

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

AKADEMİK BAŞARININ MODELLENMESİNDE ÇİZGE
MADENCİLİĞİ YAKLAŞIMI

BAŞAK BULUZ
YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

GEBZE
2017

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

AKADEMİK BAŞARININ
MODELLENMESİNDE ÇİZGE
MADENCİLİĞİ YAKLAŞIMI

BAŞAK BULUZ
YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

DANIŞMANI
YARD. DOÇ. DR. BURCU YILMAZ

GEBZE
2017

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

**GRAPH MINING APPROACH FOR
MODELING ACADEMIC SUCCESS**

BASAK BULUZ
A THESIS SUBMITTED FOR THE DEGREE OF
MASTER OF SCIENCE
DEPARTMENT OF COMPUTER ENGINEERING

THESIS SUPERVISOR
ASST. PROF. DR. BURCU YILMAZ

GEBZE
2017

GEBZE TEKNİK ÜNİVERSİTESİ	YÜKSEK LİSANS JÜRİ ONAY FORMU
----------------------------------	--------------------------------------

GTÜ Fen Bilimleri Enstitüsü Yönetim Kurulu'nun/...../..... tarih ve/..... sayılı kararıyla oluşturulan jüri tarafından/...../..... tarihinde tez savunma sınavı yapılan Başak BULUZ'un tez çalışması Bilgisayar Mühendisliği Anabilim Dalında YÜKSEK LİSANS tezi olarak kabul edilmiştir.

JÜRİ

ÜYE

(TEZ DANIŞMANI) : Yard. Doç. Dr. Burcu YILMAZ

ÜYE

: Yard. Doç. Dr. Yakup GENÇ

ÜYE

: Yard. Doç. Dr. Alev MUTLU

YEDEK ÜYE

: Yard. Doç. Dr. Ayşe Betül OKTAY

ONAY

Gebze Teknik Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun
...../...../..... tarih ve/..... sayılı kararı.

İMZA/MÜHÜR

ÖZET

Öğrencinin akademik başarısının tahmini, nedenlere ve süreçlere göre başarısının değişimi ve bu değişimin sonuçlarının incelenmesi farklı disiplinlerden birçok bilim dalını ilgilendiren genel bir araştırma konusudur. Bu konu da gerek bilgisayar bilimleri gerekse eğitim bilimleri alanlarında literatürde birçok çalışmaya rastlamak mümkündür.

Bilgisayar bilimleri çerçevesinde literatürde mevcut bulunan çalışmaların birçoğunda akademik başarının modellenmesi konusunda öğrencilerin yalnızca demografik özellikleri veya yalnızca geçmiş akademik başarıları temel alınmıştır. Bu çalışmada getirilen yaklaşımda ise öncelikle kişilerin demografik ve akademik geçmiş bilgileri bir araya getirilmiş ve geleneksel veri madenciliği (Data Mining-DM) sınıflandırma metotlarından olan Naive Bayes ve K-En Yakın Komşuluk Yöntemi kullanılarak oluşturulan modellerin performansları incelenmiştir. Ardından her bir öğrencinin demografik verileri ve akademik geçmişinden elde edilen birçok veriden oluşan veri kümesi çizgilerle ifade edilmiş ve bu çizgilerden sıklıkla tekrar eden alt-çizgiler bulunarak, başarılı ve başarısız öğrencilerin özelliklerini ifade eden çeşitli örüntüler elde edilmiştir. Böylece veri kümesinin zenginleşmesi sağlanmıştır. Alt çizgilerle zenginleştirilen veri kümesi üzerine tekrar aynı geleneksel sınıflandırma yöntemleri uygulanmış ve oluşturulan yeni modellerin performansları ayrıca incelenmiştir. Her iki sonuç karşılaştırıldığında görülmüştür ki; alt-çizgilerle zenginleştirilmiş veri kümeleri ile oluşturulan modeller ile yapılan sınıflandırmalar performans bakımından daha iyi sonuçlar vermiştir.

Anahtar Kelimeler: Akademik Başarının Modellenmesi, Veri Madenciliği, Çizge Tabanlı Veri Madenciliği.

SUMMARY

The prediction of the academic success of a student, the change of success according to causes and processes, and the examination of the consequences of the change are general research topics that deals with many disciplines. Within this subject, it is possible to reach a lot of studies in the literature both in computer science and educational science fields.

Most of the studies in the field of computer science are based on either the demographic characteristics of the students or the past academic achievements in modeling academic success. In the thesis, both demographic and past academic success information of the students were collected and the performances of the models are evaluated using Naive Bayes and K-Nearest Neighborhood Methods, which are the traditional data mining classification methods. A dataset, where the demographic and academic background of each student are expressed in sets of graphs, are constructed. The data set is enriched with the frequent repetition of these graphs. The same traditional classification methods are applied again on the dataset enriched by the sub-graphs and the performances of the new models have also been examined. It is seen that when both results are compared, the classifications accuracy of the models and the dataset enriched with frequent sub-graphs yielded better than the unenriched datasets.

Key Words: Modeling Academic Success, Data Mining, Graph Mining.

TEŞEKKÜR

Yüksek lisans eğitimim ve tez çalışmam süresince sabır, anlayış ve özveriyle desteğini, bilgisini ve kıymetli vaktini biran olsun esirgemeyen, danışmanım olarak beni onurlandıran saygıdeğer hocam Yard. Doç. Dr. Burcu YILMAZ'a; akademik hayatımın ilk gününden bugüne kadar her an yanımda olan, desteğini daima hissettiğim, her konuda bilgisini ve tecrübelerini benimle paylaşan değerli hocam Yard.Doç.Dr. Ferdi SÖNMEZ'e ; beraber çalışmaktan onur duyduğum, tez çalışmam boyunca eleştiri ve yönlendirmeleri ile yol almamda yardımcı olan mesai arkadaşım ve hocam Öğr. Gör. Enis KARAÇAM'a; mezun öğrencisi ve çalışanı olmaktan gurur duyduğum İstanbul Aydın Üniversitesi bünyesinde ki tüm hocalarıma, tez çalışmamda kullandığım verilere ulaşmamda emeğini esirgemeyen idari personelimize ve çalışma arkadaşlarıma sabırları ve destekleri için minnettarım.

Değerli vakitlerini ayırıp, çalışmama değer vererek tez jürimde bulunmayı kabul eden çok kıymetli hocalarım Yard.Doç.Dr. Yakup GENÇ, Yard.Doç.Dr. Alev MUTLU ve Yard.Doç.Dr. Ayşe Betül OKTAY' a ayrıca teşekkür ederim.

Hayatımın her anında varlıklarına şükrettiğim annem Şadıman BULUZ, babam Suat BULUZ ve ağabeyim Murat Can BULUZ'a göstermiş oldukları sabır, şefkat, anlayış ve destek için teşekkürü borç bilirim.

İÇİNDEKİLER

	<u>Sayfa</u>
ÖZET	v
SUMMARY	vi
TEŞEKKÜR	vii
İÇİNDEKİLER	viii
SİMGELER ve KISALTMALAR DİZİNİ	x
ŞEKİLLER DİZİNİ	xi
TABLolar DİZİNİ	xii
1. GİRİŞ	1
1.1. Tezin Amacı, Katkısı ve İçeriği	1
1.2. Tezin Organizasyonu	2
2. LİTERATÜR ANALİZİ	4
2.1. Veri Madenciliği	4
2.1.1. Eğitimsel Veri Madenciliği	4
2.1.2. Çizge Tabanlı Veri Madenciliği	6
2.1.3. Çizge Veri Kümesinden Tekrarlayan Alt çizgelerin Çıkarılması	7
2.2. Makine Öğrenmesi	13
2.3. Sınıflandırma Metotları	13
2.3.1. Naive Bayes Sınıflandırıcı	13
2.3.2. K-En Yakın Komşuluk Yöntemi (k-NN)	15
2.3.3. Destek Vektör Makineleri	16
2.3.4. Karar Ağaçları	17
2.3.5. Yapay Sinir Ağları	18
3. PROBLEMİN TANIMI VE HİPOTEZ	19
4. YÖNTEM	21
4.1. Veri Kümesinin Oluşturulması	21
4.1.1. Özellik Seçim Yöntemleri	24
4.1.1.1. Bilgi Kazanımı (Information Gain)	24
4.1.1.2. Kazanım Oranı (Gain Ratio)	25
4.1.1.3. Ki-kare (Chi-square)	25

4.1.2. Anahtar Sözcük Çıkarımı	26
4.1.3. Çizge Veri Yapısının Oluşturulması	27
4.1.4. Çizge Madenciliği Yöntemleri ile Örüntülerin Çıkarılması	29
4.1.4.1. SUBDUE	30
4.2. Özellik Vektörlerinin Oluşturulması	33
4.3. Sınıflandırma Metotları	33
4.3.1. Naive Bayes Sınıflandırıcı (NB)	33
4.3.2. K-En Yakın Komşuluk Yöntemi (k-NN)	35
5. DEĞERLENDİRME	38
6. SONUÇLAR ve ÖNERİLER	47
KAYNAKLAR	49
ÖZGEÇMİŞ	55

SİMGELER ve KISALTMALAR DİZİNİ

Simgeler ve Açıklamalar Kısaltmalar

VM	: Veri Madenciliği (Data Mining-DM)
EVM	: Eğitimsel Veri Madenciliği (Educational Data Mining-EDM)
ÇİVM	: Çok ilişkisel Veri Madenciliği (Multi-relational Data Mining-MRDM)
TMP	: Tümevarımsal Mantık Programlama (Inductive Logic Programming-ILP)
GÖA	: Genişlik Öncelikli Arama (Breadth First Search-BFS)
DÖA	: Derinlik Öncelikli Arama (Depth First Search-DFS)
MÖ	: Makine Öğrenmesi (Machine Learning-ML)
ÇKA	: Çok Katmanlı Algılayıcı (Multilayer Perceptron-MP)
k-NN	: k-En Yakın Komşuluk Yöntemi (k-Nearest Neighbour- k-NN)
DVM	: Destek Vektör Makinesi (Support Vector Machine-SVM)
MTU	: Minimum Tanımlama Uzaklığı (Minimum Description Length-MDL)

ŞEKİLLER DİZİNİ

<u>Sekil No:</u>	<u>Sayfa</u>
2.1: Eğitimsel veri madenciliği süreçleri.	5
2.2: Örnek bir çizge veri kümesinden minimum tekrar sayısı 3 için elde edilen alt çizge.	8
4.1: Başarısız bir öğrenci için oluşturulmuş örnek çizge yapısı.	28
4.2: Çizge yapısının gösteriminin basit ifadesi.	28
4.3: Çizgelerin metinsel ifade edilmesi örneği.	29
4.4: Subdue keşif algoritmasının sözde kodu.	31
4.5: SUBDUE yöntemi ile alt çizgelerin elde edilmesi örneği.	32
4.6: Başarısız öğrenciler için elde edilen örnek alt çizge.	32
5.1: Örnek çizge yapısı.	40
5.2: Eklenen alt-çizge sayısına göre Naive Bayes sınıflandırıcının performans değişimi.	42
5.3: Eklenen alt-çizge sayısına göre K-NN sınıflandırıcının performans değişimi.	43

TABLolar DİZİNİ

<u>Tablo No:</u>	<u>Sayfa</u>
2.1: 2009 yılına kadar görev / kategori bazında gruplanmış, yayınlanmış makale sayısı.	6
2.2: GÖA ve DÖA stratejilerini temel alan yöntemlerin karşılaştırılması.	11
4.1: Veri kümesindeki öğrencilerin programlara göre dağılımı.	22
4.2: Veri kümesindeki öğrencilerin genel başarı durumlarına göre dağılımı.	22
4.3: Veri kümesindeki öğrencilerin geçmiş dönemde aldıkları derslerin harf notu karşılıkları.	22
5.1: Özellik seçim yöntemleri ile belirlenen özellikler.	39
5.2: Ders içeriklerinden elde edilen anahtar sözcükler.	39
5.3: Çizgesiz veri kümesinde Naive Bayes sınıflandırıcı performans tablosu.	41
5.4: Çizgesiz veri kümesinde K-NN En Yakın Komşuluk yöntemi performans tablosu.	41
5.5: Çizgeler eklenmiş veri kümesinde Naive Bayes sınıflandırıcı performans tablosu.	43
5.6: Çizgeler eklenmiş veri kümesinde k-NN en yakın komşuluk yöntemi performans tablosu.	43
5.7: Çizgeli/Çizgesiz özellik kümeleri için karşılaştırmalı performans tablosu.	44
5.8: 10-kat çapraz gerçeklik için verilerin dağılımı.	45
5.9: 10-kat çapraz gerçeklik uygulandıktan sonra Naive Bayes sınıflandırıcı için elde edilen değerler.	45
5.10: 10-kat çapraz gerçeklik uygulandıktan sonra k-nn sınıflandırıcı için elde edilen değerler.	45

1. GİRİŞ

Farklı disiplinlerden birçok bilim dalının ortak payda da ele aldıkları öğrencinin akademik başarısının modellenmesi konusu kuşkusuz ki öncelikle eğitim bilimleri perspektifinden bakarak değerlendirilmelidir.

Eğitim bilimlerinde akademik başarı, öğrencinin bulunduğu okul, sınıf ve derse göre belirlenmiş sonuçlara ulaşmada göstermiş olduğu ilerleme [1] olarak tanımlansa da çağdaş anlamda başarı kavramının yalnızca akademik değişkenleri değil bilgi ve beceri gibi bilişsel davranışlar ve kişilik özellikleri, kişinin olaylara karşı tutumları gibi bilişsel olmayan davranışlarını da kapsadığı görülmektedir [2].

Başarısızlık kavramı ise daha çok çocuğun ya da gencin uzun süreli (bir eğitim-öğretim döneminden daha uzun süre) hemen her dersten, gelişim düzeyinin ve yeteneklerinin altında bir başarı göstermesi ve bu başarısızlığı bir türlü telafi edememesi durumu olarak kabul edilir [2].

Akademik başarı ve başarısızlıkların belirlenmesinde öğrenciye ait birçok farklı özelliğin etkili olduğu sosyologlar ve eğitim bilimciler tarafından uygulanan çeşitli anketler ve çalışmalarla belirlenebilse de; hangi özelliğin hangi öğrenci grubunda daha çok etkili olduğunu belirlemek, başarılı/başarısız öğrenciler arasında tekrar eden ortak özelliklerin varlığını tespit etmek ve eğer varsa bu tekrar eden özellikler arasında bir ilişki olup olmadığını ortaya çıkarmak veri madenciliğinin çalışma alanı dâhilindedir.

1.1. Tezin Amacı, Katkısı ve İçeriği

Benzer özelliklerde ve ortak koşullarda eğitim gören bir grup öğrencinin akademik başarısının modellenmesi amaçlanan çalışmada, başarıyı yalnızca eğitsel özellikler ve yeteneklerle sınırlandırmak yerine, başarıyı etkileyen diğer faktörleri de bir araya getirip geniş bir bakış açısından bakılması hedeflenmiştir. Öğrencilerin akademik başarı veya başarısızlıkları üzerinde etkisi olduğu literatürde yer alan çalışmalarca belirlenmiş olan demografik özellikler bu faktörlerin ilkidir [3]. Ele alınan bir diğer faktör ise, gerek eğitim bilimlerinde yapılan çalışmalarda gerekse eğitimsel veri madenciliği çalışmalarında üzerinde önemle durulan öğrencilerin geçmiş dönemlere ait akademik performanslarının değerlendirilmesi olmuştur.

Literatürde bu konuya ait yapılan çalışmalar incelendiğinde göze çarpan husus, çalışmalarda yalnızca demografik veriler veya akademik geçmiş veriler üzerine odaklanılmış olmasıdır. Bu tez çalışmasında ise farklı olarak iki farklı veri bir araya getirilmiş ve böylece öğrencinin akademik başarı veya başarısızlığı üzerinde etkili olabilecek farklı bakış açılarını da değerlendirilmiştir.

Öğrencilerin yazılı izni ile toplanan verilerinden özellik seçim yöntemleri ile belirlenen, akademik başarıları üzerinde anlamlı etkileri olduğu tespit edilen demografik özellikleri ve yine hali hazırda eğitim gördükleri kurumdan alınan izinle elde edilmiş olan akademik geçmişlerinde aldıkları derslerin içeriklerinden çıkarılan frekansı en yüksek olan 1-gramlar ile özellik vektörleri oluşturulmuştur.

Öğrencilere ait başarılı ve başarısız olmak üzere iki farklı kategoride toplanan veriler üzerinde öncelikle incelenen birçok çalışmada da kullanılan gözetimli öğrenme yöntemlerinden Naive Bayes ve K-en yakın komşuluk yöntemleri uygulanmıştır. Daha sonraki adımda ise, farklı bir yaklaşım olarak her bir öğrenciye ait çizgeler oluşturularak, çizge madenciliği yöntemlerinden sıklıkla tekrar eden alt çizgelerin tespiti için SUBDUE algoritması kullanılmıştır. SUBDUE algoritması ile başarılı ve başarısız öğrencilere ait sıklıkla tekrar eden örüntüler elde edilmiştir. Elde edilen örüntülerle veri kümesi zenginleştirilmiş ve zenginleştirilen veri kümesi üzerine tekrar Navie Bayes ve K-en yakın komşuluk yöntemleri uygulanarak daha önceki sonuçlar ile karşılaştırılmıştır. Karşılaştırılan sonuçlarda görülmüştür ki; alt çizgelerle zenginleştirilen veri kümesi üzerinde sınıflandırma yöntemlerinin doğru sınıflandırma performansında artış olmuştur.

1.2. Tezin Organizasyonu

Çalışmanın birinci bölümünde akademik başarı ve başarısızlığın tanımı yapılarak, çalışmanın motivasyonu ve özetle yol haritasından bahsedilmiştir. İkinci bölümde, tez konusunun temelini oluşturan eğitimsel veri madenciliği ve çizge tabanlı örüntü madenciliğinden bahsedilerek, literatürde bu alanlarda yapılan çalışmalardan bahsedilmiştir. Ardından üçüncü bölümde problemin tanımı yapılarak, ortaya sunulan hipotez ayrıntılarıyla açıklanmıştır. Dördüncü bölümde tez çalışmasında kullanılan veri kümeleri ve yöntemler tanıtılmış, konu çerçevesinde nasıl uygulandığı anlatılmıştır. Yapılan tüm çalışmaların olumlu ve olumsuz yönleri değerlendirme

bölümünde (5.Bölüm) tartışılmıştır. Son bölümde ise sonuçlar sunulmuş ve gelecek çalışmalar için yol gösterici olması bakımından çeşitli önerilerden bahsedilmiştir.

2. LİTERATÜR ANALİZİ

Bu bölümde tez konusunun da çalışma alanı içine dâhil olduğu ve bilgisayar bilimleri alanında her geçen gün üzerinde daha fazla çalışılan veri madenciliği ve makine öğrenmesi temel başlıkları üzerinde durulacaktır. Akademik başarının modellenmesi üzerine literatürde yer alan çalışmalarda kullanılan makine öğrenmesi metotlarından bahsedilecektir.

2.1. Veri Madenciliği

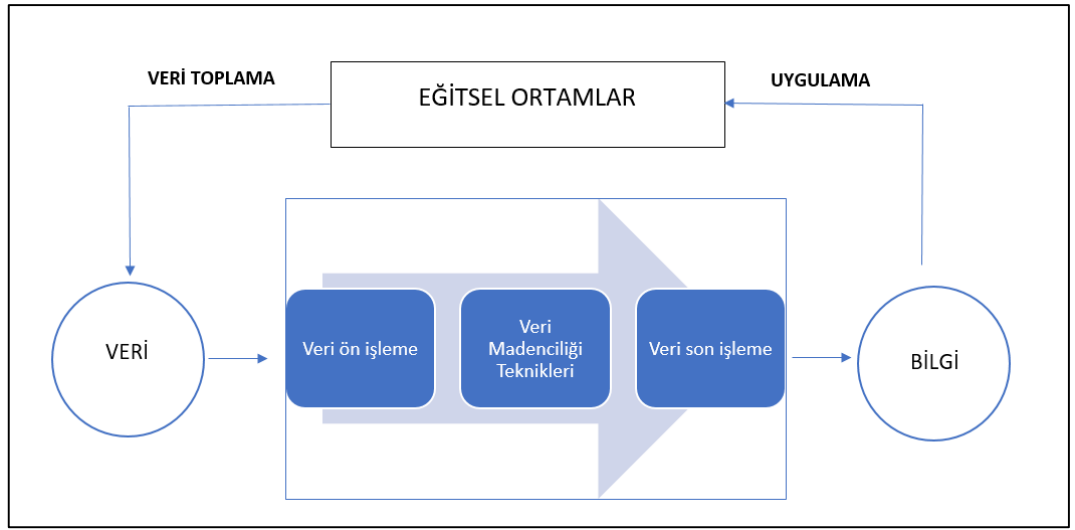
Veri madenciliği, büyük veri tabanlarını taramak, bilgi üretmek ve bilgiyi keşfetmekle görevli bilgisayara dayalı bir bilgi sistemidir [4]. Daha genel bir tanımla, veri madenciliği büyük veri tabanlarından faydalı örüntülerin ve eğilimlerin keşfedilmesi sürecidir [5]. Çeşitli yöntemler sonucu elde edilen bu yapılar bazen bir kurallar kümesi, bazen bir çizge veya bir ağ yapısı, bazen ise bir ağaç yapısı veya bir veya daha fazla denklem bütünü şeklinde karşımıza çıkabilir. Bir veri madenciliği uygulamasında ilginç eğilimlerin ve örüntülerin ortaya çıkarılması için bir veya daha fazla algoritma kullanılır ve kullanılan algoritmalar sonucu elde edilen bu bilgiler, verinin genelleştirilmiş bir modelini sunar [6]. Bu sürecin temel hedefleri; veri kalıplarını bulmak, veriler arasındaki gizli ilişkileri tespit etmek, ilişki kurallarını yapılandırmak, varlıkların sınıflandırılmasını sağlamak, homojen nesne kümeleri oluşturmak ve kolaylıkla üretilemeyen birçok bilgiyi ortaya çıkarmaktır [1].

Veri madenciliği işletmelerden hükümetlere, turizm şirketlerinden süpermarketlere kadar hemen her alana uygulanabilen multidisipliner bir çalışma alanıdır [7].

2.1.1. Eğitimsel Veri Madenciliği

Eğitimsel veri madenciliği (EVM), eğitim ile ilgili uygulamalardaki önemli ve ilginç örüntüleri keşfetmek, öğrencileri ve öğretim süreçlerini daha iyi anlamak için yöntem geliştirilmesini hedefleyen yeni gelişen bir disiplindir [8]. Eğitim yazılımlarının kullanımının artması ve hükümetlerin oluşturdukları veri tabanlarında

öğrencilerin verilerine daha çok yer vermeye başlamasıyla birlikte, öğrencilerin öğrenim süreçlerini yansıtan veriler kolaylıkla ulaşılabilir bir hal almaya başlamıştır [9]. Diğer yandan günümüzde internetin eğitim alanında aktif bir şekilde kullanılması ile birlikte öğretme-öğrenme etkileşimi hakkında çok fazla veri üretilebilen ve her yerden ulaşılabilen sanal öğrenme (e-learning) veya web-tabanlı öğrenme (web-based learning) olarak isimlendirilen yeni eğitim ortamlarının oluşması da eğitimsel veri madenciliğinin ihtiyaç duyduğu eğitimsel verilerin elde edilmesinde büyük kolaylık sağlamıştır [10].



Şekil 2.1: Eğitimsel veri madenciliği süreçleri.

Eğitim ortamlarında veri madenciliği yoluyla çözölen birçok uygulama vardır. Literatürde eğitimsel veri madenciliği alanında yapılan çalışmalar incelendiğinde sıklıkla gözlemlenen uygulama başlıkları aşağıdaki gibidir [11]:

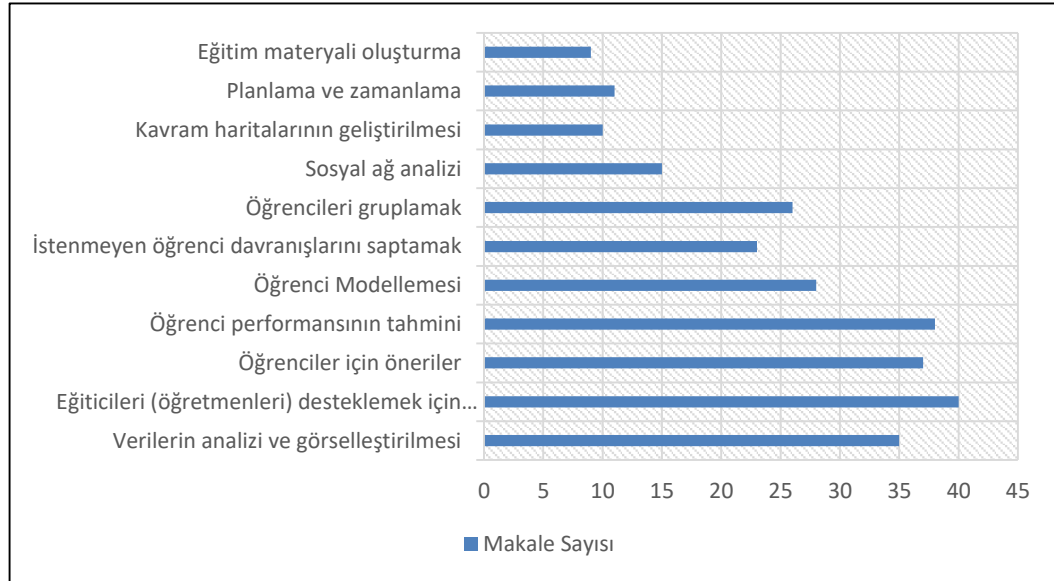
- Verilerin analizi ve görselleştirilmesi
- Eğiticileri (öğretmenleri) desteklemek için geribildirim sağlamak
- Öğrenciler için öneriler
- Öğrenci performansının tahmini
- Öğrenci modellemesi
- İstenmeyen öğrenci davranışlarını saptamak
- Öğrencileri gruplamak
- Sosyal ağ analizi

- Kavram haritalarının geliştirilmesi
- Planlama ve zamanlama
- Eğitim materyali oluşturma

Tablo 2.1’de görüldüğü gibi eğitimsel veri madenciliği alanında yapılan çalışmaların büyük bir kısmı öğrenci performansının tahmin edilmesi ve öğrenci modellemesi üzerinedir. Öğrencilerin eğitim alacakları dönem içerisinde gösterecekleri performansları ve başarılı/başarısız öğrencilerin taşıdıkları ortak özelliklerin bulunması ile öğrencilerin modellenmesi konuları eğitim bilimleri alanında her zaman merak konusu olmuştur.

Bu tez çalışmasının temel motivasyonu da öğrencilerin başarılı veya başarısız olmaları durumunda taşıdıkları benzer özelliklerin tespitinden yola çıkarak akademik başarılarının modellenmesidir.

Tablo 2.1: 2009 yılına kadar görev / kategori bazında gruplanmış, yayınlanmış makale sayısı.



2.1.2. Çizge Tabanlı Veri Madenciliği

Veri madenciliği ile ilgili çoğu yaklaşım bir varlığın nitelikleri arasındaki ilişkileri araştırır, ancak aslında varlıklar arasındaki ilişkiler daha zengin bir bilgi kaynağı ve nihai bilgilerdir. Çizge tabanlı veri madenciliği yaklaşımının da bir parçası olduğu çok ilişkisel veri madenciliği yaklaşımının da (multi-relational data mining) tek

bir veri tablosunda örüntü (pattern) aramak yerine, bir ilişkisel veri tabanından birden fazla tablo (ilişki) içeren desenleri aramak üzerine yoğunlaşmıştır. Çok ilişkisel veri madenciliği için temelde iki farklı yaklaşım geliştirilmiştir. Bu yaklaşımlar; mantık tabanlı yaklaşımlar ve çizge tabanlı yaklaşımlardır [12].

Mantık tabanlı yaklaşımlar tümevarımsal mantık programlama (TMP) alanına girmektedir. TMP, verileri tanımlayabilmek için mantıksal bir teoriyi tetikleme ihtiyacı duyar ve bu işlevi gerçekleştirmek üzere bir dizi tekniği bünyesinde barındırır [13].

Çizge tabanlı veri madenciliği en genel manada, bir çizge olarak temsil edilen verilerin ilişkisel yönlerini incelemek için kullanılan teknikler kümesini temsil eder.

Karmaşık veri kümelerini modellemek için çizgeleri kullanmanın iyi bir çözüm olduğu kimya, bilgisayarlı görme gibi farklı alanlarda çalışan birçok araştırmacı tarafından kabul edilmiştir [14]. Çizge tabanlı veri madenciliğindeki farklı birçok uygulama alanında araştırmacılar, birbirine benzer gözüken fakat birbirinden farklı çözümleri gerektiren problemler üzerinde yoğunlaşarak, farklı yaklaşımlar getirmektedirler [15]. Tez çalışması kapsamında yoğunlaşılacak problem ise çizge veri kümelerinden sıklıkla tekrarlayan alt çizgelerin çıkarılmasıdır.

Bölüm 2.1.2.1’de çizge veri kümelerinde sıklıkla tekrarlayan alt çizgelerin çıkarılması problemi ayrıntılı olarak tanımlanacak ve bu problemin çözümüne yönelik geliştirilen yaklaşımlardan bahsedilecektir.

2.1.3. Çizge Veri Kümesinden Tekrarlayan Alt çizgelerin Çıkarılması

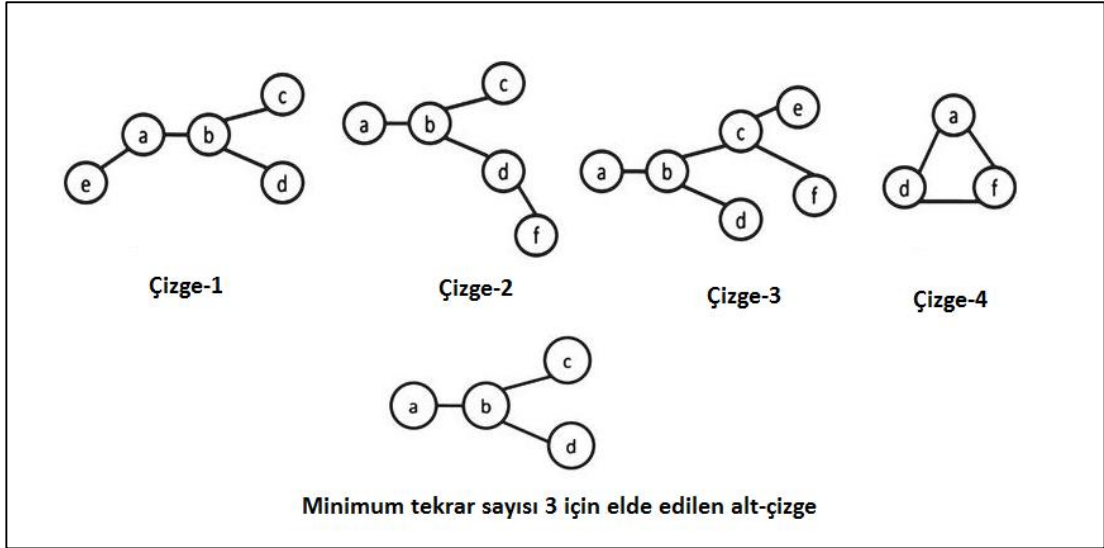
Tekrar eden alt çizgelerin keşfedilmesi problemi; çok sayıda çizge içeren belirli bir çizge veri kümesinde ki, tekrar sayısı belirlenmiş bir eşik değerinin (threshold) üzerinde olan tüm sıklıkla tekrar eden alt çizgelerin elde edilmesine dayanır [16].

Çizgelerden oluşan bir veri kümesi G ve eşik değeri ∂ ($0 < \partial \leq 1$) olmak üzere; bir A çizgesi için destek değeri $A \geq$ destek eşik değeri ise A çizgesi G kümesinde tekrarlayan bir alt çizgedir denir. Burada A ’nın destek değeri destek değeri $_A$ şeklinde ifade edilmiş ve 2.1 numaralı formülde görüldüğü gibi hesaplanmaktadır.

$$\text{destek değeri}_A = \frac{|\{A' \in G \mid A \subseteq A'\}|}{|G|} \quad (2.1)$$

A alt çizgesinin destek değeri, G çizge veri kümesindeki A alt çizgesini içeren çizgelerin tüm çizgelere oranı olarak tanımlanır [15].

Bu problemin çözümünden elde edilen alt-çizgeler veri kümesinin temel özellikleri hakkında önemli bilgiler sunar. Şekil 2.2 de örnek bir çizge veri kümesi ve elde edilen sıklıkla tekrarlayan alt çizge sunulmuştur.



Şekil 2.2: Örnek bir çizge veri kümesinden minimum tekrar sayısı 3 için elde edilen alt çizge.

Tekrarlayan alt çizgelerin bulunması problemindeki kilit konular alt-çizge adaylarının üretimi (candidate generation) ve destek değerinin belirlenmesidir (support counting).

Alt çizge izomorfizm problemi bir çizgenin, farklı bir çizgenin alt çizgesi olup olmadığını inceleyen, çözümü oldukça zorlayıcı ve hesaplama yoğunluğu fazla olan yaklaşımlara ihtiyaç duyan, NP-tam zaman karmaşıklığına sahip bir problemdir. Düğüm sayısı çok olan çizgeleri içeren veri kümelerinde bu problemin çözümünün zor ve uzun zaman alıcıdır. Bu sebeple araştırmacılar tarafından aday üretimi konusunda çok fazla sayıda farklı yaklaşım geliştirilmiştir [14],[16].

Aday üretimi konusunda literatürde temel alınan iki farklı yaklaşım vardır. İlk yaklaşım Apriori tabanlı yaklaşımlarda kullanılan, alt çizgelere sahip küçük çizgelerin birleştirilmesiyle ağaç yapısının oluşturulması yöntemidir. Bu yöntemde öncelikle ortak çizgelere sahip çizgeler belirlenerek, hangi çizgelerin birleştirileceğinin tespit edilmesi gerekmektedir. Diğer yaklaşım ise g-Span, Subdue benzeri yaklaşımlarda kullanılan en alt seviye çizgelerden başlanarak her seferinde bir kenar eklenip

geniştirilerek aday çizgelerin üretilmesi yöntemidir. Bu yöntemin dezavantajı, çizge veri kümesinde bulunmayan gereksiz çizgelerin de oluşmasına olanak sağlamasıdır. İşlem sonunda bu gereksiz çizgelerin destek değerlerinin de hesaplanması işlem süresini arttırmaktadır.

Destek değerinin belirlenmesi problemi ise zaman karmaşıklığı NP-tam olması sebebiyle araştırmacıların üzerinde birçok çalışma gerçekleştirdiği ve farklı birçok çözüm sundukları bir sorundur.

Destek değeri elde edilen yeni bir çizgenin, çizge veri kümesinde geçme oranıdır. Bu problemin çözümüne getirilen yaklaşımlardan ilki, Apriori tabanlı yaklaşımlarda budama mantığı ile birleştirilecek iki ana çizgenin destek değeri eşik değerinden düşük ise onların birleşiminden oluşacak çizge yapıya eklenmemesidir. Gaston, FFSM benzeri yöntemlerde kullanılan yaklaşım ise gömülü listelerin kullanılmasıdır. Bu yaklaşımda, alt çizgenin kenar ve düğümleri ile onu içeren çizgenin içindeki ona karşılık gelen kenar ve düğümler arasında bir eşleşme gerçekleştirilir. Böylelikle eşik değeri hesaplanırken yalnızca yeni eklenen kenar ve düğümleri aramak yeterli olmaktadır. Zaman olarak avantaj sağlayan bu yaklaşımda, büyük bir hafıza gereksinimi vardır. Literatürde yer alan bir diğer yaklaşımda ise iki ana alt çizgenin birleşiminden oluşturulan çocuk çizgenin destek değeri hesaplanırken yalnızca birleştirilen ana çizgelerin destek değerlerine bakılır. Böylelikle arama uzayı daraltılarak zamandan kazanç sağlanır. Fakat bir önceki yaklaşımda olduğu gibi alt çizgelerin geçtiği tüm çizgelerin hafızada tutulmasının gerektiğinden dolayı büyük bir hafıza gereksinimi mevcuttur. Subdue yönteminde ise farklı bir yaklaşım olarak Minimum Tanımlama Uzunluğu (MTU) kodu ve arka plan bilgisi kullanılır. Elde edilen değer ile en iyi alt çizge bulunur ve bu alt çizge ile veri kümesi sıkıştırılır.

Bahsi geçen bu iki kilit konunun çözümüne çok sayıda farklı yaklaşım geliştirilmiştir. Geliştirilen her bir yöntem farklı bir bakış açısından problemin değerlendirilmesine olanak sağlamıştır. Alt çizge madenciliği yöntemlerini birbiri arasında kıyaslayabilmek ve farklılıklarını/benzerliklerini daha iyi görebilmek için arama stratejisine göre, girdi niteliğine göre ve çıktının tamlığına göre kategorilere ayırmak mümkündür [22].

- Arama stratejisine göre: Tekrarlayan alt çizge madenciliği yöntemlerinde kullanılan 2 farklı arama stratejisi vardır. Bu stratejiler Genişlik öncelikli arama (Breadth First Search-GÖA) ve Derinlik öncelikli arama (Depth First

Search-DÖA)’dır. GÖA stratejisi seviye bazında (belirli bir seviyedeki tüm düğümleri ele alarak) bir arama gerçekleştirdiği için, zaman ve hafıza bakımından yeterli şartlar sağlandığında asgari eşik değerinin üzerindeki tüm tekrarlayan alt çizgelerin bulunmasını garanti eder. Fakat DÖA stratejisi, GÖA ile karşılaştırıldığında daha az hafıza kullanarak bir arama gerçekleştirdiğinden dolayı, özellikle son yapılan çalışmalarda daha fazla tercih edilir hale gelmiştir. DÖA stratejisinin dezavantajı ise çok fazla alt-çizge üretilmesine sebep olmasıdır [22]. Literatürde GÖA stratejisini temel alan ve en sık bahsedilen yöntemler; SUBDUE [12, 15], FARMER [23], HSIGRAM [24] ve FSG’dir [14]. DFS stratejisi ile geliştirilen yöntemlerden bazıları ise; gSpan [25], CloseGraph [26], FFSM [27], Gaston [28], ISG [29] ve GREW [30] ‘dir. Tablo 2.2.’ de GÖA ve DÖA stratejilerini temel alan ve literatürde sıklıkla bahsedilen 10 algoritmanın karşılaştırma tablosu görülmektedir [22].

- Girdi niteliğine göre: Tekrarlayan alt çizgelerin bulunması probleminde sosyal medya ağlarında olduğu gibi tek bir büyük çizge yapısı üzerinde veya genellikle kimyasal maddelerin yapılarını gösterirken kullanılan boyut olarak küçük fakat birçok çizgeden oluşan bir çizge kümesi üzerinde çalışılabilecek olabilir. Yöntemin kullanacağı girdinin niteliği, tek bir büyük çizge yapısı veya çok sayıda küçük boyutlu çizge kümesi olabilir. Tablo 2.2.’ de özellikleri belirtilen yöntemlerden tek bir büyük çizge yapısını girdi olarak kabul eden yöntemler; HISGRAM, GREW’dir. FARMER, FSG, gSpan, CloseGraph, FFSM, Gaston ve ISG yöntemleri ise çok sayıda küçük boyutlu çizgelerden oluşan çizge veri kümesini girdi olarak kabul ederler. SUBDUE yöntemi ise her iki çizge kümesi üzerinde de çalışabilmektedir.

Tablo 2.2: GÖA ve DÖA stratejilerini temel alan yöntemlerin karşılaştırılması.

ALGORİTMA	ÇİZGE GÖSTERİMİ	ALT-ÇİZGE ÜRETİMİ	ALGORİTMİK YAKLAŞIM	FREKANS ÖLÇÜMÜ
SUBDUE	Komşuluk matrisi	Seviyeli arama	Örüntü genişletme	Minimum tanımlama uzunluğu kodu ve arka plan bilgisi
FARMER	Ağaç yapısı	Seviyeli arama	MVP ve Apriori-tabanlı	Ağaç veri yapısı
FSG	Komşuluk listesi	Bir kenar genişletme	Apriori-tabanlı	İşlem tanımlayıcı listeler
GSPAN	Komşuluk listesi	En sağ genişletme	Örüntü genişletme	DÖA sözlüksel sıralama
CLOSEGRAPH	Komşuluk listesi	En sağ genişletme	Örüntü genişletme	DÖA sözlüksel sıralama
HISGRAM	Komşuluk matrisi	Yinelemeli birleştirme	Apriori-tabanlı	Maksimal bağımsız küme
GREW	Seyrek ağaç gösterimi	Yinelemeli birleştirme	Örüntü genişletme	Maksimal bağımsız küme
FFSM	Komşuluk matrisi	Birleştirme ve genişletme	Apriori-tabanlı	En uygun optimal kanonik bitişiklik matrisi ağacı
GASTON	Hash tablosu	Genişletme	Örüntü genişletme	Gömülü listeler
ISG	Üçlü kenar (Edge triplet)	Üçlü kenar genişletme	Apriori-tabanlı	İşlem tanımlayıcı listeler

- Çıktının tamlığına göre: Kullanılan yöntemler, çıktı olarak ürettikleri sıklıkla tekrar eden alt çizgelerin tamamının bulunması veya bulunmaması özelliğine göre kategorize edilebilirler. Bu iki yaklaşımın farkı aslında arama metotlarından gelmektedir. Burada iki farklı arama metodundan bahsetmek mümkündür. İlk arama metodu olan kesin olmayan arama yaklaşımında, iki çizgenin benzerliğini karşılaştırmak için yaklaşık bir ölçü kullanılır, diğer bir deyişle destek değerinin artması için karşılaştırılan iki alt çizgenin birebir aynı olması gerekmez. Bu sebeple de bu yaklaşımı temel edinen algoritmalar tüm tekrarlayan alt-çizgeleri bulabilmek için kesin bir yol değildir. Bu yaklaşımı temel edinen ve literatürde en sık bahsedilen algoritma SUBDUE'dur [12]. SUBDUE haricinde bu yaklaşımı benimsemiş diğer algoritmalar ise; GREW, CloseGraph ve ISG' dir. Diğer arama metodu ise tam-kesin arama yaklaşımıdır. Kesin olmayan arama yaklaşımından farklı olarak verilen çizge kümesi içerisindeki tekrarlayan alt çizgeleri bulmayı garanti eder. Bu yaklaşımın kesinlik olgusu alt-izomorfizm karşılaştırmalarının kapsamlı olmasını gerektirir ve bu sebeple önemli bir hesaplama maliyeti ile sonuçlanır [19]. Literatür de yapılan araştırmalarda görüldüğü kadarıyla tam-kesin arama yaklaşımı ile oluşturulan yöntemler kesin olmayan arama yaklaşımıyla oluşturulan yöntemlere göre daha yaygındır. Bu yaklaşımı temel alan bazı yöntemler; FARMER, gSpan, FFSM, Gaston, FSG, HSIGRAM' dir.

Tablo 2.2’de genel özellikleriyle karşılaştırılan 10 farklı yaklaşım temel özelliklerine göre 3 ayrı kategoride incelenmiştir. İncelenen bu yaklaşımlardan literatürde en sık kullanılanlar hakkında genel bilgi devam eden bölümde verilmiştir.

- FSG: Tek sınıflı uygulamalar için geliştirilmiş olan ve kanonik çizge gösterimlerinden yararlanılan bu yaklaşımda, her bir alt çizge için bitişiklik matrisinin kenar ve düğümlerinin tüm kombinasyonları oluşturularak tek bir tanım yaratılmaya çalışılır. Arama uzayını daraltmak amacıyla Apriori yaklaşımına benzer aday alt çizge üretme ve destek değerine göre arama uzayını daraltma yöntemini uygular.
- gSpan: Tekrarlayan alt çizgelerin bulunması probleminde derinlik öncelikli arama yaklaşımını ilk kez kullanan ve bu yaklaşımı sözlüksel sıra yöntemleri ile birleştiren bir yöntemdir. GSPAN aday alt-çizge üretme ve yanlış pozitif budama olmaksızın tüm tekrarlayan alt-çizgeleri bulmaya çalışır.
- Gaston: Gaston, birden fazla çizgeyi girdi olarak kabul eden DÖA tabanlı bir yöntemdir. SUBDUE, gSpan, CloseGraph, GREW yaklaşımlarına benzer olarak örüntü büyütme yaklaşımını temel alarak ilerler. Fakat hızlı-başlama prensibini (quick start principle) benimsediğinden ötürü bu yaklaşımlara göre daha hızlıdır. Bu yaklaşımda izomorfik çizgelerin oluşumun önlenmesi için, eklenen düğüm ve kenarların sırasını gösteren ‘çizge kodu’ olarak isimlendirilen bir karakter dizisi kullanılır.
- FFSM: FSG ve HISTOGRAM yöntemlerine benzer bir şekilde Apriori tabanlı bir metottur. FFSM tekrar etme sıklığını hesaplayabilmek için yaklaşık-optimal (sub-optimal) kanonik bitişiklik matrisi kullanan, DÖA yaklaşımını benimsemiş bir yöntemdir. Gaston yaklaşımıyla benzer olarak birden fazla çizgeyi girdi olarak kabul eder.

Tez çalışması kapsamında tekrarlayan alt-çizgelerin bulunması adımımda yukarıda genel bilgi verilen yaklaşımlardan farklı olarak SUBDUE yönteminden yararlanılmıştır ve Bölüm 4’de yöntemin çalışma mantığından ayrıntılarıyla bahsedilecektir.

2.2. Makine Öğrenmesi

Makine Öğrenmesi (Machine Learning-ML), insanlardaki öğrenme yapısının bilgisayara taşınmasını hedefleyen, bilgisayarların ‘öğrenme’ işlemini gerçekleştirebilmeleri için gerekli algoritma ve tekniklerin gelişimi ile ilgilenen, yapay zekanın bir alt dalıdır.

Son yıllarda eğitim alanında yapılan çalışmalarda makine öğrenmesi metotları çok dikkat çekmektedir. Literatürdeki çalışmaların önemli bir kısmının öğrencilerin dersleri tamamlarken ve problemlerini çözerken gösterdikleri performansları öngörmeye odaklandığı görülmektedir [31]. Karar ağaçları [32]-[36], yapay sinir ağları [37]-[39], destek vektör makineleri [40],[41],[52],[53], Naive Bayes sınıflandırıcı [41]-[45], K-en yakın komşuluk yöntemi [44]-[46] gibi birçok makine öğrenmesi metodu öğrencilerin başarılarının tahmin edilmesi ve öğrenci başarısının modellenmesi konusunda kullanılmaktadır.

2.3. Sınıflandırma Metotları

Sınıflandırma, veri kümesindeki farklı sınıflar için modeller oluşturarak sınıfı bilinmeyen bir veri parçasını bilinen bir gruba yerleştirmekte kullanılan makine öğrenmesinin popüler ve temel uygulama alanlarından biridir [47].

Bu bölümde literatürde öğrenci modellemesi ve öğrenci başarısının tahmin edilmesi hususlarında literatürde kullanılan sınıflandırma metotları öncelikle tanıtılacak, daha sonra da literatürdeki çalışmalardan bahsedilecektir.

2.3.1. Naive Bayes Sınıflandırıcı

Naive Bayes sınıflandırıcı, Bayes Teoremi’nden yararlanılarak oluşturulmuş, kolay uygulanabilir ve anlaşılabilir istatistiksel bir sınıflandırma yöntemidir. Özellik vektörü ile tanımlanmış bir örneğin verilen sınıflardan ait olma olasılığı en yüksek olan sınıfa atanmasını sağlar [48].

Metin sınıflandırması, tıbbi teşhis ve sistem performans yönetimi de dâhil olmak üzere birçok pratik uygulamada etkili olduğunu kanıtlanmış olan Naive Bayes

sınıflandırıcı [49], eğitimsel veri madenciliği alanında pratik uygulamalarda da sıklıkla kullanılmıştır.

Öğrencilerin modellenmesi, akademik başarılarının tahmini gibi eğitimsel veri madenciliği kapsamına giren konularda araştırmacılar tarafından gerçekleştirilmiş çalışmalardan bazıları aşağıda listelenmiştir.

- 2009 yılında Memphins Üniversitesi araştırmacıları tarafından MetaTutor isimli akıllı bir dersane sistemi ile etkileşimde bulunan öğrencilerin zihinsel modellerini (düşük, orta, yüksek) anlamak üzere, Naive Bayes sınıflandırıcı ve 5 farklı sınıflandırıcı algoritma uygulanarak öğrencilerin zihinsel modelleri tahmin edilmeye çalışılmıştır [41].
- Hırvatistan’da gerçekleştirilen bir diğer çalışmada ise Hırvatistan Yükseköğretim Bilgi Sistemi’nden elde edilen öğrencilere ait demografik veriler ve geçmiş akademik bilgiler içeren bir veri koleksiyonu kullanılmıştır. Çalışmanın amacı ise öğrenci başarısının tahmininde demografik özelliklerin etkisinin belirlenmesidir. Bu amaç doğrultusunda Naive Bayes ve Destek Vektör Makinesi (Support Vector Machine) olmak üzere 2 farklı sınıflandırıcı kullanılmıştır. Çalışma sonucunda akademik başarı üzerinde etkili olan 3 özelliğin öğrencilere ait geçmiş dönem ortalamaları, daha önce tamamlamış oldukları okullar ve öğrencilik hakları olduğu görülmüştür [42].
- 2010-2011 eğitim-öğretim yılında Tuzla Üniversitesi İktisat Fakültesi birinci sınıf öğrencilerinin kayıt esnasında verdikleri bilgiler, yaz döneminde bu öğrencilere uygulanan anketlerden hareketle 3 farklı sınıflandırıcı metodu kullanarak öğrencilerin başarı durumları değerlendirilmeye çalışılmıştır. Öğrencilerin başarı durumları uygulanan geçme sınavındaki geçer not ile değerlendirilmiştir. Öğrencilere ait çeşitli sosyo-demografik verilerin, giriş sınavındaki gösterdikleri performansın ve başarıyı etkileyebilecek öğrenime yönelik tutumlarına ait verilerin değerlendirilmeye alındığı bu çalışmada Naive Bayes sınıflandırıcı, karar ağaçları, çok katmanlı algılayıcı (MLP) makine öğrenmesi metotlarından yararlanılmıştır. Elde edilen sonuçlarda ilgili üç sınıflandırma metodundan en iyi sonuç veren Naive Bayes yöntemi olmuştur [43].

- Bulgaristan’da eğitim-öğretim veren bir üniversitenin çalışanları tarafından gerçekleştirilen ve asıl amacı üniversite yönetimine veri madenciliğinin eğitim alanında aktif bir şekilde kullanılabileceğini göstermek amaçlı gerçekleştirilen bu çalışmada, farklı makine öğrenmesi metotları ile öğrencilerin başarılarının tahmin edilmeye çalışılmıştır. 10330 örneklem ve her bir örneklem için 14 nitelik ile veri seti oluşturulmuştur. Bu nitelikler içerisinde öğrencilere ait cinsiyet ve yaş gibi demografik verilerin haricinde üniversite öncesi özellikleri de yer almaktadır. Öğrencilerin daha önce eğitim aldıkları ortaokul profili ve lokasyonu, son orta öğretim başarı puanı, kabul sınavı başarı puanı gibi geçmiş eğitim verilerinden de yararlanılmıştır. Veriler üzerine karar ağacı, Naive Bayes sınıflandırıcı, Bayes ağları, k-En yakın komşuluk yöntemi ve OneR ve JRip olmak üzere iki farklı kural öğrenici metod uygulanarak öğrencileri 5 farklı sınıfa ayırmak hedeflenmiştir. Bu sınıflar “mükemmel”, “çok iyi”, “iyi”, “orta”, “kötü” dür. Çalışma sonunda en başarılı sınıflandırma metodunun karar ağacı olduğu görülmüştür [44].

Yukarıda listelenen birçok araştırmada da olduğu gibi Naive Bayes sınıflandırıcı eğitimsel veri madenciliğinin çözüm bulmaya çalıştığı en önemli problemlerden olan öğrencinin akademik başarısının tahmini veya öğrencinin başarı/başarısızlık durumunun modellenmesi gibi konularda sıklıkla başvurulmuş bir yöntemdir. Tez çalışması kapsamında da kullanılan makine öğrenmesi metotlarından biridir. Naive Bayes sınıflandırıcının temelini oluşturan Bayes teoremi ve sınıflandırıcının çalışma mantığı ayrıntılarıyla Bölüm 4’de anlatılacaktır.

2.3.2. K-En Yakın Komşuluk Yöntemi (k-NN)

K-En Yakın Komşuluk Yöntemi (k-NN) diğer sınıflandırma yöntemlerinden farklı olarak uzaklık hesaplamasına dayalı geliştirilen gözetimli öğrenme yöntemidir. Bu yöntemde ilk adım olarak, sınıf değeri bulunmak istenen yeni örneğin veri setinde bulunan tüm örneklerle olan uzaklığı hesaplanır. Daha sonra uzaklıklar büyükten küçüğe sıralanarak, k en küçük uzaklık değerine sahip gözlemler ele alınır. Çoğunluk oylaması veya ağırlıklı oylama yöntemlerinden biri ile, k gözlemin içinde en sık tekrar eden sınıf tahmini sınıf değeri olarak bulunur.

K-En Yakın Komşuluk Yöntemi literatürde çok sıklıkla kullanılan bir makine öğrenmesi metodudur. Tez kapsamında da kullanılan bu yöntem ile ilgili tanımlar ve yöntemin uygulanışı Bölüm 4’de verilecektir.

Devam eden bölümde ise öğrenci başarısının modellenmesi kapsamında k-NN yöntemi kullanılarak yapılmış olan birkaç çalışmadan bahsedilecektir.

- 2010 yılında Hellenic Open Üniversitesi’nde 3 farklı sınıflandırıcıyı bir araya getirerek uzaktan eğitim öğrencilerinin performanslarını tahmin etmek üzerine yoğunlaşan bir yaklaşım geliştirilmiştir. Birebir etkileşimde bulunulamayan uzaktan eğitim öğrencilerinin performanslarını tahmin etmek, geleneksel sınıf içi eğitim alan öğrencilerin performanslarını tahmin etmekten daha zorlu bir süreç olduğu için bu çalışmada Naive Bayes, k-NN ve WINNOW algoritması bir araya getirilmiştir ve kompleks bir sınıflandırma metodu oluşturulmuştur. Geliştirilen bu metotla tahmin başarısının %73’ den %82 oranına yükseldiği gözlemlenmiştir [45].
- 12 dersten oluşan çevrimiçi dokunmatik ekranda yazı yazma kursundaki öğrencilerin final sınavındaki geçme/kalma potansiyellerinin ve alacakları başarı puanının tahmin edilmesine yönelik çalışmada sınıflandırma metodu olarak k-NN kullanılmıştır. Bu çalışmadaki amaç eğiticilerin öğrencilerinin başarı durumlarını önceden tahmin ederek, o öğrencilerin zorlandığı konulara daha çok odaklanarak başarı oranını yükseltmektir. Yapılan çalışmada başarı puanının doğru tahmin edilme oranı %85’in üzerinde tespit edilmiştir [46].

2.3.3. Destek Vektör Makineleri

Geleneksel istatistik yaklaşımında kullanılacak doğru modelin formunun başlangıçta bilindiği varsayımıyla yola çıkılarak bu modelin parametrelerini belirleme amacı güdülür. İstatistiksel öğrenme teorisinde ise klasik istatistik yaklaşımından farklı olarak, modelin formu başlangıçta bilinmeksizin olası doğru modeller arasındaki en iyi modelin bulunması amacı vardır.

Destek Vektör Makinesi (DVM), yapısal riski en aza indirme ilkesini (structural risk minimisation principle) benimseyen, asıl amacı çok boyutlu veriyi bir hiper

düzlem yardımıyla tamamen iki ayrı sınıfa ayırmak olan, gözetimli öğrenme yöntemi kapsamına dahil olan bir istatistiksel öğrenme teorisi yöntemidir [50][51].

Bu yaklaşım başlangıçta yalnızca iki sınıflı doğrusal verilerin sınıflandırılması problemine çözüm üretebilmek için geliştirilse de, zamanla doğrusal olmayan ve çok sınıflı verilerin sınıflandırılması ihtiyacını karşılamak üzere bazı stratejileri benimsemiştir.

DVM'nin diğer öğrenme metotlarından üstünlüğü ise, bir sistemin bilinmeyen veri üzerindeki başarı olasılığı olarak tanımlanan VC boyutunu hesaplayabilmesidir.

Eğitimsel veri madenciliği alanında sıklıkla üzerinde çalışılan öğrencinin akademik başarısının tahmini [40], öğrenci modellenmesi [41], beklenmedik öğrenci davranışlarının tespiti [52] ve öğrencilerin gruplanması [53] gibi temel problemlerin çözümü için destek vektör makinelerinden yararlanılmaktadır.

2.3.4. Karar Ağaçları

Karar ağaçları yapılandırılması, yorumlanması ve anlaşılması diğer yöntemlere kıyasla daha kolay, özellikle model üretme konusunda oldukça başarılı ve sınıflandırma problemlerinde sıklıkla tercih edilen görsel bir modelleme yöntemidir.

Karar ağaçları yönteminde sınıflandırma yapabilmek için karşılaşılan probleme ilişkin sahip olunan büyük miktardaki veriler çok küçük parçalara bölünerek kullanılır ve karar seçenekleri ile olasılıklara bağlı durumlar sıralanarak ortaya bir ağaç yapısı çıkarılır. Oluşturulan bu ağaç yapısı üzerindeki kurallar dizisine göre yeni gelen bir verinin sınıfı tahmin edilebilir.

Karar ağaçları, üzerinde çalışılan problem için hangi faktörlerin etken olduğunu ve bu faktörlerin birbirleri arasındaki ilişkilerini ortaya koyan, çeşitli kurallar ve bu kuralların sonuçlarını içeren hiyerarşik yapıdaki bir yöntemdir.

Büyük veri setleri üzerine kolay uygulanabilmesi sebebiyle farklı alanlarda birçok uygulamada kullanılan karar ağaçları eğitimsel veri madenciliği alanında da, eğitimcileri desteklemek için geribildirim sağlamak [54], kişiye özel ve uyarlanabilir eğitim desteği vermek [55], öğrencilerin öğrenme süreçlerini kolaylaştırmak ve öğrenme sonuçlarını en üst düzeye çıkarmak için optimal öğrenme süreçleri önermek [56], web tabanlı öğrenme sistemlerinde öğrenci performansını tahmin etmek [57], öğrencilerin memnuniyetlerini tahmin etmek [58], öğrencilerin eğitimi terk etme

olasılıklarını tahmin etmek [59], öğrencileri başarısızlık risklerine göre gruplandırmak [60], vb. birçok problemin çözümü için araştırmacılar tarafından tercih edilmiş bir makine öğrenmesi metodudur.

2.3.5. Yapay Sinir Ağları

Yapay zekâ teknolojileri insanı insan yapan en önemli yetilerinden olan düşünme, değerlendirme, öğrenme ve bir durum veya problem karşısında çözüm üretme gibi zekâ gerektiren faaliyetlerin makinelerce herhangi bir destek olmaksızın taklit edilerek çözümler üretilmesini hedeflemektedir. Bu hedef doğrultusunda geliştirilen yapay sinir ağları da, sinir hücrelerinin biyolojik yapısından yola çıkılarak tasarlanmış bir tahmin yöntemidir.

Yapay sinir ağları, insan beyni tarafından gerçekleştirilebilen etrafındaki materyallerden ve durumlardan yeni bilgiler öğrenme, öğrenmenin sonucu olarak yeni bilgiler üretebilme, üretilen bu bilgiler arasında ilişkiler kurabilme gibi eşsiz kabiliyetlerin otomatik olarak gerçekleştirilmesini sağlayan bilgisayar sistemleridir.

Biyolojik sinir ağları temel alınarak geliştirilen yapay sinir ağları, bir giriş verisi vektörü ile başlar ve bir çıkış verisi vektörü ile sona erer. Giriş vektörü ile çıkış vektörü arasındaki ilişkinin belirlenmesi yapay sinir ağları mühendisliğinin konusu dâhilindedir. Belirlenen bu ilişkinin doğrusal olma zorunluluğu olmadığı gibi çıkış değerleri belli olmayan sistem girişleri ile de uyumlu olarak çalışması sebebiyle karmaşık sorunlara bile iyi çözümler sunabilmektedirler.

Öğrencilerin sene sonu notlarının tahmini [40], çevrimiçi eğitim ortamlarında öğrenim gören öğrencilerin eğitime devam etme/etmeme tercihlerinin tespiti [61], öğrencilerin akademik performanslarının tahmini [37],[38],[62] kişiye özel ve uyarlanabilir eğitim içeriğinin tespiti [39] gibi bazı eğitimsel veri madenciliği konularında yapay sinir ağları yöntemlerinden yararlanılmıştır.

3. PROBLEMİN TANIMI VE HİPOTEZ

Bu bölümde tez çalışmasına konu olan problemin tanımı yapılmış ve ardından ortaya konulan hipotez açıklanmıştır.

Tez çalışması kapsamında öğrencilerin akademik başarılarının modellenmesinde literatürde incelenen birçok çalışmada kullanılan- sosyo-demografik veriler ve geçmiş akademik performans verilerinden farklı olarak tekrarlayan çizge madenciliği yöntemi ile elde edilmiş olan, öğrencilerin akademik başarı ve başarısızlıklarının karakteristik özelliklerini ifade eden örüntülerle zenginleştirilmiş bir veri kümesi üzerinde geleneksel makine öğrenmesi metotları uygulanmış ve bu yaklaşımla makine öğrenmesi metotlarının doğru sınıflandırma performanslarının artırılması hedeflenmiştir.

Nedenlere, süreçlere ve kişilere bağlı olarak değişiklik gösterebilen akademik başarı ve başarısızlık kavramlarının tek bir kalıpla ifade edilmesi oldukça zordur. Bu sebeple öğrencilere ait olan ve öğrencilerin akademik başarılarının modellenmesi konusunda yapılmış olan geçmiş çalışmalarca da benimsenmiş olan sosyo-demografik ve geçmiş akademik performans verilerinden yola çıkarak başarılı veya başarısız bireyler arasında sürekli olarak tekrar edebilecek özelliklerin tespit edilebilmesi üzerinde durulmuştur. Tespit edilen bu özelliklerin veri kümesini zenginleştirebileceği ve zenginleştirilmiş bu veri üzerinde yapılacak modellemenin güvenilirliğinin daha yüksek olabileceği varsayımı üzerinde durulmuştur.

Tez çalışmasına konu olan probleme genel bir çerçeveden bakıldığında asıl hedefin öğrencilerin akademik başarılarının modellenmesi olduğu görülmektedir. Bölüm 2.1.1 de genel bir tanımı verilmiş olan Eğitimsel Veri Madenciliği' nin konusu olan bu problemin çözümü için diğer çalışmalardan farklı olarak Çizge Tabanlı Veri Madenciliğinin bir alt dalı olan Tekrarlayan Alt çizge Madenciliği ve geleneksel Makine Öğrenmesi metotları bir araya getirilerek kapsamlı bir bakış açısı elde edilmiştir.

Tez çalışması oluşturulurken ortaya konulan hipotez ise aşağıdaki gibi açıklanabilir.

Hipotez 1) Veri sayısı az olan bir veri kümesi üzerinde akademik başarının modellenmesi konusunda tekrarlayan alt çizge madenciliği metodu ile zenginleştirilen

özellik vektörleri geleneksel makine öğrenmesi metotlarının doğru sınıflandırma performanslarını artırabilmektedir.

4. YÖNTEM

Bu bölümde tez çalışması kapsamında kullanılan yöntemler açıklanacaktır. Bu çalışmada kullanılan öğrencilere ait gerçek veriler ve bu verilerden tekrarlayan alt çizge madenciliği yöntemi ile elde edilen alt çizgelerin eklenmesi ile oluşturulan veri kümesi ile ilgili bilgiler Bölüm 4.1’de anlatılmıştır.

Veri kümesi oluşturulurken uygulanan temel adımlar olan özellik seçim yöntemlerinin uygulanması, ders içeriklerinden anahtar sözcüklerin çıkarılması, başarılı/başarısız öğrencilere ait çizge yapılarının oluşturulması, oluşturulan çizgelerden tekrarlayan alt çizge madenciliği yöntemi olan SUBDUE ile örüntülerin elde edilmesi ve elde edilen örüntülerle zenginleştirilen nihai özellik vektörünün oluşturulması yaklaşımları ise Bölüm 4.1’in alt başlıkları olarak devam eden bölümde yer almaktadır.

Son olarak geleneksel sınıflandırma metotlarından Naive Bayes ve K-en yakın komşuluk yöntemi ile ilgili gerekli tanımlar ve oluşturulan veri kümesi üzerine uygulanaşına Bölüm 4.2’nin alt başlıklarında yer verilmiştir.

4.1. Veri Kümesinin Oluşturulması

Öğrencinin akademik başarısının modellenmesi amaçlanan bu tez çalışmasında öğrenci grubu olarak İstanbul ilinde bulunan bir vakıf üniversitesi olan İstanbul Aydın Üniversitesi Anadolu Bil Meslek Yüksekokulu Bilgisayar Teknolojisi Bölümü programlarına kayıtlı olan öğrenciler tercih edilmiştir.

Çalışmada kullanılan verilerin elde edildiği öğrenci grubu ortak bir merkezi sınavla birbirine oransal olarak çok yakın yüzdelik dilimle benzer bölümlere kayıt yaptırmış olan ve aynı alanda mesleki eğitim alan bir grup öğrenci olacak şekilde belirlenmiştir.

Bu çalışmada önerilen yaklaşım, Bilgisayar Teknolojisi Bölümü programlarında kayıtlı olan öğrencilerin sosyo-demografik özellikleri, aile yapısı, akademik geçmişi gibi birçok bilgisinden oluşturulan gerçek bir veri kümesi üzerine uygulanmıştır.

Veri kümesinde bilgileri bulunan toplam 142 öğrencinin eğitim aldıkları programlar ve programlara göre sayısal olarak dağılımları Tablo 4.1’de, öğrencilerin geçmiş dönemlerde aldıkları derslere bağlı olarak genel başarı durumlarının sayısal

dağılımları ise Tablo 4.2.'de gösterilmiştir. Başarılı öğrenciler Pozitif (XP), başarısız öğrenciler ise Negatif (XN) olarak etiketlenmiştir.

Tablo 4.1: Veri kümesindeki öğrencilerin programlara göre dağılımı.

Program İsmi	Eğitim Dili	Öğrenci Sayısı	Toplam
Bilgisayar Programcılığı	Türkçe	83	83
Bilgisayar Programcılığı	İngilizce	49	59
Bilgisayar Teknolojisi	İngilizce	10	
TOPLAM			142

Tablo 4.2: Veri kümesindeki öğrencilerin genel başarı durumlarına göre dağılımı.

Negatif (XN)	53
Pozitif (XP)	89
	142

Öğrencilerin geçmiş dönemlerde aldıkları derslerin harf notları Tablo 4.3'de gösterildiği gibi değerlendirilmiştir. Öğrenci tarafından geçmiş dönemlerde alınmış olan dersin uluslararası kredi sistemi olan AKTS karşılığına göre ağırlıklı olarak genel başarı durumuna etkisi hesaplanmıştır.

Tablo 4.3: Veri kümesindeki öğrencilerin geçmiş dönemde aldıkları derslerin harf notu karşılıkları.

Harf notu	Puan	Katsayı	Açıklaması	Genel Başarı Durumu
AA	90-100	4.0	Mükemmel	XP
BA	85-89	3.5	Pekiyi	
BB	80-84	3.0	İyi	
CB	75-79	2.5	Yeterli	
CC	70-74	2.0	Geçer	
DC	60-69	1.5	Koşullu Başarılı/Başarısız	XN
DD	50-59	1.0	Koşullu Başarılı/Başarısız	
FD	40-49	0.5	Başarısız	
FF	0-39	0.0	Başarısız	
NA	x	0.0	Devamsız	

Öğrencilerin akademik başarı/başarısızlıkları üzerinde etkili olabileceği daha önceki çalışmalarda belirlenmiş olan özellikler Bölüm 4.1.1’de ayrıntılı olarak bahsedilecek olan 3 farklı özellik seçim yöntemi ile ortaya çıkarılmıştır. Bu özelliklere ait veriler ise öğrencilerden alınan yazılı izinlerin ardından bir araya getirilmiştir.

Öğrencilerin yazılı izinleri ile bir araya getirilen verilerden oluşan veri kümesi üzerinde Bölüm 4.2’de bahsedilen klasik sınıflandırma metotlarının uygulanması sonucunda tatmin edici bir sınıflandırma performansı elde edilememiştir. Bu performansı artırabilmek ve daha sağlıklı bir akademik başarı değerlendirme modeli oluşturabilmek için çizge tabanlı veri madenciliğinden yararlanılmıştır. Öncelikle her bir öğrenci için sosyo-demografik özellikleri ve akademik geçmişleri bir araya getirilerek bir çizge yapısı oluşturulmuştur. Bu çizge yapısı oluşturulurken veri kümesindeki verilerin mantıksal ilişkilerle birbirine bağlanması çalışmanın öne çıkan bir yapısal özelliği olmuştur.

Oluşturulan bu mantıksal çizgelerden başarı ve başarısızlıkları için elde edilen sıklıkla tekrar eden alt-çizgelerin öğrencilere ait çizge yapılarında mevcut olup/olmaması durumu da 1 veya 0 şeklinde ifade edilerek veri kümesine özellik olarak eklenmiş ve böylelikle veri kümesi zenginleştirilmiştir.

Aynı merkezi sınav ile aynı değerlendirme kriterlerine dahil olarak, çok yakın başarı yüzdelik dilimde bulunarak aynı kurumda eşit şartlarda ve aynı meslek alanı dahilinde ders içerikleri ve kapsamı bakımından yaklaşık %80 benzerlik oranına sahip olan programlarda eğitim alan öğrencilerden birebir yazılı izin alınarak verilerin bir araya getirilmesi oldukça zorlu bir süreç olmuştur. Bu zorlu süreç tez çalışmasını araştırmacıların yaptığı diğer modelleme yaklaşımlarından da farklı kılmıştır.

Daha önce yapılmış çalışmalar incelendiğinde, veri kümesi olarak kullanılan öğrenci sayısının bu tez çalışmasına nazaran oldukça yüksek olduğu görülmektedir. Örneğin, İspanya’da 2010 yılında yapılan ve öğrencilerin başarısızlık durumlarının tahmin edilmeye çalışıldığı bir çalışmada kullanılan öğrenci sayısı 670 [32]; 2005 yılında gerçekleştirilen ve öğrencilerin dönem sonu başarı durumlarının tahmin edilmeye çalışıldığı bir çalışmada kullanılan öğrenci sayısı 354 [40]; öğrencilerin okulu bırakma durumlarının tahmin edilmeye çalışıldığı bir diğer çalışmada ki öğrenci sayısı 648 [45]; öğrencilerin akademik performanslarının tahmin edilmeye çalışıldığı farklı bir çalışmada ise kullanılan öğrenci sayısı 10330’dur [44].

Bu çalışmada ise sayı olarak az fakat özellik vektörleri geniş, özellikler arasındaki ilişkiler de göz önünde bulundurulmuş olan bir veri kümesi ile oluşturulan modellerde de yüksek oranlarda başarı elde edilebileceği varsayımı benimsenmiştir.

4.1.1. Özellik Seçim Yöntemleri

Öğrencilere ait sosyo-demografik, aile yapısı, akademik geçmiş, ilgi alanları, cinsiyet gibi bilgilerin analiz edilerek sınıflayıcılar (classifiers) yardımı ile sınıflandırılarak çeşitli modeller oluşturulması tez çalışmasına konu olan problemin çözümü konusunda bir mühendislik yaklaşımıdır [63]. Bu yaklaşımda ilk adım gereksiz veya en az etkisi olan özelliklerin elenmesi için yapılacak olan özellik seçimi (feature selection) işlemidir. Özellik seçimi işlemi makine öğrenmesinin ön işleme adımıdır.

Veri madenciliği uygulamalarında sınıflandırma metotları kullanılmadan önce doğruluk oranının artırılması ve işlem zamanını azaltmak için modeli etkilemeyen özelliklerin özellik kümesinden çıkarılması için özellik seçimi yöntemleri önemli rol oynar. Bu çalışmada da bu sebeplere ek olarak çizge madenciliği yöntemlerinin zaman karmaşıklığı sebebiyle çizgenin oluşturulmasında kullanılacak özellik sayısını azaltmak için özellik seçim yöntemleri kullanılmıştır.

Bu bölümün devamında veri kümesinin özelliklerini belirlemek için çalışmada kullanılmış olan Bilgi Kazanımı (Information Gain), Kazanım Oranı (Gain Ratio) ve Ki-Kare (Chi-Square) yöntemleri ile ilgili bilgilere yer verilmiştir.

4.1.1.1. Bilgi Kazanımı (Information Gain)

Entropi tabanlı bir özellik seçimi olan Bilgi Kazanımı yöntemi, bir veri kümesi içerisindeki en ayırt edici özelliği tespit etmek üzere her bir özellik için ölçüm yapan bir metottur.

Entropi, en genel ifade ile bir sistem içerisindeki düzensizliğin ya da belirsizliğin ölçüsü olarak tanımlanabilmektedir. Entropi değeri 4.1 numaralı formülde ifade edildiği gibi hesaplanır ve bu hesaplanan değer 1'e yaklaştıkça belirsizliği artmak üzere 0-1 aralığında bir sayıya denk düşer.

$$E(D) = -\sum_{k=1}^m p_i \log_i(p_i) \quad (4.1)$$

Bu hesaplamada D veri kümesi olmak üzere p_i değeri ‘i’ sınıfının D veri kümesindeki olasılığıdır ve ‘i’ sınıfına düşen örnek sayısının toplam örnek sayısına bölünmesi sonucu elde edilir.

Hesaplama sonucunda yüksek entropiye sahip verinin daha fazla bilgi içerdiği kabul edilir.

4.1.1.2. Kazanım Oranı (Gain Ratio)

Kazanım Oranı yöntemi Bilgi Kazancı yöntemindeki sapma oranının azaltılması amacıyla geliştirilmiş bir özellik seçim yöntemidir. Farklı değerlere sahip özellikleri seçme eğiliminde olan Bilgi Kazancı yaklaşımında sonuçlar sapmalı olarak elde edilir. Bu dezavantajı ortadan kaldırmak üzere Kazanım Oranı yönteminde bölünme bilgisinden yararlanılmaktadır.

$$\text{Bölünme Bilgisi (Split Information)} = -\sum_{i=1}^v \left(\frac{|S_i|}{|S|}\right) \log_2 \left(\frac{|S_i|}{|S|}\right) \quad (4.2)$$

Kazanım Oranı, bilgi kazancı değerleri ile bölünme bilgisinin oranlanması şeklinde hesaplanan bir çeşit normalizasyon işlemini içinde barındırır. Nihai kazanım oranının hesaplanması ve en yüksek kazanım oranına sahip özelliklerin tespit edilmesi aşağıdaki formülde ifade edildiği gibi bilgi kazanımının bölünme bilgisine oranlanması ile elde edilir.

$$\text{Kazanım Oranı (Gain Ratio)} = \frac{\text{Bilgi Kazancı}(A)}{\text{Bölünme Bilgisi}(S)} \quad (4.3)$$

4.1.1.3. Ki-kare (Chi-square)

Ki-kare (Chi-square) özellik seçim yöntemi iki basamakta uygulanan ve ki-kare istatistiğini temel alan bir metottur. Öncelikle özelliklerin sınıflara bağlı ki-kare istatistik değerleri hesaplanır ve ardından serbestlik dereceleri ve önem derecelerine

göre ki-kare değerleri incelenerek veri kümesi içerisindeki tutarsız özellikler bulununcaya dek devamlı olarak özellikler ayrıştırılır.

Birbirinden farklı iki değişken arasındaki ilişkinin bağımlılık veya bağımsızlığını belirlemek için kullanılan bir yöntem olan Ki-kare yaklaşımında, hesaplanan değer o özelliğin sınıf içerisindeki bağımlılığı hakkında bilgi vermektedir. Hesaplanan ki-kare istatistik değeri 0'a yaklaştıkça özelliğin bağımsızlaştığı, değer büyüdükçe ise bu özelliğin veri kümesi içerisinde tanımlayıcılığının arttığı ifade edilebilir.

Aşağıdaki eşitliklerde ifade edilen k sınıf sayısını, A_{ij} gözlenen frekans değerini ve E_{ij} ise beklenen frekans değerini temsil etmektedir. Son eşitlikte bölüm kısmında bulunan R_i , i. Aralıktaki veri sayısı, C_j , j. Sınıftaki gözlemlerin sayısı, N ise sınıflardaki gözlemlerin toplamını ifade etmektedir.

$$X^2 = \sum_{i=1}^2 \sum_{j=1}^k \frac{(A_{ij} - E_{ij})^2}{E_{ij}} \quad (4.4)$$

$$E_{ij} = \frac{(R_i * C_j)}{N} \quad (4.5)$$

4.1.2. Anahtar Sözcük Çıkarımı

Öğrencilerin yalnızca daha önceki dönemlerde almış oldukları derslerdeki başarı durumlarının veri kümesine özellik olarak dahil edilmesi, akademik başarının modellenmesi gibi kapsamlı bir konuda ancak yüzeysel bir bilgi sunmaktadır. Derslerdeki başarı durumu dönemsel olarak öğrencinin o derse ait yetenekleri hakkında bilgi verirken, ders içerikleri öğrencilerin ilgi alanları hakkında da gizli bilgileri barındırıyor olabilir. Bu yaklaşımla bakıldığında öğrencilerin derslerde aldıkları notlardan daha çok ders içeriklerinden elde edilen 1-gramlara yoğunlaşılması gerekliliği ortaya çıkmaktadır. Böylelikle farklı dersler arasındaki ilişkiler ve öğrencilerin teknik olarak 1-gram olarak ifade edilen anahtar sözcüklerle olan ilişkilerinde incelenmesi sağlanacaktır.

İncelenen öğrenci grubunun eğitim gördüğü vakıf üniversitesinin eğitim bilgi sisteminden gerekli izinlerle elde edilen her bir derse ait ders içeriği, dersin amacı ve 14 haftalık ders planı bilgileri bir araya getirilmiştir.

Literatürdeki Bölüm 2.1.2.’ de bahsedilen ilgili çizge madenciliği algoritmaları çizge izomorfizm ve alt çizge izomorfizmin sebep olduğu zaman karmaşıklığı sebebiyle arama uzayını azaltma yoluna giderek alt çizgeleri bulmak için yaklaşım getirirler. Yine de düğüm ve kenar sayısının çok fazla olması işlem zamanını oldukça arttırmaktadır. Bu sebeple de derslerin içeriklerindeki 1-gramların filtrelmesi gerekir. Bu amaçla öncelikle gereksiz kelimeler (stop words) olarak adlandırılan ‘ama, ve, veya, vb.’ sözcükler temizlenmiş ve kalan sözcüklerden 1-gramlar elde edilmiştir.

Elde edilen 1-gramların her birinin frekansları (tekrar etme sayıları) hesaplanmış ve frekansı en yüksek olan sözcükler birer özellik olarak veri kümesine eklenmiştir. Böylelikle dersler arasındaki ilişkilerin tespiti ve öğrencilerin hangi konularda başarı sağlayabildikleri konusunda bilgi elde edilmesi sağlanmıştır.

4.1.3. Çizge Veri Yapısının Oluşturulması

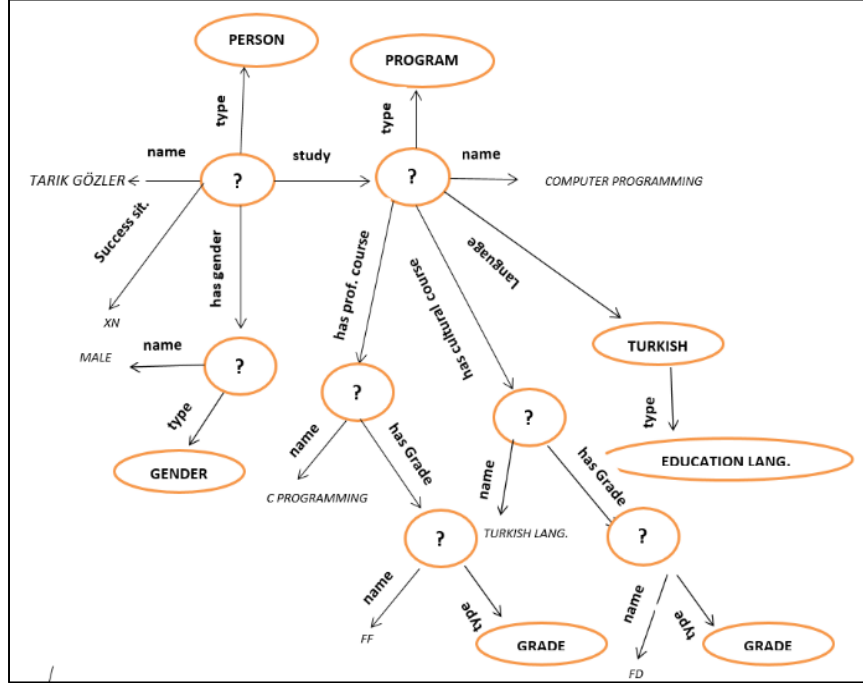
Düğüm kümesinin L_v , düğümler arasındaki ilişkiyi temsil eden kenarlar kümesinin L_E olarak temsil edildiği durumda, dört bileşenden oluşan bir çizge yapısı $g = (V, E, \mu, \nu)$ şeklinde tanımlanır. Bu tanımda;

- V düğümlerin sonlu kümesini,
- $E \subset V \times V$ kenarların sonlu kümesini,
- $\mu : V \rightarrow L_v$ düğüm etiketleme fonksiyonunu,
- $\nu : E \rightarrow L_E$ kenar etiketleme fonksiyonunu sembolize eder.

V düğümleri tanımlarken, kenarların kümesi E ise çizgenin yapısını temsil eder. $u \in V$ olmak üzere bir düğüm ve yine $v \in V$ olmak üzere u düğümüyle bağlantılı olan bir başka düğüm olmak üzere, bu iki düğüm arasındaki bağlantıyı sağlayan kenar $e = (u, v)$ $(u, v) \in E$ olacak şekilde tanımlanır [64].

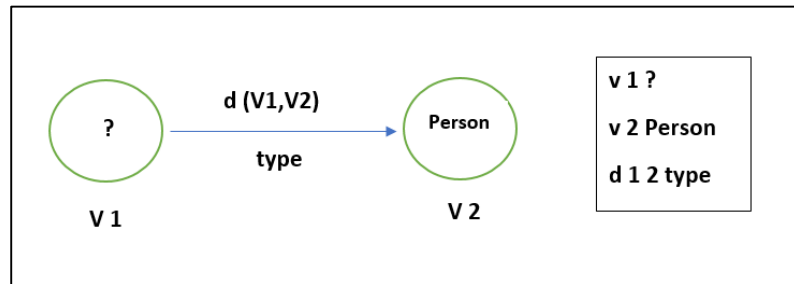
Tez çalışmasında her bir öğrenci için kişinin kendi sosyo-demografik ve geçmiş eğitim bilgilerden yararlanılarak bir çizge veri yapısı oluşturulmuştur. Oluşturulan çizgede düğümler kavramları, kenarlar ise kavramlar arasındaki ilişkileri ifade etmektedir. Çizge yapısını oluşturan kenar ve düğümler bilgi içerebilir. Eğer bu bilgi isim ya da sayı ise çizge etiketlenmiş çizge olarak adlandırılır.

Şekil 4.1’de görsel olarak ifade edilen örnek bir çizge yapısı gösterilmektedir. Burada düğümler arasındaki kenarlar üzerinde bilgiler tutulmaktadır. Bu bilgiler iki düğüm arasındaki ilişkiyi ifade edilmek üzere tutulmuş etiketlerdir.



Şekil 4.1: Başarısız bir öğrenci için oluşturulmuş örnek çizge yapısı.

Şekil 4.1’de simule edilen çizge yapısının Bölüm 4.1.4’ de bahsedilecek olan tekrarlayan alt çizge madenciliği yöntemlerinden biri olan SUBDUE yaklaşımı için uygun formatta metinsel olarak ifade edilmesi gerekmektedir. Bu metinsel ifadenin bir kısmı Şekil 4.2’de gösterilmektedir.



Şekil 4.2: Çizge yapısının gösteriminin basit ifadesi.

Her bir öğrencinin öğrencinin geçmiş dönemlerde aldığı notlardan yola çıkarak belirlenen başarı/başarısızlık durumuna göre ise her bir öğrenci XN (Negatif:

Başarısız) veya XP (Pozitif: Başarılı) olarak başlangıçta etiketlenmiştir. Bu etiketlerden yola çıkarak SUBDUE yaklaşımı ile başarılı / başarısız öğrencilere ait çeşitli örüntüler elde edilecektir.

Şekil 4.3’de ifade edilen çizge yapısında *v* ile ifade edilen her bir ifade bir düğümü, *d* ise düğüm numaraları verilmiş 2 düğüm arasındaki ilişkiyi göstermektedir. Bu gösterimin basit bir hali Şekil 4.2’de gösterilmiştir.

```
XN
% TARIK GÖZLER
v 1 ?
v 2 Person
d 1 2 type
v 3 TARIK GÖZLER
d 1 3 name
v 4 ?
d 1 4 study
v 5 program
d 4 5 type
v 6 Computer Programming
d 4 6 name
v 7 ?
d 1 7 hasGender
v 8 Gender
v 9 Male
d 7 8 type
d 7 9 name
v 10 ?
d 4 10 edu_language
v 11 Education_Language
d 10 11 type
```

Şekil 4.3: Çizgelerin metinsel ifade edilmesi örneği.

4.1.4. Çizge Madenciliği Yöntemleri ile Örüntülerin Çıkarılması

Teorik olarak tanımlanmış olan çizge veri yapısı, özellikle internet kullanımı ve sosyal ağların sosyal hayatın içerisine nüfuz etmesiyle birlikte, diğer bir deyişle sanal ağların gelişmesiyle, veri madenciliğinin birçok alanında güçlü veri yapıları olarak kullanılmaya başlanmıştır.

Özellikle ağ yapısı şeklinde birbiriyle bağlantılı verilerin kolaylıkla ifade edilebildiği bir veri yapısı olan çizgeler; moleküllerin kimyasal yapılarının ifade edilmesinde, ilaç bileşenlerinin ifade edilmesinde, sosyal ağ uygulamalarında sıklıkla kullanılmaktadır. Kullanım alanı günden güne genişleyen bu veri yapılarının analizine olanak sağlayan çizge tabanlı örüntü madenciliği de veri madenciliğinin bir alt dalı olarak karşımıza çıkmaktadır.

Tez çalışması kapsamında sınıflandırma metotları için anlamlı olabilecek ve özellik seçim yöntemleri ile belirlenen bazı özelliklerden oluşturulmuş veri kümesinin, çizgelerden elde edilebilecek olan ilişkisel verilerle zenginleştirilerek sınıflandırma metotlarının tahmin başarılarını artırabileceği yaklaşımı benimsenmiştir. Bu yaklaşımdan hareketle 142 adet öğrencinin birçok farklı özelliğinden oluşturulmuş olan bir veri kümesinden elde edilen 142 adet küçük çizge yapısı ile bir çizge kümesi yaratılmıştır.

Çok sayıda küçük çizge içeren çizge kümesinden sıklıkla tekrarlayan alt çizgelerin bulunması yaklaşımları üzerinde yoğunlaşarak başarılı ve başarısız öğrencilerin temel karakteristiklerini ifade edebilecek bazı örüntüler elde edilmeye çalışılmıştır.

Tekrarlayan alt çizgelerin bulunması amacıyla literatürde kullanılan birçok yaklaşımdan bahsetmek mümkündür. Bu yaklaşımlar ile ilgili bilgiler ve birbirleri ile ilgili olan kıyaslamaları Bölüm 2.1.2.1’de mevcuttur. Bu çalışmada tercih edilen yaklaşım ise aç-gözlü yaklaşımı benimseyen gözetimsiz bilgi keşif yöntemlerinden biri olan SUBDUE yöntemi olmuştur.

4.1.4.1. SUBDUE

SUBDUE yöntemi yalnızca tekrar eden alt çizgeleri bulmakla kalmayıp, bulunduğu alt çizgeleri de sıkıştırarak ilginç, gözle görülmesi mümkün olmayan ilişkileri de ortaya çıkarmakta oldukça başarılı ilişkisel öğrenme yöntemidir. Bir bilgi keşif sistemi olan SUBDUE, açgözlü yaklaşımını benimseyen arama tabanlı çizge madenciliği alanındaki öncü çalışmalardan biridir [12].

Tek bir büyük çizge veya birçok küçük çizgeden oluşan bir çizge kümesi üzerinde çalışabilen SUBDUE yaklaşımında girdi çizgeleri etiketli veya etiketsiz de olabilir. Bu metot Minimum Tanımlama Uzunluğu (MTU) prensibine göre girdi veri kümesini sıkıştırarak, alt çizgeleri elde eden bir yaklaşımı benimser [65].

SUBDUE yöntemi benzersiz etiketler içeren tüm köşelerden oluşan alt yapılardan başlayarak, hesaplaması sınırlandırılmış bir ıřın araması gerçekleştirir. Olası aday alt yapıları oluşturabilmek için örnek çizgeler yardımıyla alt yapıları bir kenar veya bir köşe ile genişletme yoluna gider. Alt yapılar daha sonra veri kümesinin Tanımlama Boyutunu (Description Length) ne kadar iyi sıkıştırılabildiklerine göre

değerlendirilirler. Girdi veri kümesi G 'nin tanımlama boyutu, altyapı S kullanılarak aşağıdaki formülle hesaplanır.

$$I(S) + I(G|S) \quad (4.6)$$

Burada S , G veri kümesini sıkıştırmak için kullanılan altyapıdır. $I(S)$ ve $I(G|S)$ ise sıkıştırma işleminden sonra, veri kümesi G ve alt yapı S 'nin kodlama uzunluğunu ifade etmektedir. Daha fazla genişleme için alınan altyapıların sayısı arama ışının boyutunu ifade eder.

Algoritma ilk altyapıyı bulduktan sonra, çizgeyi sıkıştırır ve aynı işlemi tekrar edebilmek için süreci yineleyebilir. SUBDUE, yapılarında ufak farklılıklar olan alt yapıların keşfedilmesine izin veren, tam olmayan bir eşleşme gerçekleştirebilen bir yapıya sahiptir.

```

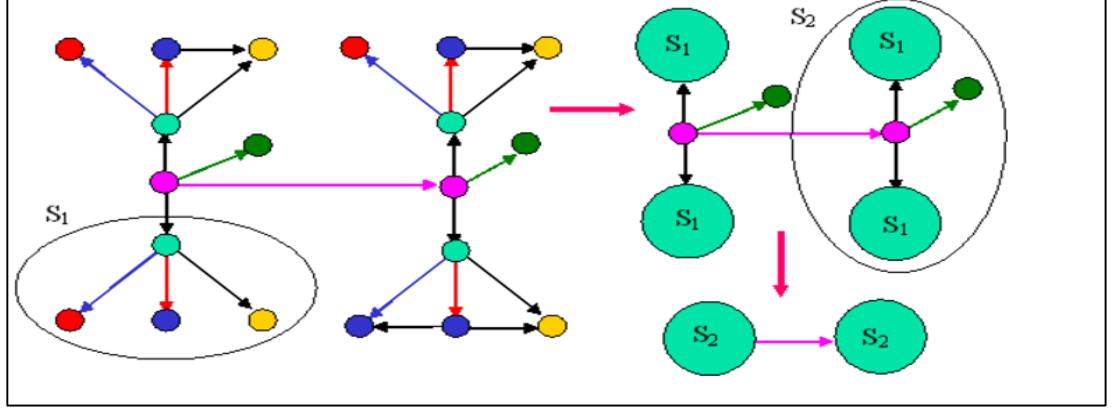
SUBDUE( $G, limit, beam$ )
   $D = \{\}$ 
   $S = \text{vertices}(G)$ 
  while (computation <  $limit$ ) and ( $S \neq \{\}$ )
    order  $S$  from best to worst using MDL and background knowledge rules
     $S = \text{first } beam \text{ substructures of } S$ 
     $b = \text{first}(S)$ 
     $D = D \cup \{b\}$ 
     $E = \{b \text{ extended by one edge in all possible ways } \}$ 
     $S = S \cup E$ 
  return  $D$ 

```

Şekil 4.4: SUBDUE keşif algoritmasının sözde kodu.

Şekil 4.4'de sözde kodu verilmiş olan bu yöntemde verinin hiyerarşik yapısını oluşturmak için tekrar eden parçalar sıkıştırılarak, benzer alt çizgelerin bulunması yoluna gidilir [66],[67]. Öncelikle tek bir düğümden başlanarak, her iterasyonda açgözlü mantığı ile o an için en uygun başka bir kenar ekleyerek alt parça büyütülür. Bir eşik değeri kullanılarak her bir alt çizge için frekans hesaplanır ve bu eşik değerinin üstünde frekansa sahip tekrar eden alt çizgeler seçilir.

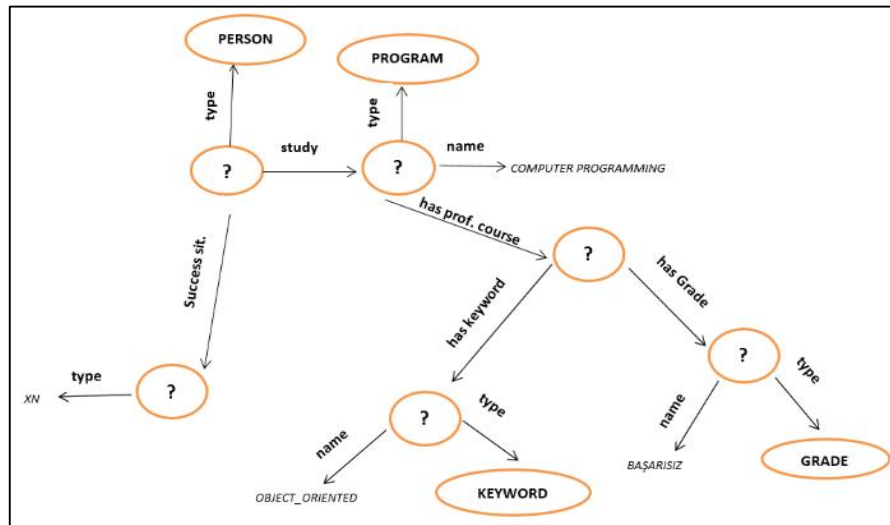
Şekil 4.5’de SUBDUE yöntemi girdi olarak kabul edilen bir çizge veri kümesindeki 4 tane S_1 örüntüsüne ait örneğin keşfedilmesi temsil edilmiştir. Bir sonraki iterasyonda S_2 örüntüsü keşfedilmiş ve çizge S_2 örüntüsü ile sıkıştırılmıştır.



Şekil 4.5: SUBDUE yöntemi ile alt çizgelerin elde edilmesi örneği [68].

Tez çalışmasında SUBDUE yöntemi kullanılmadan önce; her bir öğrenci için oluşturulan çizgeler, öğrenciye ait verilerden oluşturulmuş ve öğrencinin daha önceden almış olduğu derslere ait genel not ortalamasının başarılı veya başarısız olma durumuna göre pozitif (XP) veya negatif (XN) olacak şekilde iki sınıf olarak etiketlenmiştir. Sonrasında çizgelerden SUBDUE yöntemi ile her sınıf için sıklıkla tekrar eden alt çizgeler çıkartılmıştır. Başarılı veya başarısız olarak belirlenen öğrenciler arasındaki ortak örüntüler (alt çizgeler) elde edilmeye çalışılmıştır.

Şekil 4.6’ da SUBDUE yöntemi elde edilmiş olan başarısız öğrenciler için elde edilmiş frekans değeri yüksek bir örüntü örneği görülmektedir.



Şekil 4.6: Başarısız öğrenciler için elde edilen örnek alt çizge.

4.2. Özellik Vektörlerinin Oluşturulması

Tez çalışmasında iki farklı tip özellik vektörleri bir araya getirilmiştir. Birinci özellik vektöründe literatürde akademik başarının modellenmesi konusunda yapılan çalışmalarda üzerinde sıklıkla çalışılan ve akademik başarı/başarısızlık üzerine doğrudan etkisi olduğu tespit edilen yaş, cinsiyet, aile yaşantısı, ebeveynlerin sağlık ve eğitim durumları gibi sosyo-demografik veriler ile öğrencilerin geçmiş dönemlerde göstermiş oldukları akademik performans verileri bir araya getirilmiştir. Bir araya getirilen bu özellikler, özellik seçim yöntemlerince sınıflandırma metotları üzerinde etkisi olduğu tespit edilen özelliklerdir. İlk adımda oluşturulan bu özellik vektörü üzerinde sınıflandırma metotları uygulandığında, veri kümesinin de az sayıda örnek içermesinden kaynaklı olarak tatmin edici sonuçlar elde edilememiştir. Bu sebeple veri sayısı az olan kümelerde de modelleme başarısını artırmak amacıyla akademik başarı üzerine etkisi olabilecek ilişkisel veriler üzerine de yoğunlaşmıştır.

İkinci özellik vektörü öğrencilerden doğrudan elde edilen veriler yerine eldeki verilerden oluşturulan çizgelerden tekrarlayan çizge madenciliği yöntemi Subdue ile tespit edilmiş alt çizgelerden frekansı en yüksek olan 20 tanesi seçilerek oluşturulmuştur. Özellik vektörünün her boyutu başarılı ve başarısız öğrencilerden elde edilen örüntüleri ifade etmektedir. Öğrencinin ilgili örüntüyü içerip içermediği bilgisine göre de özellik vektörü oluşturulmuştur.

4.3. Sınıflandırma Metotları

Bu bölümde öğrencilerin başarısının tahmin edilmesi için çizgelerle zenginleştirilmiş veri kümelerine uygulanan gözetimli öğrenme yöntemleri incelenmiştir.

4.3.1. Naive Bayes Sınıflandırıcı (NB)

X bir veri kümesi olmak üzere, X ‘kanıt’ olarak isimlendirilir ve n tane özellik üzerine yapılmış bir dizi hesaplama ile açıklanır. H ise, X veri kümesindeki elemanların C sınıfına ait olduğunu kabul eden hipotezlerdir. Bayes teoremini temel alan sınıflandırma problemlerinde yapılmak istenen, kanıtlanmış veya gözlenen veriler göz önüne alındığında H hipotezinin doğru olabilme olasılığının belirlenmesidir.

$P = (H|X)$ şeklinde ifade edilen olasılık durumu sonsal olasılıktır ve H'nin X üzerinde koşullu olasılığını anlatır. Aynı şekilde $P = (X|H)$ şeklinde ifade edilen koşullu olasılıkta X'in H üzerinde koşullu olasılığının ifadesidir. $P = (H)$ ise önsel olasılık olarak isimlendirilir ve H'nin bağımsız olasılığını ifade eder.

Bayes Teoremi koşullu olasılık hesabında kullanılır ve (8) numaralı formülde gösterildiği gibi hesaplanır [7].

H ve X rastgele olaylar olmak üzere,

$$P(H|X) = P(X|H)P(H)/P(X) \quad (4.7)$$

$P(H)$, H olayının bağımsız olasılığını; $P(X)$, X olayının bağımsız olasılığını; $P(X|H)$ H olayının gerçekleştiği bilindiği durumda X olayının gerçekleşme olasılığını ve $P(H|X)$ ise X olayının gerçekleştiği bilindiği durumda H olayının gerçekleşme olasılığını ifade eder.

Naive Bayes sınıflandırıcı yukarıda tanımlanan Bayes teoremini temel alan, istatistiksel bir sınıflandırma metodudur. Bir örneğin (instance) hangi sınıfa ait olabileceğini olasılıksal olarak ön görebilen bir yapıya sahiptir. Bu anlamda Naive Bayes sınıflandırıcı hedef değişkenle diğer bağımsız değişkenler arasında olan ilişkiyi analiz etme yeteneğine sahip, tahmin edici ve tanımlayıcı bir sınıflandırma yöntemidir [69]. Bir özellik değerinin verilen bir sınıfa etkisinin diğer özelliklerin değerlerinden bağımsız olduğunu varsayar. Bu varsayımda ilgili hesaplamaları basitleştirir.

Naive Bayes sınıflandırıcının çalışma prensibi sırasıyla aşağıdaki adımlarla ifade edilebilir.

- D tüm kayıtları ve kayıtların ait oldukları sınıf etiketlerini içeren bir eğitim kümesi olmak üzere, her bir kayıt n-boyutlu bir $X=(x_1, x_2, \dots, x_n)$ şeklinde ifade edilen özellik vektörüyle gösterilmektedir.
- $C_1, C_2, \dots, C_m$ olarak ifade edilen m adet sınıf olduğu varsayıldığında; sınıflandırıcı verilen her X için, X'in en yüksek sonsal (posterior) olasılık değerine sahip olduğu sınıfa ait olduğunu tespit eder. Sınıflandırıcının X'in C_i sınıfına ait olduğunu tespit etmesi için kesinlikle ve kesinlikle;

$$1 \leq j \leq m, j \neq m \text{ için } P(C_i|X) > P(C_j|X)$$

şartını sağlaması gerekmektedir.

- $P(X)$ değeri tüm sınıflar için sabit olduğundan dolayı, burada $P(X|C_i)P(C_i)$ değerinin maksimum olması gerekmektedir. Eğer sınıfa ait önsel (prior) olasılık değeri bilinmiyorsa tüm sınıflar eşit kabul edilir. Yani, $P(C_1) = P(C_2) = \dots = P(C_m)$ olacaktır ve bu durumda $P(X|C_i)$ değerinin maksimum olması gerekmektedir. Burada sınıfın önsel olasılık değerinin hesabı; C_i sınıfındaki D eğitim setine ait eğitim verilerinin sayısı $|C_{i,D}|$ olarak gösterilmek üzere, $P(C_i) = |C_{i,D}|/|D|$ şeklinde hesaplanır.
- Üzerinde çalışılan veri setinin özellik vektörünün geniş olması sebebiyle $P(X|C_i)$ değerinin hesaplanması çok maliyetli olabilir. Bu hesaplama maliyetini azaltmak amacıyla naif bir yaklaşım olan sınıf koşullu bağımsızlığı geliştirilmiştir. Bu yaklaşımda verinin sınıf etiketi göz önüne alındığında özellik değerlerinin koşullu olarak birbirinden bağımsız olduğu varsayımı benimsenir. Böylelikle $P(X|C_i)$ değeri kolaylıkla (4.8) numaralı formülde ifade edildiği gibi hesaplanabilir.

$$\begin{aligned} P(X|C_i) &= \prod_{k=1}^n P(x_k|C_i) \\ &= P(x_1|C_i) * P(x_2|C_i) * \dots * P(x_n|C_i) \end{aligned} \quad (4.8)$$

- Yukarıdaki adımların sırasıyla gerçekleştirilmesi sonucu ilerleyen Naive Bayes sınıflandırıcı metodu tez çalışmasında kullanılan iki gözetimli öğrenme metodundan biridir.

4.3.2. K-En Yakın Komşuluk Yöntemi (k-NN)

K-En yakın komşuluk yöntemi Naive Bayes yönteminden farklı olarak mesafeye dayalı sınıflandırma yaklaşımını benimseyen bir gözetimli öğrenme algoritmasıdır. Sınıflandırılacak örneğin öğrenme kümesindeki verilere benzerliklerinin hesaplanması sonucu sınıflandırma problemine çözüm üretilmeye çalışılır. Algoritma yapısı gereği

çok boyutlu özellik uzayında örnekleri tahmin etmek veya örneklerin hangi sınıfa ait olduğunu tespit etmek için en yakın komşu örneklerden faydalanır.

En yakın komşular tespit edilirken sınıflandırılacak örnek ile eğitim kümesindeki örnekler arasında mesafe ölçümü yapılır. Yapılan mesafe ölçüm değerleri en yakından en uzağa olacak şekilde sıralanır. Böylelikle sınıflandırılacak örneğe en yakın mesafedeki eğitim kümesi örneğinden en uzak olana doğru bir sıralama da yapılmış olmaktadır.

Algoritmanın verimli çalışabilmesinde önemli etkisi olduğu belirlenen üç farklı faktör vardır. Bu faktörler,

- En yakın komşu sayısı (k),
- Eşik değeri,
- Benzerlik ölçütü,
- Öğrenme kümesindeki normal davranışlı veri sayısının yeterli olmasıdır.

Gürültülü veriden az etkilenen, uygulaması, anlaşılması ve kullanımı oldukça kolay olan K-en yakın komşuluk yöntemi, özellikle en yakın komşu sayısı(k) değerinden ve seçilen uzaklık ölçütünden oldukça etkilenmesi gibi dezavantajlara sahiptir.

Tez çalışması kapsamında kullanılan bir diğer gözetimli öğrenme metodu olan K-en yakın komşuluk yönteminde uygulanan adımları aşağıdaki gibi sıralayabilmek mümkündür.

- Test kümesi T ve öğrenme kümesi S olmak üzere; $T=\{t_1, t_2, \dots, t_n\}$ test kümesindeki bir elemanın $S=\{s_1, s_2, \dots, s_m\}$ öğrenme kümesinin her bir verisine olan uzaklığı $1 \leq i \leq n$ ve $1 \leq j \leq m$ olmak üzere (4.9) numaralı formülle hesaplanır.

$$sim(t_i, s_j) = \frac{t_i \cdot s_j}{\|t_i\| \|s_j\|} \quad (4.9)$$

- Her bir verinin S öğrenme kümesindeki verilere olan uzaklıkları küçükten büyüğe sıralanır ve belirlenen ilk “k” tanesi alınarak uzaklık ortalamaları (4.10) numaralı formülle hesaplanır.

$$sim_{ortalama}(t_i) = \frac{\max[\sum_{j=1}^k sim(t_i, s_j)]}{k} \quad (4.10)$$

- Hesaplanan uzaklık ortalama deęerleri arasından, belirlenmiř olan eřik deęerinden daha büyük olanlar normal, daha küçük olanlar ise anormal olarak sınıflandırılmıř olurlar.

5. DEĞERLENDİRME

Tezde önerilen yaklaşım, yazılı izinlerle kişilerden ve eğitim aldıkları vakıf üniversitesinden toplanmış gerçek veri kümelerinde test edilmiştir. Çalışma boyunca yapılan tüm işlemler 2.9 Ghz Intel Core İ7-7500U, 12 Gb Ram özelliklerine sahip bir bilgisayar üzerinde yapılmıştır. Özellik seçim yöntemleri, anahtar sözcük çıkarımı, çizge yapısının oluşturulması ve sınıflandırma metotlarının tamamı Windows 10 işletim sisteminde gerçekleştirilmiştir. Tekrarlayan alt çizgelerin tespiti konusunda yapılan hesaplamalar ise aynı bilgisayar üzerine kurulu Virtual Box sanal makine yazılımı aracılığı ile Ubuntu işletim sisteminde 4Gb Ram ayrılarak çalıştırılmıştır.

Anahtar sözcük çıkarımı ve çizge yapısının oluşturulmasındaki tüm hesaplama işlemleri Visual Studio 2013' de araştırmacı tarafından geliştirilen uygulama aracılığı ile uygulanmıştır. Özellik seçim yöntemleri ve sınıflandırma metotlarının uygulanmasında ise RapidMiner Studio paket programından faydalanılmıştır.

Tekrarlayan alt çizgelerin tespiti konusunda yapılan hesaplamalar Ubuntu işletim sistemi üzerinde Subdue-5.2.2 isimli çizge tabanlı bilgi keşfi sağlayan, görsel ara yüzü olmayan komut satırları ile yönetilen bir program ile gerçekleştirilmiştir.

Önerilen yaklaşımın ilk adımında literatür taraması yapılarak eğitimsel veri madenciliği alanında birçok çalışmaya konu olan 'akademik başarının modellenmesi' problemi için kullanılan yöntemler ve veri kümeleri incelenmiştir. İncelenen bu çalışmalar doğrultusunda akademik başarı üzerinde etkisi olduğu belirlenmiş özellikler ortaya çıkarılmıştır ve bu özellikler bir araya getirilerek öğrencilere uygulanmak üzere bir bilgi formu düzenlenmiştir.

Hazırlanan bilgi formu, gerçek bir veri kümesi oluşturmak amacıyla İstanbul ilinde faaliyet gösteren bir vakıf üniversitesi olan İstanbul Aydın Üniversitesi bünyesindeki Anadolu Bil Meslek Yüksekokulu Bilgisayar Teknolojisi Bölümü programlarında kayıtlı olan 142 adet öğrenciye uygulanmıştır. Öğrencilerin geçmiş dönem akademik başarı durumları incelendiğinde 92 tanesinin başarılı, 50 tanesinin ise başarısız olduğu görülmektedir. Öğrenciler aynı meslek alanında ve eğitim kurumunda, ders içerikleri bakımından %80 benzerlik oranına sahip programlarda kayıtlıdır. Ayrıca ortak merkezi bir sınav sonucu, çok yakın bir yüzdelik dilim içerisinde bulunarak bu eğitim kurumuna girmeye hak kazanmış öğrencilerdir.

Bilgi formu aracılığıyla öğrencilerden kendi yazılı izinleri doğrultusunda sosyo-demografik özellikleri, aile yapısı, akademik geçmişi gibi kişisel bilgileri toplanmıştır. Böylelikle geliştirilen yaklaşım için gerçek bir veri kümesi hazırlanması sağlanmıştır.

Öğrencilerden toplanan veriler üzerine uygulanan Bilgi Kazanımı, Kazanım Oranı ve Ki-kare özellik seçim yöntemleri ile akademik başarı/başarısızlıkları üzerinde hangi özelliklerin daha etkili olabileceği belirlenmiştir. Tespit edilen ve daha sonraki adımlarda sınıflandırma yöntemleri için kullanılacak özellikler Tablo 5.1’de verilmiştir.

Tablo 5.1: Özellik seçim yöntemleri ile belirlenen özellikler.

Öğrenciye ait demografik veriler	Öğrenciye ait akademik veriler
Cinsiyet Aile evlilik durumu Aile sağlık durumu Anne & Baba eğitim durumu Anne & Baba mesleği Anne & Baba çalışma durumu Anne & Baba iş sektörü Kardeş sayısı Adres bilgisi (İlçe)	Program adı Eğitim dili Genel & Mesleki geçmiş dönem ders notları Genel başarı durumu

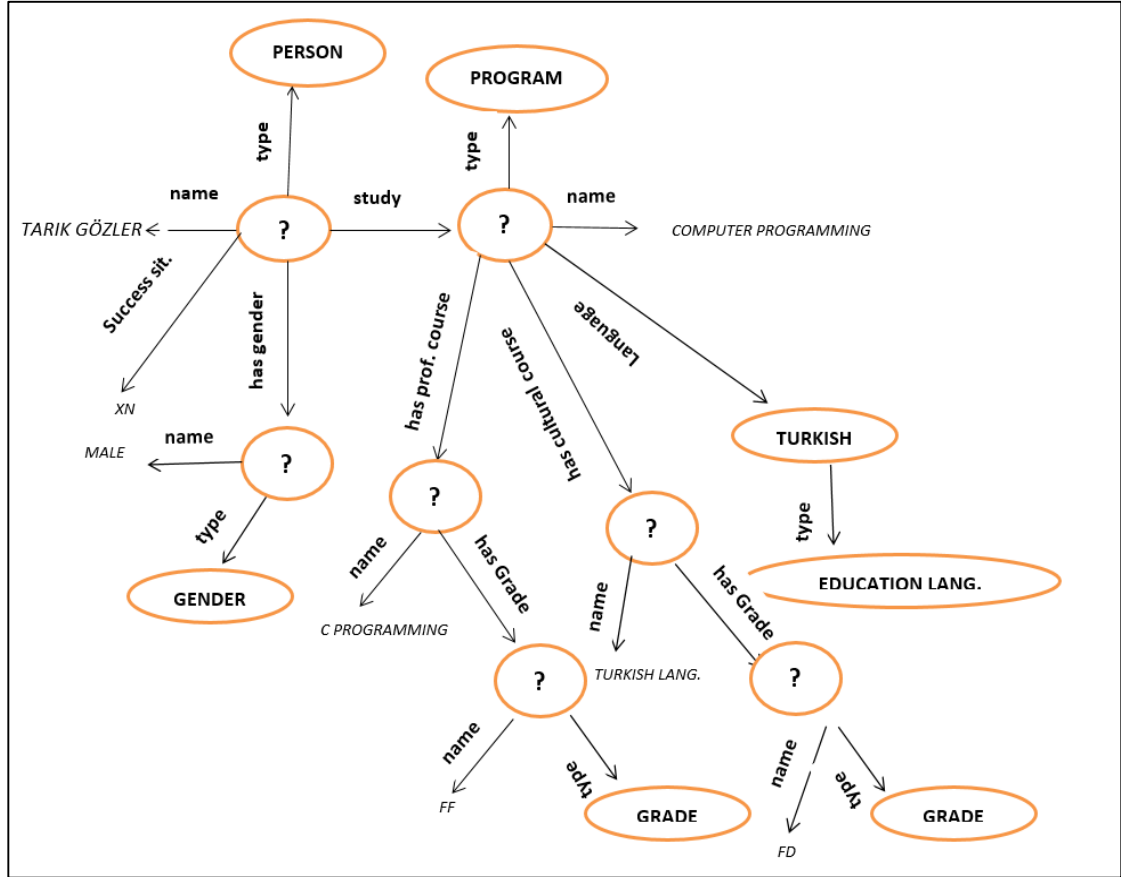
Öğrencilerin sosyo-demografik verilerinin yanı sıra akademik geçmişleri de akademik başarının modellenmesi konusunda daha önceki yapılan çalışmalarda görüldüğü üzere büyük önem taşımaktadır. Buradan hareketle sırasıyla öğrencilerin akademik geçmişlerinde almış oldukları ders içeriklerinin incelenerek frekansa bağlı olarak anahtar sözcüklerin çıkarılması ve bu sözcüklerin birer özellik olarak veri kümesine eklenmesi ile birlikte dersler arasındaki ilişkilerin tespiti ve öğrencilerin derslerdeki başarıları ile ilgili bilgi elde edilmesi sağlanmıştır.

Sıklıkla tekrarlayan alt çizgelerin çıkarılması algoritmalarının uzun işlem zamanları sebebiyle anahtar kelimelerin filtrelenmesi gerekmiştir. Ders içeriklerinden anahtar sözcükler elde edilirken, bu sözcüklerin tekrar sayıları temel alınmıştır. Frekansı en yüksek olan 25 anahtar sözcük tespit edilmiştir. Elde edilen 25 anahtar sözcük Tablo 5.2’de listelenmiştir.

Tablo 5.2: Ders içeriklerinden elde edilen anahtar sözcükler.

Bilgisayar	Program	Kullanıcı	Değişken	Analiz
Algoritma	Ağ	Sistem	İşletim sistemi	Nesnetabanlı
Donanım	Teknoloji	Güvenlik	Döngü	Fonksiyon
Yazılım	Veritabanı	Veri	Görsel	Komut

Bir sonraki aşamada ise her öğrenci için Tablo 5.1’de bahsedilen ve sonrasında ders içeriklerinden elde edilen 1-gramlarının da eklenmesi ile genişletilen özelliklerden yararlanılarak Şekil 5.1’dekine benzer çizgeler elde edilmiştir.



Şekil 5.1: Örnek çizge yapısı.

Oluşturulan çizge yapılarına Subdue sıklıkla tekrarlayan alt çizgelerin çıkarılması yöntemi uygulanırken iterasyon sayısı, aday alt-çizgelerin minimum ve maksimum boyutları ve eşik değeri gibi belirleyici özellikler etkili olmaktadır.

Çalışmada tüm bu belirleyici özelliklerin aşağıda listelenmiş değerler olarak kabul edilmesi durumunda 86 adet alt-çizge elde edilmiştir. Uygulamada kullanılan değerler aşağıdaki gibidir.

- İterasyon sayısı: 1
- Aday alt çizgelerin minimum boyutu: 4
- Aday alt çizgelerin maximum boyutu: 20
- Eşik değeri: 0.3
- Budama (Prune): Aktif (True)

Çizge sayısının fazla olduğu bu çalışmada belirleyici özelliklerin yukarıda listelenen değerler olarak kabul edilmesi durumunda, oluşturulan 86 alt-çizge çalışma için kullanılan bilgisayar üzerinde yaklaşık 4 haftalık bir süreçte elde edilmiştir.

Bu çalışmada öncelikle özellik seçim yöntemleri yardımıyla belirlenen özellik vektörü kapsamında bir araya getirilen veriler üzerine Naive Bayes sınıflandırıcı uygulanmış ve tatmin edici bir sonuç elde edilememiştir.

Tablo 5.3: Çizgesiz veri kümesinde Naive Bayes sınıflandırıcı performans tablosu.

	Başarılı (Doğru)	Başarısız (Doğru)	Sınıf doğruluğu
Başarılı (Tahmin)	81	8	% 91,01
Başarısız (Tahmin)	11	42	%79,25
	%88,04	%84,00	

Tablo 5.3’de çizgeler eklenmemiş veri kümesi üzerine uygulanan sınıflandırıcının tahmin oranları gösterilmektedir. İlgili veri kümesi için Naive Bayes sınıflandırıcının doğru sınıflandırma oranı %86,62’dir.

Bölüm 4.3.2’de uygulama adımları verilmiş olan K-en yakın komşuluk algoritması tez çalışması kapsamında henüz alt-çizgelerle zenginleştirilmemiş öğrencilerin yalnızca sosyo-demografik ve geçmiş akademik performans verilerinden oluşan veri kümesi üzerine uygulandığında ise %70,83 gibi kötü bir performans sergilemiştir. Burada k değerinin en ideal 3 olduğu görülmüş ve k=3 olarak belirlenmiştir.

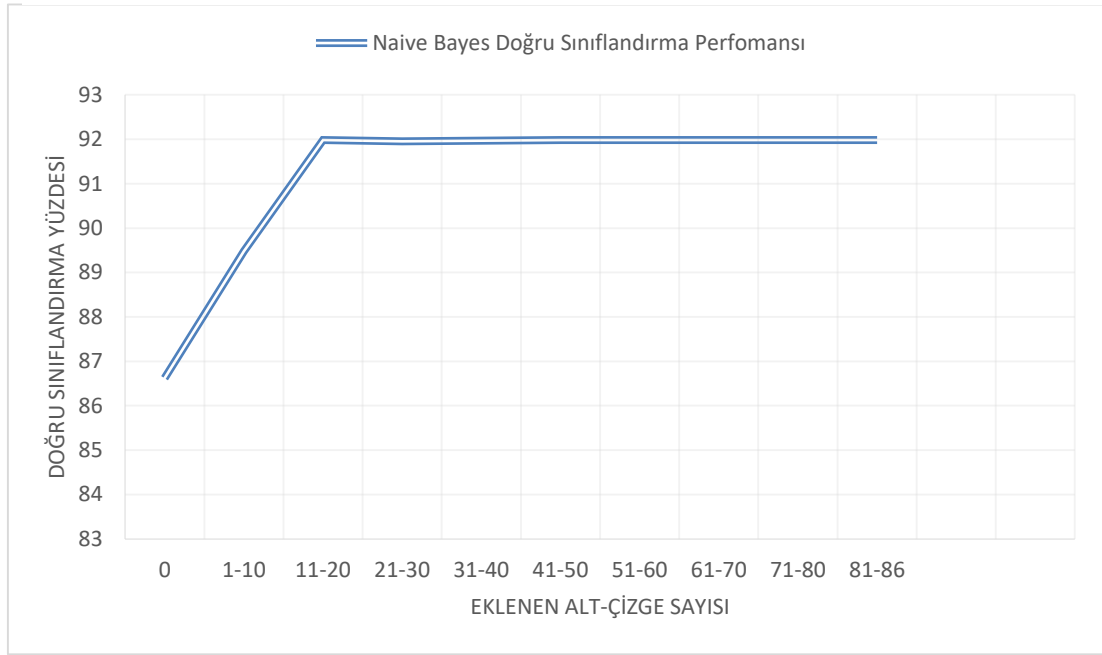
Tablo 5.4: Çizgesiz veri kümesinde k-NN En Yakın Komşuluk yöntemi performans tablosu.

	Başarılı (Doğru)	Başarısız (Doğru)	Sınıf doğruluğu
Başarılı (Tahmin)	23	2	% 92,00
Başarısız (Tahmin)	12	11	%47,83
	%65,71	%84,62	

Her iki sınıflandırma metodunda da tatmin edici sonuçların elde edilememesi üzerine veri kümeleri sıklıkla tekrar eden alt-çizgelerle zenginleştirilerek tekrar her iki gözetimli öğrenme metodu için doğru sınıflandırma performansları ölçülmüştür.

Subdue yöntemi ile elde edilen alt-çizgelerden frekans değeri en yüksek olan 20 tanesi seçilmiş ve özellik vektörüne eklenmiştir. Bu kısıtlamanın sebebi Şekil 5.2’de ve Şekil 5.3’de görüldüğü üzere her iki sınıflandırma metodunda da 20 adetten fazla alt-çizgenin veri kümesine dahil edilmesi durumunda sınıflandırma performanslarında artış gözlenmemiş olmasıdır.

Öncelikle doğru sınıflandırma oranının artırılması hedefi ile tekrarlayan alt-çizgeler ile zenginleştirilen veri kümesi üzerine tekrar Naive Bayes sınıflandırma yöntemi uygulanmıştır. Bu hususla ilgili elde edilen sonuçlar Tablo 5.6’da gösterilmiştir.



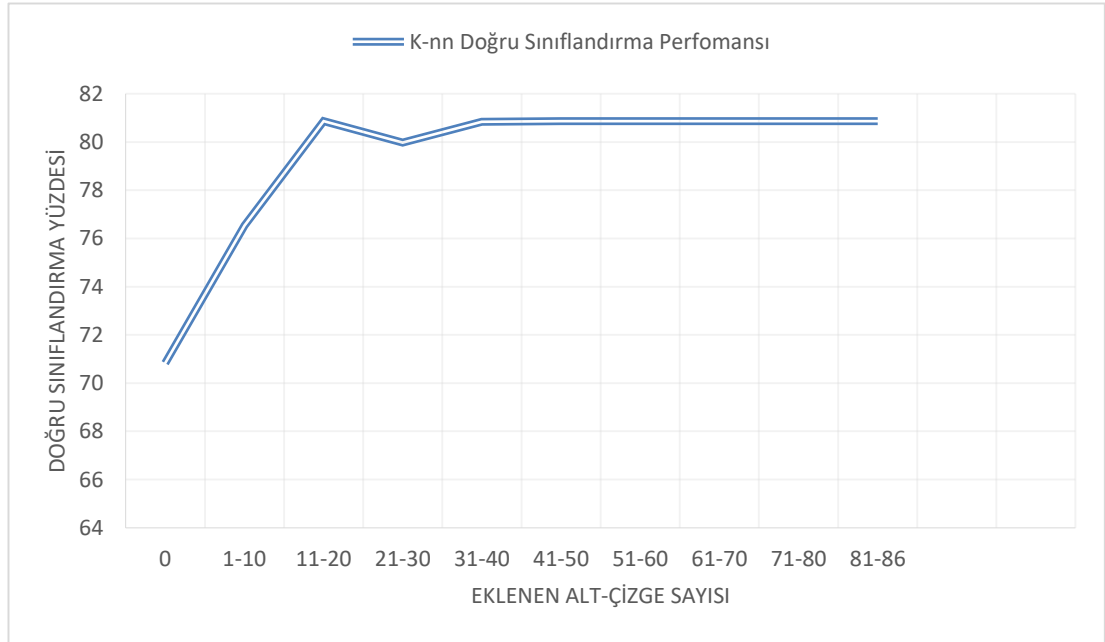
Şekil 5.2: Eklenen alt-çizge sayısına göre Naive Bayes sınıflandırıcının performans değişimi.

Tekrarlayan alt çizgelerle eklenmiş veri kümesine tekrar Naive Bayes sınıflandırıcı uygulandığında, doğru sınıflandırma oranının %86,62’den % 92,95’e yükseldiği görülmüştür.

Tablo 5.5: Çizgeler eklenmiş veri kümesinde Naive Bayes sınıflandırıcı performans tablosu.

	Başarılı (Doğru)	Başarısız (Doğru)	Sınıf doğruluğu
Başarılı (Tahmin)	87	5	% 94,56
Başarısız (Tahmin)	5	45	%90,00
	%94,56	%90,00	

Bir sonraki adım olarak frekansı en yüksek olan 20 alt-çizgenin veri kümesine dahil edilmesiyle elde edilen ve ilk veri kümesine göre oldukça zenginleştirilen veri kümesine yine k değeri 3 seçilerek aynı algoritma uygulandığında ise doğru tahmin oranının %81,25 ‘ e yükseldiği görülmüştür.



Şekil 5.3: Eklenen alt-çizge sayısına göre K-NN sınıflandırıcının performans değişimi.

Tablo 5.6: Çizgeler eklenmiş veri kümesinde k-NN en yakın komşuluk yöntemi performans tablosu.

	Başarılı (Doğru)	Başarısız (Doğru)	Sınıf doğruluğu
Başarılı (Tahmin)	27	1	%96,29
Başarısız (Tahmin)	8	12	%60,0
	%77,14	%92,3	

Bu çalışmada öncelikle çizgelerin katılmadığı özellik vektörlerinin sınıflandırma başarısı gözlemlenmiştir. NB sınıflandırıcısının doğru sınıflandırma (accuracy) oranı % 86,62 ve k değeri 3, bölünme oranı (split ratio) 0,66 için k-NN sınıflandırıcısının doğru sınıflandırma oranı ise % 70,83 olarak tespit edilmiştir.

Sonraki aşamada çizgelerle zenginleştirilmiş özellik vektörlerine NB ve k-NN yöntemleri uygulanmış ve karşılaştırmalı sonuçlar Tablo 5.7’de sunulmuştur.

Tablo 5.7: Çizgili/Çizgesiz özellik kümeleri için karşılaştırmalı performans tablosu.

	Özellik Kümesi (Çizgesiz)	Özellik Kümesi (Çizgili)
k-NN	%70,83	%81,25
Naive Bayes	%86,62	%92,95

Naive Bayes sınıflandırıcısının doğru sınıflandırma oranı %86,62’den %92,95’e; K-NN sınıflandırıcısının doğru sınıflandırma oranı ise %70,83’ten %81,25’e kadar yükseldiği gözlemlenmiştir.

Sıklıkla tekrar eden çizgelerin eklenmesiyle zenginleştirilen veri kümesinde sınıflandırma performanslarının artışlarının güvenilirliklerinin kontrolü için bir kez de veri kümesi 10-kat çapraz gerçeklik (10-fold cross validation) uygulanmıştır. Uygulanan bu adımda zaman karmaşıklığının çok büyük olmasından dolayı yalnızca 4 iterasyon gerçekleştirilmiştir.

10- kat çapraz gerçeklik için öncelikle 142 örnekten oluşan veri kümesi Tablo 5.8 de görüldüğü gibi 10 parçaya bölünmüştür ve ardından her bir iterasyon için sırasıyla aşağıda belirtilen adımlar gerçekleştirilmiştir:

- Eğitim setinde (Training set) Tablo 5.8 de en alt satırda sayısı belirtilen kadar alt çizge edilmiştir,
- Elde edilen alt çizgelerden frekansı en yüksek 20 tanesi belirlenmiştir,
- Frekansı belirlenen alt çizgelerin hangilerinin tüm 142 örnekte var olup olmadığı incelenerek veri kümesi tekrar düzenlenmiştir,
- Düzenlenen yeni veri kümesine Naive Bayes ve k-NN uygulanarak doğru sınıflandırma oranları (accuracy) hesaplanmıştır.

Tablo 5.8: 10-kat apraz gereklik iin verilerin daėılımı.

	1.İTERASYON	2.İTERASYON	3.İTERASYON	4.İTERASYON
14	TEST KÜMESİ	EĐİTİM KÜMESİ	EĐİTİM KÜMESİ	EĐİTİM KÜMESİ
14	EĐİTİM KÜMESİ	TEST KÜMESİ	TEST KÜMESİ	TEST KÜMESİ
14		EĐİTİM KÜMESİ	EĐİTİM KÜMESİ	EĐİTİM KÜMESİ
14			EĐİTİM KÜMESİ	EĐİTİM KÜMESİ
14			EĐİTİM KÜMESİ	EĐİTİM KÜMESİ
14			EĐİTİM KÜMESİ	EĐİTİM KÜMESİ
14			EĐİTİM KÜMESİ	EĐİTİM KÜMESİ
14			EĐİTİM KÜMESİ	EĐİTİM KÜMESİ
16			EĐİTİM KÜMESİ	EĐİTİM KÜMESİ
	76 ADET ALT ÇİZGE ELDE EDİLDİ.	79 ALT ÇİZGE ELDE EDİLDİ.	81 ALT ÇİZGE ELDE EDİLDİ.	72 ALT ÇİZGE ELDE EDİLDİ.

10-kat apraz gerekliėin uygulandıėı veri kümesinde Naive Bayes ve k-NN sınıflandırma metotlarında 4 iterasyon iin elde edilen sonuçlar Tablo 5.9 ve Tablo 5.10’da belirtilmiřtir.

Tablo 5.9: 10-kat apraz gereklik uygulandıktan sonra Naive Bayes sınıflandırıcı iin elde edilen deėerler.

İLK DEėER	84 Çizge	%92,95
1.İTERASYON	76 Çizge	%87,32
2.İTERASYON	79 Çizge	%90,84
3.İTERASYON	81 Çizge	%92,25
4.İTERASYON	72 Çizge	%87,32
ORTALAMA		%89,43

Tablo 5.10: 10-kat apraz gereklik uygulandıktan sonra k-NN sınıflandırıcı iin elde edilen deėerler.

İLK DEėER	84 ÇİZGE	%81,25
1.İTERASYON	76 ÇİZGE	%75
2.İTERASYON	79 ÇİZGE	%79,16
3.İTERASYON	81 ÇİZGE	%81,25
4.İTERASYON	72 ÇİZGE	%75
ORTALAMA		%77,6

Her iki sınıflandırma metodunda da gözlemlenen performans artışı, tez alışmasının genel motivasyonunu oluřturan hipotezi doėrular niteliktedir.

Tekrarlayan alt-çizge madenciliği ve makine öğrenmesi metotlarının bir arada kullanılması, akademik başarının modellenmesi problemi için olumlu sonuçlar doğurmuştur.

Sonuçta elde edilen başarıda veri kümesinin zenginliğinin oldukça büyük payı olduğu düşünülmektedir. Literatürde ilgili alanda yapılan çalışmalar incelendiğinde dikkat çeken, tek bir bakış açısı ile akademik başarı üzerine modeller geliştirmek yaklaşımı bu çalışmada benimsenmemiştir. Aksine, başarının yalnızca sosyo-demografik özelliklere ya da yalnızca akademik geçmişe göre değerlendirilmesinin yanıltıcı sonuçlar verebileceği ihtimaline karşın, her iki veri kümesi de bir araya getirilmeye çalışılmıştır.

Akademik geçmiş verilerinde ise yalnızca alınan notların değerlendirilmesi yerine daha önce öğrenci tarafından tamamlanmış derslerin içeriklerinden de faydalanılarak dersler arasında ilişkisel bağlar kurulmak hedeflenmiştir.

Son olarak tüm bu elde edilen veriler, veri madenciliğinin bir alt dalı olan çizge madenciliği metotlarıyla elde edilen çizge örüntüleri ile de zenginleştirilerek, veri kümesinin sayıda azlığına rağmen geleneksel makine öğrenmesi sınıflandırma metotlarının performansının artırılması sağlanmıştır.

6. SONUÇLAR ve ÖNERİLER

Veri madenciliği, veri setlerinde var olan ve tahmin edilemeyen örüntüleri keşfetme yeteneğine sahip otomatik veya yarı otomatik bir süreçtir. Bu süreç sonucunda elde edilen örüntülerin anlamlı ve fayda sağlayıcı özellikte olması beklenir. Örüntülerin elde edilmesi sürecinde makine öğrenmesi, örüntü tanıma, istatistik, veri tabanı ve çeşitli görselleştirme teknikleri bir araya getirilir.

Eğitsel veriler üzerinde uygulanan veri madenciliği yöntemleri ile öğrencilerin eğitim süreçlerinde karşılaştıkları derslere yönelik başarı/başarısızlık durumlarının önceden tahmini, kişiye özgü öğrenim planlarının yaratılması gibi problemlere çözümler üretilebilmektedir.

Literatürde özellikle tez çalışmasına konu olan öğrenci modellemesi konusunda gerçekleştirilmiş birçok çalışmaya rastlanmaktadır. Araştırmacılar bu konuda yaptıkları çalışmalarda genellikle öğrenci sayısı fazla, özellik vektörleri küçük ve yalnızca tek bir tarafa odaklanmış sığ bir bakış açısı benimsenmiş olan veri kümeleri kullanmışlardır.

Gerçek hayatta ise öğrencilerin sosyal yaşantılarının, akademik alt yapılarının, aile yaşantılarının, v.b. kriterlerin akademik başarılarına olumlu veya olumsuz etkilerinin olduğu açıkça gözlenmektedir. Bu fikirden yola çıkarak, tek bir bakış açısı yerine öğrencilerin sosyal, demografik ve akademik alt yapıları ile ilgili birçok farklı özelliği bir araya getirilerek oluşturulacak bir veri kümesi üzerinde daha başarılı bir modelleme yapılabileceği ön görülmüştür. Burada akademik altyapıları değerlendirilirken ise yalnızca başarı puanlarının değil ders içeriklerinin eklenmesi ile de öğrencinin ilgi alanları ve derslerin birbirleri ile ilişkileri hakkında bilgi sağlayabileceği ön görülmüştür.

Farklı bakış açılarından birçok özelliği içinde barındıran bir özellik vektörüne sahip olan bu çalışmada özellikle öğrenci sayısının azlığı sınıflandırma başarısında istenilen sonuçlara ulaşılmasını engellemiştir. Bu sebeple daha başarılı modeller oluşturabilmek hedefi ile tekrarlayan alt-çizgelerin de veri kümesine eklenmesi ile ilişkisel verilerin çoğaltılması, böylelikle bir anlamda başarı ve başarısızlık durumlarında tekrarlayan örüntülerle sistemin daha anlamsal çalışması sağlanmıştır.

Böylelikle veri sayısı az uygulamalarda da sınıflandırma başarısının, başarının karakteristiği hakkında bilgi veren örüntülerle zenginleştirilen özellik vektörleri ile

yükseltilebileceđi görölmüştür. Ayrıca literatürde yapılan incelemelerde daha önce eğitimsel veri madenciliđi uygulamalarında çizge madenciliđi yöntemlerine başvurulmadıđı görölmüştür. Bu anlamda da eğitimsel veri madenciliđi alanında yenilikçi bir yaklaşım getirilmiştir.

KAYNAKLAR

- [1] Peña-Ayala A., (2014), "Educational data mining: A survey and a data mining-based analysis of recent works", *Expert systems with applications*, 41(4), 1432-1462.
- [2] Dragičević M., Bach, M. P., Šimičević V., (2014), "Improving University Operations with Data Mining: Predicting Student Performance", *World Academy of Science, Engineering and Technology, International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering*, 8(4), 1101-1106.
- [3] Trstenjak B., Donko D., (2014), "Determining the impact of demographic features in predicting student success in Croatia", *37th Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 1222-1227, Opatija, Croatia, 26-30 May.
- [4] Vlahos G. E., Ferratt T. W., Knoepfle G., (2004). "The use of computer-based information systems by German managers to support decision making", *Information & Management*, 41(6), 763-779.
- [5] Larose D. T., (2014), "Discovering knowledge in data: an introduction to data mining", 2nd. Edition, John Wiley & Sons.
- [6] Roiger R. J., (2017), "Data mining: A tutorial-based primer", 2nd. Edition, CRC Press.
- [7] Han J., Pei J., Kamber M., (2011), "Data mining: concepts and techniques", 3rd. Edition, Elsevier.
- [8] Baker R. S. J. D., (2010), "Data mining for education", *International encyclopedia of education*, 7(3), 112-118.
- [9] Koedinger K., Cunningham K., Skogsholm A., Leber B., (2008), "An open repository and analysis tools for fine-grained, longitudinal learner data", In *Educational Data Mining 2008*, 157.
- [10] Castro F., Vellido A., Nebot À., Mugica F., (2007), "Applying data mining techniques to e-learning problems", *Evolution of teaching and learning paradigms in intelligent environment*, 62, 183-221.
- [11] Romero C., Ventura S., (2010), "Educational data mining: a review of the state of the art", *IEEE Transactions on Systems, Man, and Cybernetics Part C (Applications and Reviews)*, 40(6), 601-618.
- [12] Holder L. B., (2009), "Encyclopedia of Data Warehousing and Mining", Second Edition, IGI Global.

- [13] Džeroski Sašo., (2010), "Data Mining and Knowledge Discovery Handbook", 887-911.
- [14] Kuramochi M., Karypis G., (2001), "Frequent subgraph discovery", IEEE International Conference on Data Mining (ICDM), 313-320, San Jose, CA, USA, 29 Nov.-2 Dec. 2001.
- [15] Yılmaz,B., (2010), "Çizge Tabanlı Örneklerde Kullanıcı Destekli Veri Madenciliği", Doktora Tezi, Gebze Yüksek Teknoloji Enstitüsü.
- [16] Jiang C., Coenen F., Zito M., (2013), "A survey of frequent subgraph mining algorithms", The Knowledge Engineering Review, 28(1), 75-105.
- [17] Han J., Pei J., Kamber M., (2006), "Data mining: concepts and techniques", 2nd. Edition, Elsevier.
- [18] Gonzalez J., Jonyer I., Holder L. B., Cook D. J., (2000), "Efficient mining of graph-based data", Technical Report No: WS-00-06, Department of Computer Science and Engineering, University of Texas at Arlington, USA.
- [19] Kuramochi M., Karypis G., (2004), "An efficient algorithm for discovering frequent subgraphs", Technical Report No: TR 02-026, Department of Computer Science and Engineering, University of Minnesota, USA.
- [20] Chen C., Yan X., Zhu F., Han J., (2007), "gapprox: Mining frequent approximate patterns from a massive network", Seventh IEEE International Conference(ICDM), 445-450.
- [21] Zhang S., Yang J., (2008), "Ram: Randomized approximate graph mining", International Conference on Scientific and Statistical Database Management, 187-203, Hong Kong, China, July 9-11.
- [22] Krishna V., Suri N. R., Athithan G., (2011), "A comparative survey of algorithms for frequent subgraph discovery", Current Science, 100(2), 190-198.
- [23] Nijssen S., Kok J., (2001), "Faster association rules for multiple relations", In International Joint Conference on Artificial Intelligence, 891-896, Seattle, Washington, USA, August 4th – 10th.
- [24] Kuramochi M., Karypis G., (2005), "Finding frequent patterns in a large sparse graph", Data mining and knowledge discovery, 11(3), 243-271.
- [25] Yan X., Han J., (2002), "gspan: Graph-based substructure pattern mining", 2002 IEEE International Conference on Data Mining, 721-724, Maebashi City, Japan, 9-12 Dec.
- [26] Yan X., Han J., (2003), "CloseGraph: mining closed frequent graph patterns", Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, 286-295, Washington, DC, USA, August 24 – 27.

- [27] Huan J., Wang W., Prins J., (2003), "Efficient mining of frequent subgraphs in the presence of isomorphism", Third IEEE International Conference, 549-552, Melbourne, FL, USA, 22-22 Nov.
- [28] Nijssen S., Kok J. N., (2004), "A quickstart in frequent structure mining can make a difference", Proceedings of the tenth ACM SIGKDD International Conference on Knowledge discovery and data mining, 647-652, Seattle, WA, USA, August 22 – 25.
- [29] Thomas L., Valluri S., Karlapalem K., (2009), "Isg: Itemset based subgraph mining", Technical Report No: IIIT/TR/2009/179, Centre for Data Engineering, International Institute of Information Technology, INDIA.
- [30] Kuramochi M., Karypis G., (2004), "Grew-a scalable frequent subgraph discovery algorithm", Fourth IEEE International Conference on Data Mining, 439-442, Brighton, UK, 1-4 Nov.
- [31] Xu J., Moon K. H., van der Schaar M., (2017), "A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs", IEEE Journal of Selected Topics in Signal Processing, 10(5), 742 – 753.
- [32] Marquez-Vera C., Romero C., Ventura S., (2010), "Predicting school failure using data mining", 4 th International Conference on Educational Data Mining, 271, Eindhoven, the Netherlands, July 6-8.
- [33] Chan C. C., (2007), "A framework for assessing usage of web-based e-learning systems", Second International Conference on Innovative Computing, Information and Control, 147, Kumamoto, Japan, 5-7 Sept.
- [34] Thomas E. H., Galambos N., (2004), "What satisfies students? Mining student-opinion data with regression and decision tree analysis", Research in Higher Education, 45(3), 251-269.
- [35] Dekker G., Pechenizkiy M., Vleeshouwers J., (2009), "Predicting students drop out: A case study", 2nd. International Conference On Educational Data Mining, 41, Cordoba, Spain, July 1-3.
- [36] Bresfelean, V. P., Bresfelean M., Ghisoiu N., Comes C. A. (2008). "Determining students' academic failure profile founded on data mining methods", 30th International Conference on Information Technology Interfaces, 317-322, Dubrovnik, Croatia, 23-26 June.
- [37] Calvo-Flores M. D., Galindo E. G., Jiménez M. P., Piñeiro O. P., (2006), "Predicting students' marks from Moodle logs using neural network models", Current Developments in Technology-Assisted Education, 1(2), 586-590.
- [38] Ibrahim Z., Rusli D., (2007), "Predicting students' academic performance: comparing artificial neural network, decision tree and linear regression", 21st Annual SAS Malaysia Forum, 1-6, Kuala Lumpur, Malezya, 5th Sept.

- [39] Wang Y. H., Liao H. C., (2011), “Data mining for adaptive learning in a TESL-based e-learning system”, *Expert Systems with Applications*, 38(6), 6480-6485.
- [40] Kotsiantis S. B., Pintelas P. E., (2005), “Predicting students marks in hellenic open university”, *Fifth IEEE International Conference on Advanced Learning Technologies*, 664-668, Kaohsiung, Taiwan, 5-8 July.
- [41] Rus V., Lintean M., Azevedo R., (2009), “Automatic Detection of Student Mental Models during Prior Knowledge Activation in MetaTutor”, *2nd. International Conference on Educational Data Mining*, 161, Cordoba, Spain, Jul 1-3.
- [42] Trstenjak B., Donko D., (2014), “Determining the impact of demographic features in predicting student success in Croatia”, *37th International Convention on Information and Communication Technology, Electronics and Microelectronics*, 1222-1227, Opatija, Croatia, 26-30 May.
- [43] Osmanbegović E., Suljić M., (2012), “Data mining approach for predicting student performance”, *Economic Review*, 10(1), 3-12.
- [44] Kabakchieva D., (2013), “Predicting student performance by using data mining methods for classification”, *Cybernetics and Information Technologies*, 13(1), 61-72.
- [45] Kotsiantis S., Patriarcheas K., Xenos M., (2010), “A combinational incremental ensemble of classifiers as a technique for predicting students’ performance in distance education”, *Knowledge-Based Systems*, 23(6), 529-535.
- [46] Tanner T., Toivonen H., (2010), “Predicting and preventing student failure—using the k-nearest neighbour method to predict student performance in an online course environment”, *International Journal of Learning Technology*, 5(4), 356-377.
- [47] Harrington P., (2012), “Machine learning in action”, 5th. Edition, Greenwich, CT: Manning.
- [48] Rish I., (2001), “An empirical study of the naive Bayes classifier”, In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, 3(22), 41-46.
- [49] Mitchell T. M., (1997), “Machine learning”, Burr Ridge, IL: McGraw Hill, 45(37), 870-877.
- [50] Vapnik V., (2013), “The nature of statistical learning theory”, 2nd. Edition, Springer science & business media.
- [51] Yu H., Kim S., (2012), “Svm tutorial—classification, regression and ranking”, In *Handbook of Natural computing*, Springer Berlin Heidelberg, 479-506.
- [52] Kotsiantis S. B., Pierrakeas C. J., Pintelas P. E., (2003), “Preventing student dropout in distance learning using machine learning techniques”, *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, 267-274, Oxford, UK, 3-5 Sept.

- [53] Quevedo J. R., Montañés E., (2009), "Obtaining Rubric Weights for Assessments by More than One Lecturer Using a Pairwise Learning Model", 2nd. International Conference on Educational Data Mining, 289, Cordoba, Spain, Jul 1-3.
- [54] Ranjan J., Khalil S., (2008), "Conceptual framework of data mining process in management education in India: An institutional perspective", *Information Technology Journal*, 7(1), 16-23.
- [55] Guo Q., Zhang M., (2009), "Implement web learning environment based on data mining", *Knowledge-Based Systems*, 22(6), 439-442.
- [56] Wang Y. H., Tseng M. H., Liao H. C., (2009), "Data mining for adaptive learning sequence in English language instruction", *Expert Systems with Applications*, 36(4), 7681-7686.
- [57] Chan C. C., (2007), "A framework for assessing usage of web-based e-learning systems", *Second International Conference on Innovative Computing, Information and Control*, 147, Kumamoto, Japan, 5-7 Sept.
- [58] Thomas E. H., Galambos N., (2004), "What satisfies students? Mining student-opinion data with regression and decision tree analysis", *Research in Higher Education*, 45(3), 251-269.
- [59] Dekker G., Pechenizkiy M., Vleeshouwers J., (2009), "Predicting students drop out: A case study", *2nd International Conference On Educational Data Mining*, 41-50, Cordoba, Spain, July 1-3.
- [60] Superby J. F., Vandamme J. P., Meskens N., (2006), "Determination of factors influencing the achievement of the first-year university students using data mining methods", *Workshop on Educational Data Mining*, 32, 234.
- [61] Lykourantzou I., Giannoukos I., Nikolopoulos V., Mpardis G., Loumos V., (2009), "Dropout prediction in e-learning courses through the combination of machine learning technique", *Computers & Education*, 53(3), 950-965.
- [62] Wang T., Mitrovic A., (2002), "Using neural networks to predict student's performance", *International Conference on Computers in Education*, 969-973, Auckland, New Zealand, 3-6 Dec.
- [63] Doshi M., (2014), "Correlation Based Feature Selection (Cfs) Technique To Predict Student Performance", *International Journal of Computer Networks & Communications*, 6(3), 197.
- [64] Cook D. J., Holder L. B. (Eds.), (2006), "Mining graph data", 1st. Edition, John Wiley & Sons.
- [65] Holder L. B., Cook D. J., Djoko S., (1994), "Substructure Discovery in the SUBDUE System", *Workshop on Knowledge Discovery in Databases*, 169-180, Seattle, WA, July 31-Aug 1.

- [66] Zhao X., Chen Y., Xiao C., Ishikawa Y., Tang J., (2016), “Frequent Subgraph Mining Based on Pregel”, The Computer Journal, 59(8), 1113-1128.
- [67] Talukder N., Zaki M. J., (2016), “A distributed approach for graph mining in massive networks”, Data Mining and Knowledge Discovery, 30(5), 1024-1052.
- [68] Web 1, (2017), <http://ailab.wsu.edu/subdue/>, (Erişim Tarihi : 04/03/2017).
- [69] AL-HUDAIRY H. H. M. A. A., (2004), “Data mining and decision making support in the governmental sector”, Doktora Tezi, University of Louisville.

ÖZGEÇMİŞ

Başak BULUZ 1990 yılında İstanbul’da doğdu. 2008 yılında %50 başarı bursu ile başladığı İstanbul Aydın Üniversitesi Fen-Edebiyat Fakültesi Matematik-Bilgisayar Bölümünü ve yüksek onur öğrencisi olması sebebiyle 2010 yılında çift anadal programını yapmaya hak kazandığı İstanbul Aydın Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği(İngilizce) Bölümünü 2013 yılında onur öğrencisi olarak tamamladı. Aynı zamanda 2008 yılında başladığı Anadolu Üniversitesi İşletme Fakültesi İşletme Bölümünü de 2015 yılında başarıyla tamamladı.

2014 yılında yüksek lisans eğitimine Gebze Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalında başladı. 2013 yılından bu yana İstanbul Aydın Üniversitesi Anadolu Bil Meslek Yüksekokulu Bilgisayar Programcılığı (İngilizce) Programında öğretim görevlisi ve 2014 yılından bu yana aynı programda program başkanı olarak görev almaktadır.