Sınıflandırılmış Kısıtlı Boltzmann Makinesi ile Kredi Risk Analizi

Credit Risk Analysis with Classification Restricted Boltzmann Machine

Mustafa Bayraktar[1], Mehmet S. Aktaş[1], Oya Kalıpsız[1]
Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi, İstanbul
mustafa.bayraktar@std.yildiz.edu.tr
mehmet@ce.yildiz.edu.tr
oya@ce.yildiz.edu.tr

Orkun Susuz[2], Selçuk Bayracı[2]
Ar-Ge Merkezi
Cybersoft, İstanbul
selcuk.bayraci@cybersoft.com.tr
orkun.susuz@cybersoft.com.tr

***Özetçe*—Bankacılık sektörünün büyüklüğü ve önemi itibari ile taşıdığı riskler, denetleyici kurumlar tarafından yakından izlenmektedir. Bankacılık sektörünü etkileyen en önemli risk olarak kredi risk kavramı öne çıkmaktadır. Bankalar kredi verdikleri müşterileri iyi tanımaya ve verdikleri kredilerin geri ödenebilmesi konusunda sorun yaşamamaya çalışırlar. Fakat günümüzde bankalardan kredi kullanan kişi sayısının milyonlara ulaşması sonucunda güvene ve müşteriye dayalı kredi verme süreci oldukça güçtür. Bu nedenle, kredi skorlama teknikleri olarak adlandırılan ve müşteri bilgilerini kullanarak müşterilere kredi verilip, verilmeyeceğini ölçen teknikler günümüzde bankalar arasında yaygın olarak kullanılmaktadır. Bu çalışmada kredi skorlama sistemlerinde kullanılabilecek algoritmalar incelenmiştir. Çalışma kapsamında, kredi talebinde bulunan müşterilerin kredi isteklerinin onaylanması veya geri çevrilmesini kolaylaştıracak bir prototip uygulama geliştirilmiştir. Geliştirilen uygulama Sınıflandırılmış Kısıtlı Boltzmann Makinesi ve Çok Katmanlı Yapay Sinir Ağları gibi derin öğrenme yöntemleri ile yaygın olarak kullanılan makine öğrenimi algoritmalarının karşılaştırılmasını sunmaktadır. Prototip uygulamanın performans testleri yapılmış ve sistemin kullanılabilirliğini gösteren olumlu sonuçlar elde edilmiştir.**

Anahtar kelimeler — Kredi Risk Analizi, Sınıflandırılmış Kısıtlı Boltzmann Makinesi, Yapay Sinir Ağları

***Abstract*—Banks are closely monitored by supervisory institutions because of the size and importance of the banking sector. One of the most important risks affecting the banking sector is the concept of credit risk. The banks try to make sure that the customers are well-financed and they can repay the given loans. But nowadays, as the number of people who use credit from banks reach to millions, trusting and customer-based lending process is very difficult. For this reason, techniques which are called credit scoring techniques and which measure whether given credit or not to customers using customer information are widely used among banks today. In this study, algorithms that can be used in credit scoring systems are examined. Within the scope of the study, a prototype implementation was developed that would facilitate the approval or denial of credit requests by customers. The developed application presents a comparison of commonly used machine learning methods with deep learning methods such as Classification Restricted Boltzmann Machine and Multilayer Artificial Neural Networks. Performance tests of the prototype application have been performed and positive results have been obtained showing the availability of the system.**

Keywords — Credit Risk Analysis, Classification Restricted Boltzmann Machine, Artificial Neural Networks

I. GİRİŞ

Kredi riskinin doğru değerlendirilmesi kredi veren kuruluşlar için son derece önem taşımaktadır. Kredi skorlama, kredi başvurusunda bulunan müşteriye kredi vermek veya vermemek yönündeki karar için finansal kurumlar tarafından gerçekleştirilen analizlere yardımcı olarak yaygın kullanıma sahip bir tekniktir[1]. Başvuru sahibinin kredi skoru veya değerliliği üzerindeki kesinlik kazanan karar olası kayıpları en aza indirerek finansal kurumların kredi verme hacmini arttırmaya yardımcı olmaktadır.

Kredi risk analizinde kredinin temerrüde düşmesini etkileyebilecek çok sayıda faktör bulunmaktadır. Bu faktörlerin içinden bir alt kümenin belirlenmesi ve modelin seçilen bu faktörler üzerinde koşturulması öğrenme algoritmalarının başarısını olumsuz yönde etkileyebilmektedir. Derin öğrenme tekniklerinin kredi skorlama amacıyla sıklıkla kullanılmasının başlıca sebebi model oluşturma safhasında geleneksel makine öğrenimi yöntemlerin aksine faktör seçme işleminin olmamasıdır.

Bu çalışma kapsamında, kredi risk analizi için derin öğrenme yöntemlerinden; Sınıflandırılmış Kısıtlı Boltzmann Makinesi (ClassRBM)[2] ve Çok Katmanlı Yapay Sinir Ağları (YSA)[3] kullanılmıştır. Buna ek olarak makine öğrenmesi algoritmalarından; K En Yakın Komşu (KNN)[4], C4.5 Karar Ağacı[5], Destek Vektör Makineleri (SVM)[6] ve Lojistik Regresyon (LR)[7] yöntemleri kullanılmıştır.

Algoritmaların karartılmış bankacılık veri setleri üzerinde koşturulması sonucu elde edilen sonuçlar, Duyarlılık, Özgüllük, Doğruluk, Gini Katsayısı ve ROC eğrisi altında kalan alan(AUC) olmak üzere 5 farklı metriğe göre karşılaştırılmıştır.

Bu bildirinin 2. Bölümünde, kredi skorlama hakkında daha önceden yapılmış benzer çalışmalar anlatılmaktadır. 3. Bölümde önerilen derin öğrenme ve makine öğrenimi yöntemleri özetlemektedir. 4. Bölümde, önerilen mimarinin kullanılabilirliği göstermek için geliştirdiğimiz prototip uygulamanın detayları sunulmaktadır. Uygulama ile ilgili yaptığımız değerlendirme 5. Bölümde yer almaktadır. Son bölüm olan 6. Bölümde çalışmanın sonuçları tartışılmakta ve geleceğe dönük çalışmalar özetlenmektedir.

978-1-5386-1501-0/18/$31.00 © 2018 IEEE

II. LİTERATÜR TARAMASI

Literatür incelendiğinde, kredi skorlama konusu üzerinde yapılmış birçok çalışma bulunmaktadır.

Jakup M. Tomczak[8] tarafından gerçekleştirilen çalışmada, kredi risk analizinin değerlendirilmesi için 4 farklı bankacılık veri seti üzerinde Sınıflandırılmış Kısıtlı Boltzmann Makinesinin performansı ölçülmüştür. Kullanılan yöntemin, büyük ve lineer olarak ayrılamayan veri setlerinde, geleneksel makine öğrenmesi yöntemlerine göre daha başarılı sonuçlar ürettiği gözlemlenmiştir.

Desai[9] tarafından yapılan araştırmada, YSA, Doğrusal Diskriminant Analizi (LDA) ve Lojistik Regresyon olmak üzere 3 farklı algoritma kullanılmıştır. Çalışmada kullanılan veri seti 3 farklı kredi birliğinden toplanan ve kredi talebinde bulunan müşterilere ait bilgilerden oluşmaktadır. Çalışma sonucunda, kötü sınıfa ait kredilerin sınıflandırılmasında YSA yüksek doğruluk oranı sunmaktadır.

Moares[10] tarafından yapılan çalışmada, müşterilerin kredi profillerini iyi veya kötü olarak sınıflandırmak için C4.5 Karar Ağacı ve YSA kullanılmıştır. Yapılan çalışma sonucunda, C4.5 Karar Ağacı ile %90.07 doğruluk oranı elde edilmiştir. YSA ise %95,58'lik bir doğruluk oranı sunmuştur.

III. METODOLOJİ

Bu çalışmanın sistem mimarisi 3 ana başlık altında incelenmektedir. İlk olarak ham veri setinin alınması ve ön işlemeden geçirilmesi işlemi yapılmaktadır. Daha sonra kullanıcı tarafından seçilen bir algoritma ile veri seti modellenmektedir. Son olarak, elde edilen model test edilerek modelin performansının ölçüm işlemi gerçekleştirilmektedir.

Sistemin çalışma yapısı Şekil 1'de gösterilmektedir. Öncelikle ham veri seti üzerinde eksik veri tamamlama ve özellik seçimi işlemi yapılır(1). Ön işlenmiş veri normalizasyon ve dummification (kukla değişken dönüşümü) işlemlerinden geçirildikten sonra modellenmeye hazır hale getirilmiş veri seti elde edilir(2). Veri derin öğrenme yöntemleri veya makine öğrenimi yöntemlerinden biri ile modellenir(3). Modellenen veri üzerinden bir desen(pattern) çıkarılır(4). Elde edilen çıktılar doğruluk, duyarlılık, özgüllük, Gini katsayısı ve AUC metrikleri ile değerlendirilir(5).

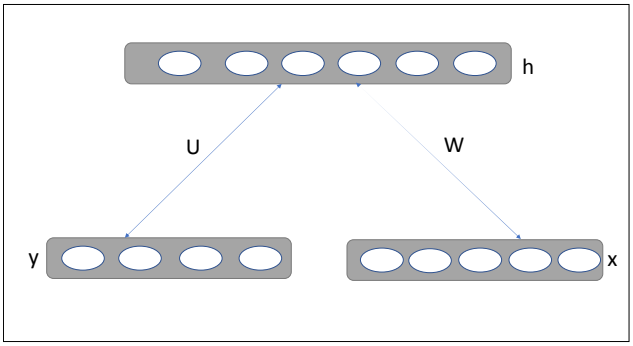
A. ClassRBM algoritması

ClassRBM algoritması stokastik kararlar veren nöronlardan oluşan, simetrik bağlantılı bir yapay sinir ağıdır. Birbirine yönsüz olarak bağlı 3 katmandan oluşmaktadır. Birinci katman, veri setindeki öznitelik değerlerini ($x \in \{0,1\}^D$) barındıran görünür giriş katmanıdır. İkinci katman, gizli değişkenlerden ($h \in \{0,1\}^D$) oluşmaktadır. Üçüncü katman ise veri setindeki etiket değerlerini($y \in \{1,2,\ldots,K\}$) barındıran çıktı katmanı olarak isimlendirilmektedir. Çıktı K uzunluklu bir ikili(binary) vektör ile tanımlanmaktadır. Böylece çıktı(sınıf etiketi) K ise y_k 1 değerini, y_k dışındaki tüm elemanlar 0 değerini almaktadır[2].

ClassRBM ağında sadece katmanlar arasında bağlantı bulunmaktadır. Aynı katmandaki nöronlar arasında herhangi bir bağlantı bulunmamaktadır. Her bir durum farklı enerji seviyesi ile ilişkilendirilmektedir. Durumların enerji seviyesinin hesaplanması için Formül-1 kullanılmaktadır.

$$E(x,y,h) = -h^T Wx - b^T x - c^T h - d^T y - h^T Uy \qquad (1)$$

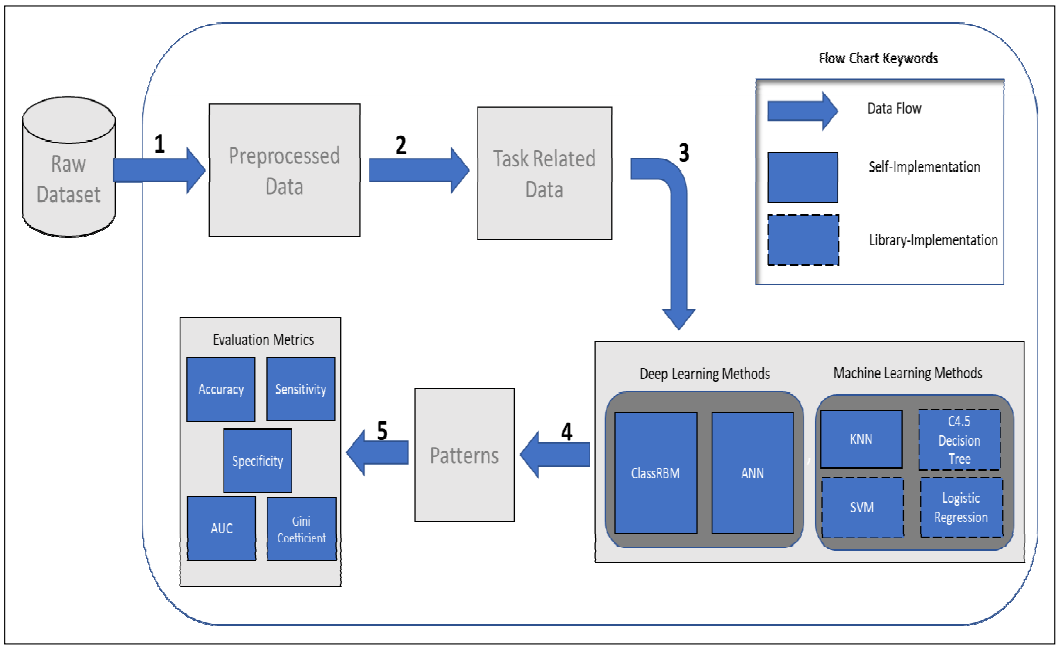
Formüldeki parametre değerleri $\theta = \{b,c,d,W,U\}$'dur. Buradaki b,c,d değerleri sırasıyla giriş, gizli ve çıkış(sınıf değeri) katmanlarının bias değerlerine karşılık gelmektedir. W, giriş katmanı ile gizli katman arasındaki ağırlık değeri, U ise çıkış katmanı ile gizli katman arasındaki ağırlık değeridir.



Şekil. 2. ClassRBM Ağ Yapısı

ClassRBM ağı Şekil 2'de görüldüğü gibi yönsüz(undirected) bir ağdır. Diğer yönlü ağlarda (Ör:ANN), gradient descent süreci ile oluşan hatalar ağırlıklara yansıtılarak ağ eğitilmektedir. ClassRBM ağının eğitimi için ise Contrastive Divergence algoritması kullanılır. Contrastive Divergence algoritmasının sözde kodu Şekil 3'de yer almaktadır.

M adet gizli nörona sahip bir ClassRBM ağı, görünür ve gizli katmanların ortak dağılımının parametrik bir modelidir. ClassRBM ağının en önemli avantajı yeterli sayıda gizli nöron ile ikili vektör üzerindeki herhangi bir dağılımı temsil edebilir. Ele alınan problemde x ikili vektörü kredi başvurusunda bulunan müşterinin karakteristiğini temsil etmektedir. Çıktı değerini temsil eden y, kredinin onaylanıp onaylanmadığını belirtmektedir. Gizli katmanı belirten h ise, kredi başvurusunda bulunan müşteriyi temsil eden tüm alanın dağılımını göstermektedir



Şekil. 1. Sistemin Çalışma Yapısı

Algoritma 1: Contrastive Divergence

Girdi: eğitim verisi çifti (y_i,x_i) , öğrenme katsayısı λ ve örnekleme için iterasyon sayısı T

% Gösterim: a ← b'in anlamı a b'ye atandı.
% sigm, sigmoid fonksiyonuna karşılık gelmektedir.

%İlklendirme
$y^0 \leftarrow y_i$, $x^0 \leftarrow x_i$, $h_0 \leftarrow sigm(c + Wx^0 + Uy^0)$

for i **in range**(T):
  % İleri besleme
  $h^i \leftarrow sigm(c + Wx^i + Uy^i)$

  % Geri besleme
  $y^i \leftarrow sigm(h^iU+d)$, $x^i \leftarrow sigm(h^iW+b)$
end for

for $\theta \in \Theta$ **do**
  $\theta \leftarrow \theta - \lambda[\frac{\partial}{\partial\theta}E(y^0,x^0, h^0) - \frac{\partial}{\partial\theta}E(y^T,x^T, h^T)]$
end for

Şekil. 3. ClassRBM'in eğitim aşaması için kullanılan contrastive divergence algoritmasının sözde kodu

B. Çok katmanlı yapay sinir ağları

Çok katmanlı yapay sinir ağları, yapay nöronlardan oluşur ve insan beynin basit bir modelini gerçeklemektedirler. Fakat bu ağlar gerçek beyin yapısı ile karşılaştırıldığında çok basit kalmaktadır. Yapay sinir ağları, eğitim veri setinden faydalanarak yapısında bulunan nöronlar arsındaki bağlantıların ağırlıklarını bulmaya çalışır. Optimum ağırlık değerlerinin bulunması doğruluk oranını artırır[3].

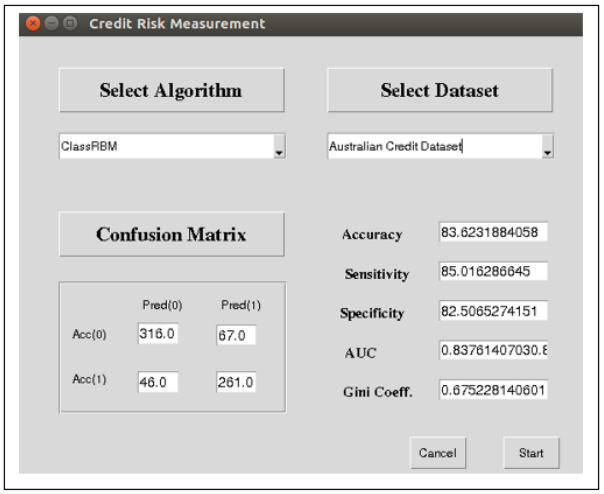
IV. PROTOTİP UYGULAMA

Çalışma kapsamında, bankacılık alanındaki veri setleri üzerinde koşturulan ve en uygun kredi skorlama modelinin tespitini sağlayan bir çözüm geliştirilmiştir. Çalışmada, güncel derin öğrenme ve makine öğrenimi algoritmaları kullanılmıştır. Algoritmaların çalıştırılması sonucu elde edilen modellerden faydalanılarak bankaların müşterilerinin kredi taleplerini değerlendirme aşamasında karar verme süreçlerinin hızlandırılması hedeflenmiştir.

Çalışma kapsamında geliştirilen uygulama Python 3.6 dili ile kodlanmıştır. Uygulamanın minimum sistem gereksinimleri 1GB RAM ve 100MB disk alanıdır.

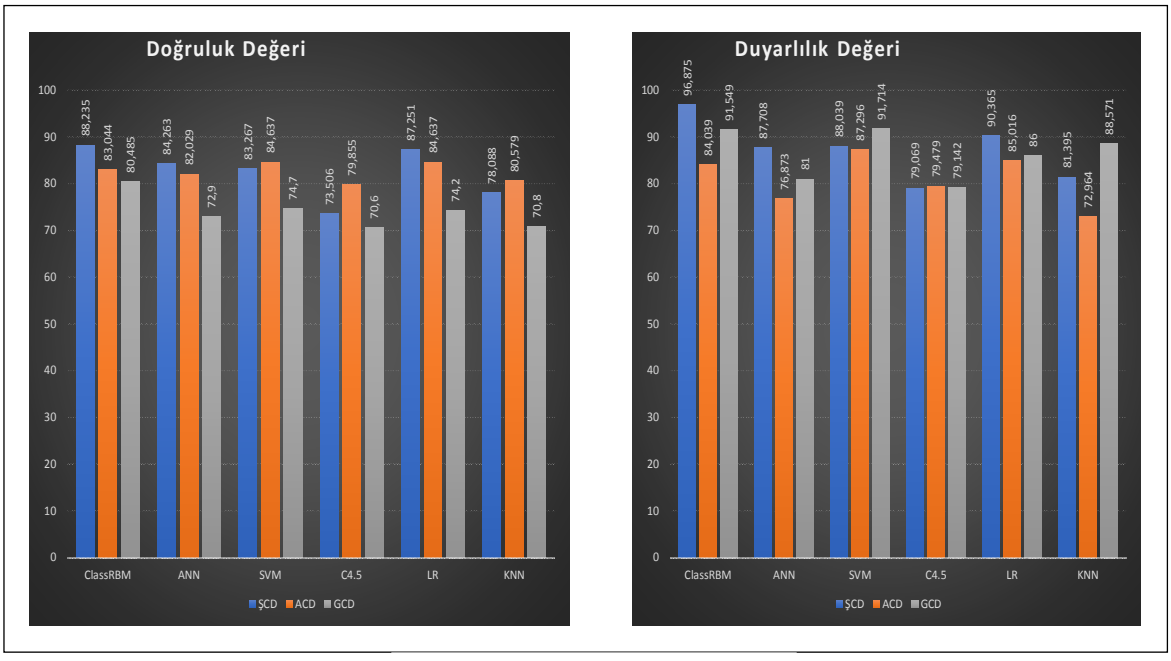
Derin öğrenme yöntemleri uygulanırken karmaşık matematiksel hesaplamaların yapılması amacıyla PyTorch 0.2.0 sürümlü kütüphane kullanılmıştır. PyTorch bilimsel hesaplamalar için kullanılan, açık kaynak kodlu bir kütüphanedir. Sadece Linux tabanlı işletim sistemleri üzerinde çalışabilmekte ve yalnız Python programlama dili ile birlikte kullanılabilmektedir. Hem CPU hem de GPU'lar üzerinde çalışabilmektedir.

Makine öğrenimi yöntemleri uygulanırken Python programlama dilinin Scikit-learn 0.19.0 sürümlü kütüphanesi kullanılmıştır. Scikit-learn veri analizi için yaygın olarak kullanılan açık kaynak kodlu bir kütüphanedir. Birçok sınıflama ve kümeleme yöntemini barındırmasının yanı sıra, verideki eksik değerleri doldurmak, öznitelik seçmek, verileri normalize etmek gibi çeşitli veri ön işleme adımlarına de sahiptir.



Şekil. 4. Sistem için tasarlanan kullanıcı arayüzü

Şekil 4'de tasarlanan sisteme ait kullanıcı arayüzü görülmektedir. Start butonuna basıldığında seçilen veri seti üzerinde algoritmalar çalıştırılmakta ve oluşturulan modeller 10 Cross Validation işlemi ile test edilmektedir. Test işlemi modellerin oluşturulmasından hemen sonra otomatik olarak gerçekleşmekte ve test sonucu elde edilen karmaşıklık matrisi ekrana yazdırılmaktadır. Her bir modele ait karmaşıklık matrisi ve bunun yanında modele ait doğruluk, duyarlılık, gini katsayısı ve ROC eğrisi altında kalan alan (AUC) bilgileri kullanıcıya sunulmaktadır.



Şekil. 5. Sistemin Başarı Durumu

V. DEĞERLENDİRME

Bu bölümde sistem oluşturulurken kullanılan teknolojik araçlar ve sistemin çalışma performansından söz edilmiştir.

Sistem kodu gerçekleştirilirken Python 3.6 sürümü kullanılmıştır. Derin öğrenme yöntemleri uygulanırken karmaşık matematiksel hesaplamaların yapılması amacıyla PyTorch 0.2.0 sürümü kullanılırken, makine öğrenmesi algoritmalarını gerçeklemek için Python programlama dilinin scikit-learn 0.19.0 sürümlü kütüphanesi kullanılmıştır.

Sistemimiz şu özelliklerdeki makine üzerinde çalıştırıldı; MSI GE62 Laptop, Intel i7-5700HQ (2.70GHz) işlemci, 8GB RAM, Ubuntu 16.04 LTS işletim sistemi.

Sistemin performansı, farklı veri setleri üzerinde çalışan algoritmaların ürettiği doğruluk(accuracy), duyarlılık (sensitivity), özgüllük(specificity),gini katsayısı ve ROC eğrisi altında kalan alana göre değerlendirilmiştir. Sistemin başarısı ölçülürken, K katlamalı çapraz doğrulama (K Fold Cross Validation) yöntemi kullanılmıştır.

Şekil 5'de sistemin 3 farklı veri seti üzerinde elde ettiği doğruluk ve duyarlılık değerleri, Şekil 6'da ise özgüllük ve AUC değerleri gösterilmektedir. Sistemin üzerinde çalıştığı veri setleri karartılmış bankacılık veri setleridir. ŞCD (Şekerbank Credit Dataset) 9'u kategorik, 26'sı nümerik olmak üzere 35 öznitelikten oluşmaktadır. ACD (Australian Credit Dataset) 8'i kategorik 6'sı nümerik olmak üzere 14 öznitelikten oluşmaktadır. GCD (German Credit Dataset) ise 17'si kategorik,3'ü nümerik olmak üzere 20 öznitelikten oluşmaktadır.

Bunun yanında ClassRBM algoritmasında model oluşturma safhasında makine öğrenmesi yöntemlerinin aksine faktör seçme işlemi yoktur. Böylece veride gizli bulunan ve doğrusal olmayan desenleri çıkartması açısından makine öğrenimi algoritmalarına göre daha başarılıdır.

Makine öğrenmesi yöntemlerinin eğitim süresi, derin öğrenme yöntemlerine göre ihmal edilebilir düzeydedir. Makine öğrenmesi yöntemleri CPU üzerinde efektif bir şekilde çalışabiliyorken, derin öğrenme yöntemleri cevap süresini kısaltmak için GPU'lara ihtiyaç duymaktadır.

VI. SONUÇLAR VE GELECEKTEKİ ÇALIŞMALAR

Bu araştırma kapsamında, bankacılık alanındaki veri setleri üzerinde koşturulan ve en uygun kredi skorlama modelinin tespitini sağlayan bir çözüm geliştirilmiştir. Araştırma kapsamında güncel derin öğrenme ve makine öğrenimi algoritmaları kullanılmıştır. Algoritmaların çalıştırılması sonucu elde edilen modellerden faydalanılarak bankaların müşterilerinin kredi taleplerini değerlendirme aşamasında karar verme süreçleri hızlandırabilmektedir.

Üçüncü bölümde detaylı bir şekilde incelenen derin öğrenme algoritmalarının, lineer olarak ayrılamayan bankacılık veri setleri üzerinde, makine öğrenimi algoritmalarına göre daha başarılı sonuçlar verdiği gözlemlenmiştir. Bu, derin öğrenme yöntemlerinin model oluşturma safhasında faktör seçme işlemi olmamasından kaynaklanmaktadır.

Gelecek çalışmalarda, Derin İnanç Ağları (Deep Belief Network) ve Oto Kodlayıcı (AutoEncoders) gibi güncel derin öğrenime algoritmalarının sisteme dâhil edilmesi planlanmaktadır.

KAYNAKÇA

[1] Demirbulut, Y. E., Aktaş, M. S., Kalıpsız, O., & Bayracı, S. İstatistiksel ve Makine Öğrenimi Yöntemleriyle Kredi Skorlama.

[2] Larochelle, H., Mandel, M., Pascanu, R., & Bengio, Y. (2012). Learning algorithms for the classification restricted boltzmann machine. *Journal of Machine Learning Research*, *13*(Mar), 643-669.

[3] Minns, A. W., & Hall, M. J. (1996). Artificial neural networks as rainfall-runoff models. *Hydrological sciences journal*, *41*(3), 399-417.

[4] Weinberger, K. Q., Blitzer, J., & Saul, L. K. (2006). Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems* (pp. 1473-1480).

[5] Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.

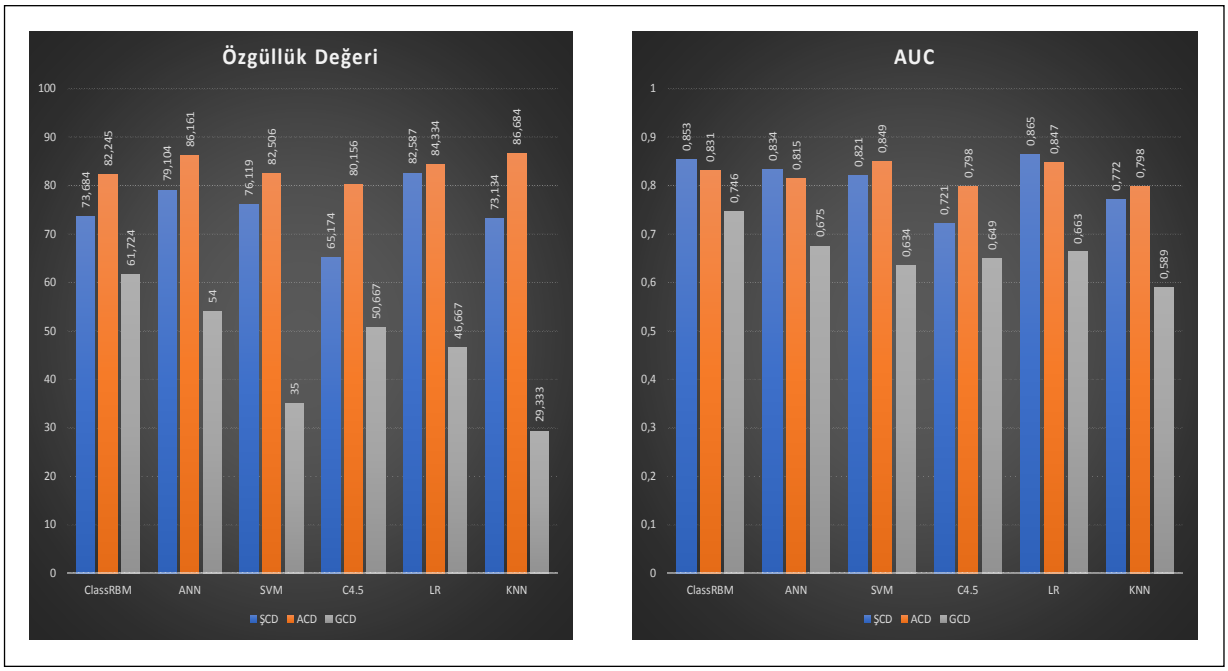
[6] Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural processing letters*, *9*(3), 293-300.

[7] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

[8] Tomczak, J. M., & Zięba, M. (2015). Classification restricted Boltzmann machine for comprehensible credit scoring model. *Expert Systems with Applications*, *42*(4), 1789-1796.

[9] Desai, V. S., Crook, J. N., & Overstreet, G. A. (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research*, *95*(1), 24-37.

[10] Sousa, M. D. M., & Figueiredo, R. S. (2014). Credit analysis using data mining: application in the case of a credit union. *JISTEM-Journal of Information Systems and Technology Management*, *11*(2), 379-396.



Şekil. 6. Sistemin Başarı Durumu