Logo

Description automatically generated

**MARMARA ÜNİVERSİTESİ  
İŞLETME FAKÜLTESİ**

**YÖNETİM BİLİŞİM SİSTEMLERİ BÖLÜMÜ**

**ARA SINAV**

**YBS304-VERİ MADENCİLİĞİ**

**KONU BAŞLIĞI**

**HAZIRLAYAN**

**İSİM SOYİSİM**

**NUMARA**

**ÖĞRETİM ÜYESİ**

**NAME VORNAME**

**2020**

**İÇİNDEKİLER**

**ŞEKİLLER**

**EKLER**

1. **GİRİŞ**

Günümüzde veri işleme yöntemleri çok gelişmiştir. Bunun sebebi bu alana yapılan sürekli veri analizi ile ilgili araştırmaların olmasıdır. Bu çalışma da 2 farklı veri analizi metodunun ‘Student‘ veri seti ile eğitip birbirleriyle karşılaştırmasını konu alacaktır.

Student veri seti; 2 farklı okulun öğrencileri hakkında yaşı, velisi, notları gibi bir çok bilgiyi barındıran bir veri setidir. Bu veri seti, bir açık veri sitesinden elde edilmiştir.

1. **ÇALIŞMANIN AMACI**

Bu çalışmanın amacı kısaca şu şekilde bahsedilebilir: Student veri setini kullanarak bir öğrencinin notlarını methodlar tarafından tahmin edilmesini sağlamak. Bunun sonucunda kullanılan iki farklı metodun (logistik regresyon ve diskriminan analiz) karşılaştırmasını yaparakhangi modelin student veri seti için daha uygun olduğunu saptamak.

1. **PROBLEMİN TANIMLANMASI**

Öğrencilerin gelecekteki notlarının bilinmesi, öğrencilerin eksikliklerinin saptanıp önceden giderilerek daha başarılı birer birey haline getirilmesini sağlayabilir. Bundan dolayı ilk olarak yapılması gereken bir öğrencinin gelecekte alacağı notlarını tahmin etmek olmalıdır ve bu çalışmanın amacı da budur.

1. **MODELLERİN OLUŞTURULMASI**

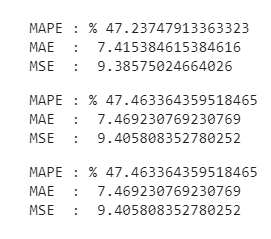
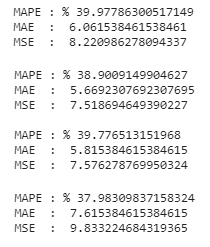
Modeller oluşturulurken ilk olarak veri setini metodlar için hazırlamaydı. Bunun için de veri seti içindeki 3 farklı kolona dağılmış olan notların toplanarak tek bir kolon haline getirildi. Sonrasında verinin numerik olmayan kolonları şifrelenerek numerik hale getirildi ve tüm kolonlar 0 – 1 değerleri arasına normalize edildi. Normalize edilmesindeki sebep bir kolonun öteki bir kolona baskın olmasını engellemekti.

Methodlar çalışmanın ilk bölümlerinde oluşturulmuş olan datanın 80% ini kendini eğitmek, %20 sini ise test etmek için kullanmıştır. Ayrıca modeller kendini eğitirken birçok farklı parametrelerde tekrar tekrar çalıştırılıp en uygun olan parametreler belirlenmeye çalışılmıştır. Modeller sırasıyla şu parametrelerde çalışmıştır:

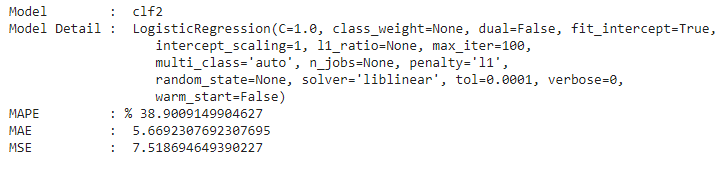
* **Logistik Regresyon:**
* penalty = l2, solver = default
* penalty = l1, solver = liblinear
* penalty = elasticnet, solver = saga
* penalty = none, solver = newton-cg
* **Diskriminant Analizi:**
* solver = svd
* solver = lsqr
* solver = eigen

1. **SONUÇLARIN KARŞILAŞTIRILMASI**

Logistik Regresyon Diskriminant Analizi



Yukarıdaki tabloda da görüldüğü gibi her model test datasında test edilip sonucun MAPE(Mean Absolute Percentage Error), MAE(Mean Absolute Error) ve MSE(Mean Squared Error) değerlerine bakılmıştır. Sonuçlar karşılaştırılınca en iyi model Aşağıdaki tabloda da görüldüğü üzere logistik regresyon metodu (penalty = l1, solver = liblinear) modeli olmuştur.



1. **DEĞERLENDİRME**

Yapılan çalışmanın sonuçları göz önünde bulundurulursa, her bir logistik regresyon modeli neredeyse her error tipi için de diskriminant analizine göre çok daha iyi sonuçlar elde etmiştir. Burdan ise student veri setinin logistik regresyon modeli ile diskriminant analiz metoduna göre çok daha iyi çalıştığı, neredeyse her solver ve penalty ile uyumlu çalıştığı (değerlerinin yakın geldiği), kısaca logistik regresyon modellerinin diskriminant analizi modellerine göre student datası için daha uygun olduğu sonucuna ulaşılabilir.