TÜBİTAK 2209-A ÜNİVERSİTE ÖĞRENCİLERİ
YURT İÇİ ARAŞTIRMA PROJELERİ
DESTEK PROGRAMI

YAPAY ZEKÂ TABANLI BÜYÜK VERİ YÖNETİM ARACININ TASARIMI VE UYGULAMASI

KARADENİZ TEKNİK ÜNİVERSİTESİ
YAZILIM MÜHENDİSLİĞİ BÖLÜMÜ

TEMATİK ALANI

YAZILIM MÜHENDİSLİĞİ
BİLİŞİM TEKNOLOJİLERİ MÜHENDİSLİĞİ

FATİH ARSLAN
Proje Yürütücüsü

DOÇ. DR. HAMDİ TOLGA KAHRAMAN
Proje Danışmanı

1. ÖZET

Teknolojik gelişmelerin en hızlı yaşandığı ve hissedildiği alanların başında yapay zekâ gelmektedir. Yapay zekânın gelişimi üzerinde etkili olan başlıca öğelerden biri ise veridir. Veri, geçmişten günümüze bilginin oluşmasındaki asıl kaynaktır. İnsanın bir hücresinde tutulan verinin terabaytlar mertebesinde olduğu düşünüldüğünde biyolojik olarak insanın veri depolama kapasitesinin günümüz teknolojisinden üstün olduğu söylenebilir. Bunun yanında veri işleme teknolojisinde son yıllarda yaşanan gelişmeleri göz ardı etmek mümkün değildir. Çok çekirdekli, paralel işlem yapabilen, yüksek hızlı ve büyük kapasiteli yeni nesil işlemciler yanında veriyi bilgiye dönüştürmede insan ve diğer canlılara benzer karar mekanizmaları kullanan algoritmalar sayesinde çok güçlü veri madenciliği araçları geliştirilmektedir. Şirketler ürünlerinden daha fazla kazanç elde edebilmek, çalışmalarında verimliliği artırmak ve karar mekanizmalarını güçlendirebilmek için veri madenciliği için geliştirilmiş araçlardan faydalanmaktadırlar. Bu süreçte veri kalitesi ve bütünlüğü için harcadıkları zamanı ve bütçeyi artırmaktadırlar. Bunun nedeni, veri kalitesi ve bütünlüğünün firmalara katmakta olduğu değerin fark yaratmasıdır (Deloitte, 2009).

Yapay zekânın bir alt disiplini olan veri madenciliği teknolojik gelişmelerden etkilenerek büyük veri madenciliğine evrilmiştir. Bu değişimin temelinde ise internet-tabanlı alış-verişten uzay araçları ile yapılan haberleşmeye, endüstriyel otomasyon uygulamalarından sosyal medya uygulamalarına kadar sayısız alanda büyük bir veri yığının ortaya çıkması ve bunların elektronik ortamda saklanması ihtiyacı gelmektedir. Yarı iletken teknolojisindeki gelişmeler, veriyi depolama ve işleme kapasitesi yüksek elektronik malzemelerin üretilmesini sağlamıştır. Verideki ve işlem yapma kapasitesindeki artış ise veriyi işleyerek anlamlı bilgiye dönüştürmede kullanılan yapay zekâ algoritmalarının tekrar gözden geçirilmesine neden olmuştur. Geçmişte, günümüze kıyasla küçük sayılabilecek veri yığınları üzerinde etkili sonuçlar üreten algoritmalar bugün ise büyük veri işleme gereksinimini aynı şekilde karşılayamamaktadırlar. Günümüzde büyük veriyi işleyerek en kısa sürede en uygun sonucu bulmak giderek zor bir hale gelmiştir. En uygun sonuç, kabul edilebilir bir maliyet ile en kaliteli olan sonuçtur. Başarılı bir veri madenciliği için veriyi işleyecek algoritmadan daha önemlisi verinin kendisidir. Verinin, problem uzayını homojen bir şekilde örneklemesi ve doğru olması gerekir. Yani başarılı bir veri madenciliği etkisi elde etmek için kaliteli bir veriye ihtiyaç vardır. Kaliteli bir veriyi elde etmek için ise verinin incelenmesi, analiz edilmesi ve işlenmeye hazır hale getirilmesi gerekir. Veri analiz süreci, bilimsel araştırma sürecinin ve veri madenciliğinin en önemli basamaklarından biridir. Bu süreçte toplanan veriler amaca en uygun istatiksel, matematiksel ya da yapay zekâ teknikleriyle işlenir veya analiz edilir.

Veri analiz sürecinde karşılaşılan problemlerin başında, eksik ya da gürültülü/hatalı verilerin tespit edilmesi ve düzeltilmesi gelmektedir. Bu sorunun çözümünde kullanılmak üzere geçmişten günümüze farklı yöntemler geliştirilmiştir. Eksik veri ile analize devam etme, eksik gözlemleri analiz dışı bırakma, eksik gözlemler yerine veri atama veya çeşitli istatistiksel yöntemlerle eksik verileri tamamlama gibi yöntemler bu durumlarda sıkça kullanılmaktadırlar (Little, 1988; Duncan, Duncan ve Li, 1998; Downey ve King, 1998; Bal, 2003; Carpita ve Manisera, 2011). Bu yöntemler içerisinde araştırmacılar tarafından en çok kullanılan yöntemler, liste bazında silme ve çiftler bazında silme gibi eksik verileri analiz dışı bırakma yöntemleridir. Ancak yapılan çalışmalar bu yöntemlerin örneklemde kayba, güvenirlikte azalmaya, tahminlerde yanlılığa neden olduğunu (Oğuzlar, 2001; Allison, 2009; Satıcı ve Kadılar, 2009; Van Der Ark ve Vermunt,2010; Cumming, 2013) ve yanlılıktan kaynaklı olarak da örneklemin evreni temsil etme derecesinin düştüğünü göstermektedir (Little, 1988; Demir ve Parlak, 2012). Belirtilen bu sebeplerden dolayı, son yıllarda, bu yöntemler yerine, beklenti maksimizasyonu ve çoklu atama gibi modern yöntemler önerilmektedir. Çünkü bu yöntemler, silme yöntemleri gibi geleneksel kayıp veri yöntemlerinin aksine, yanlılığın azaltılması, etkili parametre tahminlerinin yapılması ve daha büyük istatistiksel gücün sağlanması hususunda daha etkili sonuçlar vermektedir (Enders, 2013).

Bu proje çalışmasında eksik ve gürültülü verilerin tespit edilmesi ve düzeltilmesi amacıyla yapay zekâ tabanlı çalışan, özgün ve güçlü bir veri yönetim aracı geliştirilecektir. Bu araç sayesinde veri setlerinin analiz edilmesi, bu veri setlerindeki eksik ve gürültülü verilerin tespit edilmesi ve düzeltilmesi sağlanacaktır. Geliştirilecek yazılım aracının özgünlüğü ise eksik ve gürültülü verileri düzeltme sürecinde modern, güçlü ve melez yapay zekâ algoritmalarını kullanacak olmasıdır. Bu süreçte sezgisel optimizasyon algoritmalarından (genetik algoritma (Holland, 1975), yapay arı kolonisi algoritması

(Karaboğa ve Baştürk, 2007), ortak yaşayan organizmalar (Cheng ve Prayogo, 2014)) ve meta-sezgisel tahmin ve sınıflandırma algoritmalarından (Kahraman, 2016)) faydalanılacaktır. Geliştirilecek uygulamanın basit bir ara yüz ile kullanılması ve problemlere ait veri setlerinin kolaylıkla düzenlenebilmesi sağlanacaktır. Veri setindeki eksiklikler ve gürültülü veriler veri yönetim yazılımı tarafından otomatik olarak tespit edilip araştırmacılara rapor halinde sunulacaktır. Hata tespiti yapıldıktan sonra, program hatalı verilerin yerine geçebilecek en uygun değerleri bulup değiştirme işlemini yapacaktır. Bu süreçte melez bir tahmin ve sınıflandırma tekniği olan "meta-sezgisel k-nn algoritması" ve doğrusal olmayan regresyon problemlerinin çözümlenmesi amacıyla literatürde en yaygın kullanılan algoritma olan yapay sinir ağları kullanılacaktır. Geliştirilecek veri yönetim aracının kolay erişilebilir olması açısından internet ortamında çalışacak şekilde web-uygulaması olarak gerçekleştirilmesi planlanmaktadır. Günümüzde birçok kişinin kolayca erişim sağlayabildiği web sunucuları sayesinde uygulamalar daha geniş kesime hitap edip, daha fazla kullanıcı ile buluşmaktadır. Web uygulamalarında;

- Kullanıcılar, internet vasıtasıyla her yerden uygulamaya erişim sağlayabilir.
- Erişim sırasında herhangi bir yazılım indirme gereksinimi duyulmaz.
- Sunucu tarafı güncellenmesi halinde tüm kullanıcılar son sürümü kullanırlar.
- Platform esnekliği sağlar. Her işletim sisteminde (MacOsx, Windows, Linux, vs…) ve her tarayıcıda sorunsuzca çalışabilir.

Bu proje çalışmasında geliştirilecek olan yapay zekâ tabanlı veri yönetim aracı, hem web sunucusundan kullanıcılara hizmet verecek hem de büyük veri işleme sistemlerine (Hadoop, Spark gibi frameworkler.) entegre edilebilen bir araç olacaktır. Bu yönüyle oldukça esnek ve modüler bir yazılım aracı niteliği taşıyacaktır.