forest | Which Is Right for You?*. Erişim tarihi: 29 Mayıs 2025, https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/
- IBM. (t.y.). *What Is Random Forest?*. Erişim tarihi: 29 Mayıs 2025, https://www.ibm.com/think/topics/random-forest

- Medium. (t.y.). *Random Forest Explained: A Visual Guide with Code Examples*. Erişim tarihi: 29 Mayıs 2025, https://medium.com/data-science/random-forest-explained-a-visual-guide-with-code-examples-9f736a6e1b3c
- Wikipedia contributors. (t.y.). *Precision and recall*. Erişim tarihi: 29 Mayıs 2025, https://en.wikipedia.org/wiki/Precision_and_recall
- DataCamp. (2024). *AUC and the ROC Curve in Machine Learning*. Erişim tarihi: 29 Mayıs 2025, https://www.datacamp.com/tutorial/auc
- Gürcan Öğündür. (t.y.). *Doğruluk (Accuracy), Kesinlik (Precision), Duyarlılık (Recall) ya da F1 Score*. Erişim tarihi: 29 Mayıs 2025, https://medium.com/@gulcanogundur/do%C4%9Fruluk-accuracy-kesinlik-precision-duyarl%C4%B1l%C4%B1k-recall-ya-da-f1-score-300c925feb38
- Ghareeb, S., Hussain, A. J., Al-Jumeily, D., Khan, W., Al-Jumeily, R., Baker, T., Al Shammaa, A., & Khalaf, M. (2022). *Evaluating student levelling based on machine learning model's performance*. Discover Internet of Things, 2(3). Erişim tarihi: 29 Mayıs 2025, https://link.springer.com/article/10.1007/s43926-022-00023-0