Kredi Onayı İçin Bir Sınıflandırma Algoritması Önerisi
A Classification Algorithm Advice for Credit Approval

İsmail Haberal, Bilgisayar Mühendisliği Bölümü, Başkent Üniversitesi
ihaberal@baskent.edu.tr

Umut Tosun, Bilgisayar Mühendisliği Bölümü, Başkent Üniversitesi
utosun@baskent.edu.tr

Özet
Otomatik kredi onayı, finansal uygulamalarda önemli bir yere sahiptir. Kredi, finans kurumları tarafından müşteriye faiziyle birlikte geriye ödenmesi için verilen borç paradır. Bu borçla ilgili risk, müşterinin borcunu geriye ödeyememe olasılığını öngörebilmek, finans kurumu için önemlidir. Bunun için karar verme aşamasında kullanılacak yöntem iyi seçilmelidir. Bu çalışmada finansal kredi onaylarının sınıflandırma teknikleri ile kontrolü incelenmiştir. Veri kümesi olarak Avustralya Kredi Onay Veri Kümesi kullanılmıştır. Dört farklı sınıflandırma tekniği (LDA, kNN, Bayes, SVM) bu veri kümesine ayrı ayrı uygulanarak, elde edilen sonuçlardan hangi tekniğin daha iyi olduğu karşılaştırılmıştır.

Anahtar Kelimeler
Kredi Onayı, Veri Madenciliği, Veri Analizi, Sınıflandırma Algoritmaları

Abstract
Automatic credit approval, has an important place in financial applications. Credit to customers by financial institutions lending money with interest is to be paid back. This debt-related risk, the customer's inability to pay back the debt is important for financial institutions to be able to predict the probability. For this reason, decision-making stage must be well chosen. In this study, the classification of financial loan approval and control techniques have been investigated. Australian Credit Approval Dataset was used. Four different classification techniques (LDA, kNN, Bayesian, SVM) were applied to the data set respectively and the results obtained were compared.

Keywords
Credit Approval, Data Mining, Data Analysis, Classification Algorithms

I. GİRİŞ

Bankacılık ve finansal işlemler ile ilgili gelen başvuru taleplerine kredi onayı vermek önemli bir işlemdir. Bu başvuruları tek tek değerlendirip sonuçlandırmak oldukça fazla zaman gerektirmektedir. Her başvuru belirlenen kriterlere göre tek tek incelenip bu kriterlere uyup uymadığına göre değerlendirilmelidir. Bu oldukça yorucu ve zaman alan bir durumdur. Sonucun belirlenmesi ve müşteriye geri bildirimi uzun bir süreci gerektirecektir. Değerlendirme esnasında, gözden kaçan durumların da olabilmesi muhtemeldir. Müşterinin veya kurumun mağdur olması ihtimalleri vardır. Olumsuz durumları önlemek için, bazı kontroller içeren bir yapı oluşturmak gerekir. Başvuru talebinde bulunan kişinin geçmişe yönelik veya anlık bilgilerinin biliniyor olması bu işlemin sonuçlanmasında önemli bir adımdır. Başvuran kişinin gelir-gider durumu, daha önce kredi kullanıp kullanmadığı gibi kriterler karar vermeyi kolaylaştıracak bilgilerdir. Karar verme aşamasında bu bilgilerden yola çıkan bir sisteme ihtiyaç duyulmaktadır. Gelen yeni bir başvurunun olumlu ya da olumsuz sonuçlanmasını belirlenen kriterler ölçüsünde saptayan otomatik bir karar sistemi, en az hatayla bu görevi üstlenecektir [2]. Bu çalışmada bir kredi başvuru kümesi sınıflandırılmıştır. Bu sınıflandırmaya dayanarak gelen bir başvurunun uygun olup olmadığı tespit edilmeye çalışılmıştır. Sınıflandırma yapmak için Fisher Linear Discriminant Analysis (FLDA), k-En Yakın Komşu (kNN), Naive Bayes, Destek Vektör Makinesi (SVM) algoritmaları kullanılmıştır. Bu algoritmalar ile Avustralya Kredi Onayı veri kümesi analiz edilmiş ve en iyi sonucu hangi algoritmanın verdiği tespit edilmeye çalışılmıştır.

II. MATERYAL VE METODLAR

A. Veri Kümesi

Avustralya Kredi Onayı veri kümesi, UCI Makine Öğrenmesi Ambarından temin edilmiştir. Veri kümesi, ilk kez 1987'de Ross Quinlan tarafından, kredi kartı uygulamalarında çalışmak üzere sağlanmıştır.[1] Veri kümesi, 6 numerik, 8 kategorik olmak üzere toplam 14 özellik ve 1 tane de

sınıf değeri olmak üzere toplam 15 özelliğe sahiptir. Veri kümesi Tablo III' te özet olarak gösterilmiştir. Veri içerisindeki isim ve diğer değerler, gerçek değerlerini korumak için sembolik olarak değiştirilmiştir. Veri kümesi, uzun ve kısa değerlerden oluşan iyi bir karışıma sahiptir ve birkaç kayıp değer vardır. Veri kümesi, 307 (%44.5)'si pozitif (kredi onaylanmış), 383 (%55.5)'ü negatif (kredi red) veri olmak üzere toplam 690 kayıttan oluşmaktadır. Grafik olarak veri kümesinin sınıf dağılımı Şekil 1'de gösterilmiştir. 37 (%5) örnek bazı kayıp(eksik) değerler içermektedir. Sınıf dağılımı Tablo I' de, eksik değer istatistiği ise Tablo II' de verilmiştir. Eksik değerler, Statlog projesinde düzenlenerek kategorik ve sürekli olarak veri kümesi içerisine yeniden eklenmiştir. [3]

B. Sınıflandırıcılar

Sınıflandırıcılar, otomatik karar verme sistemi için çok önemli bir yere sahiptir [4], [5]. Gelen verinin hangi sınıfa ait olduğunu belirlememizi sağlar. Sınıflandırma işlemini yapan bir çok algoritma vardır. Bunlar, farklı veri kümeleri için farklı sonuçlar verebilir. Bunun için, elimizdeki veriye en uygun sınıflandırıcıyı kullanmamız, sağlıklı karar vermek için önemlidir.
Sınıflandırıcılar veri kümesine iki şekilde uygulanabilir; birincisi tüm veriye uygulama, ikincisi ise veri boyutu azaltılmış (future reduction) veriye uygulama. Boyut azaltma algoritması olarak Temel Bileşen Analizi (Principal Component Analysis-PCA) kullanılmıştır.

Temel Bileşen Analizi (PCA)

Verilen d boyutlu uzaydaki bir girdinin, izdüşüm yöntemiyle k<d boyutlu yeni bir uzaya, bilgi kaybı en az olacak şekilde eşlenmesidir [6]. Burada sınıf bilgisine ihtiyaç yoktur. Temel kriter, verideki değişimdir. Değişimin en yüksek olduğu yani veri noktaları arasındaki farkın en iyi ortaya çıktığı nokta, boyut azaltma noktamızdır (principal component). Bunu hesaplamak için, veri kümesine şu adımlar sırasıyla uygulanır. Önce her özelliğin ortalaması hesaplanır, her bir örneğin bu ortalamaya uzaklığı hesaplanarak elde edilen matrisin kovaryansı alınır. Kovaryans matrisinden de özdeğerler elde edilir. En iyi K değeri seçimi için, her K adet özdeğer toplamı denklem 1'de görüldüğü gibi, toplam özdeğere oranlanır (X_K).

$$X_K = \frac{\sum_{i=1}^{K} \lambda_i}{\sum_{i=1}^{N} \lambda_i} \qquad (1)$$

Özellik sayısı x eksenini, bu özellik karşılığında bulunan (XK) değeri de y eksenini temsil etmek üzere grafik çizilir.
Grafikte eğimin belirgin olarak değişime uğradığı(veri kaybı) yani farkın en belirgin olduğu noktadaki nitelik değeri (K), veri kümesi için en iyi K değeri olarak seçilir.

Tablo I: Avustralya Kredi Onayı Veri Kümesi İçin Sınıf Dağılımı

Sınıf	Dağılım
+	307 (%44.5)
-	383 (%55.5)

Tablo II: Avustralya Kredi Onayı Veri Kümesi içindeki Kayıp Değerler

Özellik	Eksik Değer
A1	12
A2	12
A4	6
A5	6
A6	9
A7	9
A14	13

Tablo III : Avustralya Kredi Onayı Veri Kümesi

Özellik	Açıklama	Tip
A1	Sex	girdi
A2	Age	girdi
A3	Mean time at addresses	girdi
A4	Home status	girdi
A5	Current occupation	girdi
A6	Current job status	girdi
A7	Mean time with employers	girdi
A8	Other investments	girdi
A9	Bank account	girdi
A10	Time with bank	girdi
A11	Liability reference	girdi
A12	Account reference	girdi
A13	Monthly housing expense	girdi
A14	Savings account balance	girdi
A15	Class(Reject/ Accept)	girdi

Fisher Doğrusal Diskriminant Analizi (FLDA)

Sınıflandırma için gözetimli bir boyut azaltma yöntemidir. W katsayısıyla tanımlanan öyle bir yön bulmak istiyoruz ki, w üzerine izdüşümleri alındığında iki sınıfın örnekleri birbirinden olabildiğince ayrılsın. İz düşümden sonra iki sınıfın iyi ayrışmış olması için ortalamalarının olabildiğince uzak olmasını ve bir sınıfın örneklerinin olabildiğince küçük bir alanda toplanmasını isteriz.

$$w=S_w^{-1}(m_1-m_2) \quad (2)$$

$$m_i=\frac{1}{K}\sum_{i=1}^{K} X_i \quad (3)$$

$$S_i=\sum(x-mi)(x-mi)^T \quad (4)$$

$$s_w=\sum_{i=1}^{K} S_i \quad (5)$$

Elde edilen yeni w değeri bütün örneklere uygulanarak örneklerin izdüşümü elde edilir:

$$y_i=w^T.x_i \quad (6)$$

İki sınıflı FLDA analizinde optimum w* değeri denklem 2 ile bulunabilir. Denklem 3 ile de her sınıfın ortalama vektörü bulunabilir. Burada x_i sınıfa ait her bir elemanı, K da eleman sayısını ifade eder. Denklem 4 ile çok değişkenli özellik uzayı x'te saçılma matrisleri tanımlanır. Denklem 5'te saçılma matrislerinin toplamı ile sınıf-içi saçılma matrisleri bulunur. Örneğin iki sınıflı FLDA analizinde $Sw = S_1 + S_2$ ile sınıf-içi saçılma matrisi elde edilir. Sonuç olarak örnek x değerlerinin, denklem 6'da olduğu gibi, doğrusal bir çizgiye izdüşümüyle skaler bir y değeri bulunur.

k-En Yakın Komşu (kNN)

Sınıflandırması bilinmeyen bir örneği, kendisine en yakın olan komşu sınıfa dahil etme yöntemidir. Bunun için iki nokta arasındaki uzaklığı hesaplayan metotlar kullanılır. Bu çalışmada denklem 7'de tanımlanan Öklit Uzaklığı metodu kullanılmıştır.

$$d(x,y)=\sqrt{\sum_{i=1}^{n}|xi-yi|^2} \quad (7)$$

Naive Bayes Sınıflandırıcılar

Sınıflandırması bilinmeyen bir örneğin, var olan sınıflardan hangisine ait olduğunu hesaplama metodudur. P(C1|X), X örneğinin C1 sınıfı içerisinde olma olasılığını gösterir [6].

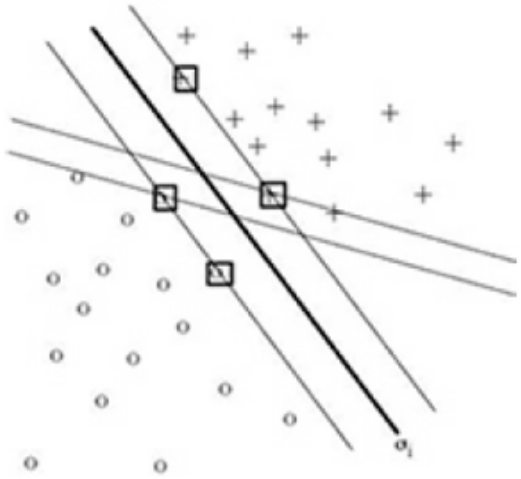
$$\frac{P(C1|X)}{P(C2|X)}>1 \quad (8)$$

$$\frac{P(C_1|X)}{P(C_2|X)}=\frac{P(X|C_1).P(C_1)}{P(X|C_2).P(C_2)} \quad (9)$$

Hangi sınıf değeri yüksek ise, X o sınıfın elemanı olarak kabul edilir. Örneğin, iki sınıflı bir örnek için Naive Bayes metoduna göre bulunan olasılıklar denklem 8 ve denklem 9'a göre yazılarak elemanın hangi sınıfa ait olduğu tespit edilir.

Destek Vektör Makinesi (SVM)

Sınıflandırma için kullanılan en önemli yöntemlerden biridir. Sınıflandırma için bir düzlemde bulunan iki grup arasında bir sınır çizilerek iki grubu ayırmak mümkündür (Şekil 1). Bu sınırın çizileceği yer ise iki grubun da üyelerine en uzak olan yer olmalıdır. İşte SVM bu sınırın nasıl çizileceğini belirler. Bu işlemin yapılması için iki gruba da yakın ve birbirine paralel iki sınır çizgisi çizilir ve bu sınır çizgileri birbirine yaklaştırılarak ortak sınır çizgisi üretilir [7].



Şekil 1: SVM düzlemi

SVM, hem doğrusal olarak ayırt edilebilen hem de edilemeyen veri kümesini sınıflandırabilir. Doğrusal olmayan bir eşlem ile n boyutlu veri kümesi m > n olacak şekilde m boyutlu yeni bir veri kümesine dönüştürülür. Yüksek boyutta doğrusal sınıflandırma işlemi yapılır. Uygun bir dönüşüm ile her zaman veri bir hiper düzlem ile iki sınıfa ayrılabilir. Hiper düzleme en yakın öğrenme verileri destek vektörleri olarak adlandırılır

C. Uygulama Yöntemi

Uygulama için, veri kümesi 2-kat yapılıp, öğrenme verisi %50 ve test verisi %50 olarak tercih edilmiştir. Bu işlem PCA uygulanmış veriye

de uygulanır. Her bir sınıflandırıcı için bu veriler (PCA uygulanmış veri ve PCA uygulanmamış orijinal veri) kullanılmıştır. Sınıflandırıcıların doğruluk, özgüllük, duyarlılık ve hassasiyet değerleri hesaplanmıştır. Veri kümesi için gerçek doğru-yanlış değerleri ile, sistem tahmininin pozitif-negatif sonuçlarından oluşan Karışıklık Matrisi'i Tablo IV'deki gibi gösterilmektedir.

Doğruluk: Populasyon içerisindeki true sonuçların (doğru positif ve doğru negatif) oranını ifade eder. Denklem 10'da gösterilmiştir.

$$\frac{TP+TN}{TP+TN+FP+FN} \qquad (10)$$

Özgüllük: Sistemin negatif tahmininin gerçek negatiflere oranını ifade eder. Denklem 11'de gösterilmiştir.

$$\frac{TN}{TN+FP} \qquad (11)$$

Duyarlılık: Sistemin pozitif tahmininin gerçek pozitiflere oranını ifade eder. Denklem 12'de gösterilmiştir.

$$\frac{TP}{TP+FN} \qquad (12)$$

Hassasiyet: Doğru pozitif sonuçların bütün pozitiflere oranını ifade eder. Denklem 13'te gösterilmiştir.

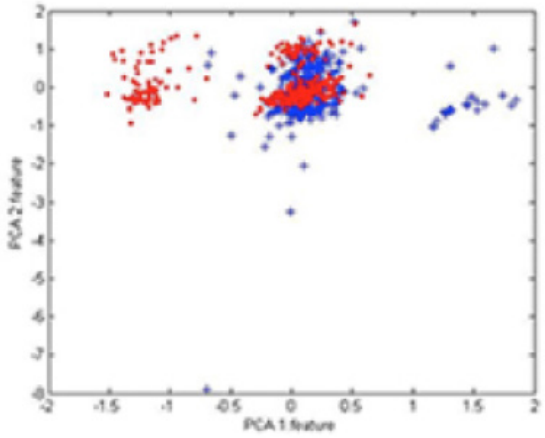
$$\frac{TP}{TP+FP} \qquad (13)$$

Tablo IV: Karışıklık Matrisi (TP: Doğru Pozitif FP: Yanlış Pozitif FN: Yanlış Negatif TN: Doğru Negatif)

		Gerçek		
		Doğru	Yanlış	
Sistem	Doğru	TP	FP	*Hassasiyet*
	Negatif	FN	TN	
		Duyarlılık	*Özgüllük*	*Doğruluk*

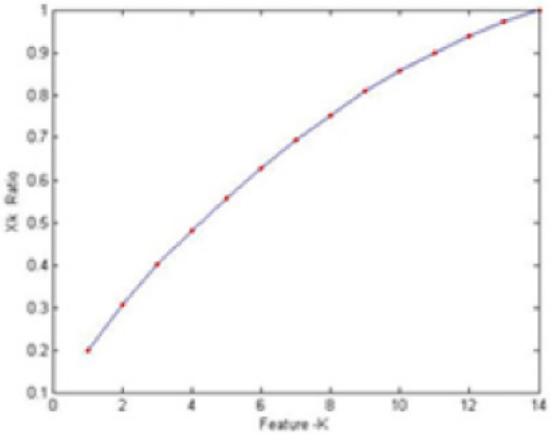
III. BULGULAR

Veri kümesine PCA uygulandıktan sonra ortaya çıkan yeni özellik değerlerine göre sınıf dağılımı Şekil 2'de gösterilmiştir.
Burada mavi renkli noktalar, pozitif sınıf değerli özellikleri, kırmızı renkli noktalar ise negatif sınıf değerli özellikleri ifade etmektedir. x ekseni 1.özellik, y ekseni ise 2.özellik değerini temsil etmektedir. Bu grafikte PCA uygulanmış verinin ilk iki özellik değeri çizilmiştir.



Şekil 2: Sınıf dağılım grafiği

Veri kümesine PCA uyguladıktan sonra en iyi K değeri olarak Şekil 3'deki grafikten 9 değeri tespit edilmiştir. PCA uygulanmış veri kümesinin, sınıflandırıcı hesabı yapılırken, bu veri kümesinin ilk 9 özellik işleme sokulmuştur.



Şekil 3: K değeri

Orijinal veri kümesi ve PCA uygulanmış veri kümesi için hesaplama yapılan sınıflandırıcılar ve elde edilen değerler Tablo V'de verilmiştir. Burada PCA(-) değeri, orijinal veri kümesi, PCA(+) değeri ise PCA uygulanmış veri kümesi (K=9) göstermektedir.
Veri kümesi üzerinde sınıflandırıcı hesabı yapılarak bir karışıklık matrisi oluşturulmuştur ve bu matristen duyarlılık, özgüllük, hassasiyet ve doğruluk değerleri hesaplanmıştır.

Yapılan analizler de CA uygulanmış veri ile PCA uygulanmamış veriden elde edilen sonuçlar farklılıklar göstermektedir.

Bazı algoritma için sonuç yüksek çıkarken, bazıları için düşük çıkmıştır. Yukardaki tablodan da anlaşılacağı gibi, bu veri kümesi için, Fisher Doğrusal Diskriminant (FLDA) sonucu doğru-

luk %46 , hassasiyet %46, özgüllük %0 değeri, diğer algoritmalara göre daha düşük çıkmıştır. Hassasiyet değeri %100 çıkmıştır. PCA uygulanmış data için de FLDA'nın bu değerleri en düşük değer olarak sonuçlanmıştır. Bu değeler doğruluk %51 , duyarlılık %43, özgüllük %58, ve hassasiyet %46 'dır.

k-En Yakın Komşu (kNN) algoritması sonucu, PCA uygulanmış data için, doğruluk %94 ve duyarlılık %99 değerleri diğer algoritmalara göre daha yüksek çıkmıştır. Duyarlılık %88, hassasiyet değeri de %99 hesaplanmıştır. Bu algoritma, PCA uygulanmamış data için ise elde edilen değerler bazı algoritmalara göre daha düşük çıkmıştır. Bu değerler doğruluk %57, duyarlılık %76, özgüllük %83, ve hassasiyet %83 olarak hesaplanmıştır.

Naive Bayes algoritması, PCA uygulanmamış data için elde edilen duyarlılık değeri %90, diğer algoritmalara göre daha yüksek çıkmıştır. Doğruluk değeri %79, özgüllük değeri %75, hassasiyet değeri %61 olarak elde edilmiştir. PCA uygulandıktan sonra Naive Bayes algoritması uygulanarak, doğruluk %73, duyarlılık %82, özgüllük %70, hassasiyet %52 sonuçları elde edilmiştir.
Karar Destek Makinesi (SVM) algoritması ile PCA uygulanmamış verilere uygulandığında özgüllük değeri %94, diğer algoritmalara göre daha yüksek sonuçlanmıştır. Doğruluk değeri %85, kNN algoritmasından sonra gelen en yüksek değer olarak elde edilmiştir. Diğer değerler ise, duyarlılık %77, hassasiyet %94 olarak sonuçlanmıştır. PCA uygulandıktan sonra SVM algoritmasında, doğruluk %77, duyarlılık %85, özgüllük %73, hassasiyet %59 değerleri elde edilmiştir.

Tablo V: Bulgular

	FLDA		kNN		Bayes		SVM	
	PCA(-)	PCA(+)	PCA(-)	PCA(+)	PCA(-)	PCA(+)	PCA(-)	PCA(+)
Duyarlılık(%)	46	43	49	88	90	82	77	85
Özgüllük(%)	0	58	76	99	75	70	94	73
Hassasiyet(%)	100	46	83	99	61	52	94	59
Doğruluk(%)	46	51	57	94	79	73	85	77

sIV. SONUÇ VE TARTIŞMA

PCA algoritması bu veri kümesi için net bir boyut azaltma yapamadığını göstermiştir. Dolayısıyla, PCA ile boyut azaltma bu veri kümesine uygulanmayabilir. Otomatik kredi skorlama ve onaylama, hızlı ve akıllı karar verme açısından finans kurumları için günümüzde çok önemli bir role sahiptir. Kredi başvurularının otomatik olarak sonuçlandırılması için, önceden bir sınıflandırma yapılması ve gelen başvuru da bu sınıflandırmalardan hangisine uygun ise ona göre sonuçlandırılması gerekmektedir.
FLDA algoritması, Avustralya Kredi Onayı verisinin sınıflandırması için en düşük değerlere sahip olduğu için, bu verinin sınıflandırmasında başarılı olamamaktadır. Veri kümesinin grafiksel dağılımında, sınıfların birbiri içerisine girip dağılması, lineer bir çizgiyle ayrılmasını olanaksızlaştırabilir. Buna bağlı olarak da sınıflandırma değerleri düşük çıkmış olabilir. kNN algoritması, PCA uygulandıktan sonra doğruluk ve özgüllük değerleri daha yüksek çıkmıştır. Bu veri kümesi sınıflandırılması için en iyi algoritmanın PCA uygulanmış kNN algoritması olduğunu söyleyebiliriz. SVM algoritması da bu veri kümesi için, kNN algoritmasından sonra iyi sonuçları veren algoritma olduğu görülmektedir. Veri kümesinin birbirinden net bir şekilde ayrılmamış olması, birbirine karışmış olması, SVM algoritmasının bu veriyi ayırmada en etkili metotlardan biri olmasını sağlamış olabilir.

KAYNAKLAR

[1] Ross Quinlan. Simplifying decision trees. Int J Man-Machine Studies 27,pp. 221-234, Dec 1987

[2] Sum Sakprasat, Mark C. Sinclair. Classification Rule Mining for Automatic Credit Approval using Genetic Programming. IEEE Congress on Evolutionary Computation, Page(s):548-555, 2007

[3] "Statlog Datasets: comparison of results". Department of Informatics, Nicolaus Copernicus University, Available: http://www.is.umk.pl/projects/datasetsstat.
html

[4] Ethem Alpaydın. Combined 5 x 2 F test for comparing supervised classification learning algorithms. Neural Computation, 11:1885–1892, 1999

[5] Janez Demsar. Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research 7, 1-30, 2007

[6] Ethem Alpaydın.Yapay Öğrenme, Boğ.Ünv Yayınları

[7] Cheng-Lung Huang, Mu-Chen Chen, Chieh-Jen Wang. Credit scoring with a data mining approach based on support vector machines. Expert Systems with Applications 33, 847–856, 2007