T.C.
BEYKENT ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ BİLİM DALI

LİSE ÖĞRENCİLERİNİN ÜNİVERSİTEYE GİRİŞ BAŞARILARININ EĞİTSEL VERİ MADENCİLİĞİ İLE TAHMİN EDİLMESİ

Yüksek Lisans Tezi

Tezi Hazırlayan:

Sakın CAN

İstanbul, 2021

T.C.
BEYKENT ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ BİLİM DALI

LİSE ÖĞRENCİLERİNİN ÜNİVERSİTEYE GİRİŞ BAŞARILARININ EĞİTSEL VERİ MADENCİLİĞİ İLE TAHMİN EDİLMESİ

Yüksek Lisans Tezi

Tezi Hazırlayan:

Sakın CAN

Öğrenci No:

17080200036

ORCID ID

0000-0001-6968-1156

Danışman:
Dr. Öğr. Üyesi Zeynep ALTAN

İstanbul, 2021

YEMIN METNI

Yüksek Lisans Tezi olarak sunduğum "Lise Öğrencilerinin Üniversiteye Giriş Başarılarının Eğitsel Veri Madenciliği ile Tahmin Edilmesi" başlıklı bu çalışmanın, bilimsel ahlak ve geleneklere uygun şekilde tarafımdan yazıldığını, yararlandığım eserlerin tamamının kaynaklarda gösterildiğini ve çalışmanın içinde kullanıldıkları her yerde bunlara atıf yapıldığını belirtir ve bunu onurumla doğrularım. **27/04/2021**

Sakın CAN

T.C.
BEYKENT ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ MÜDÜRLÜĞÜ
TEZLİ YÜKSEK LİSANS SINAV TUTANAĞI

..../..../.....

Enstitümüz ***Bilgisayar Mühendisliği*** Anabilim Dalı ***Bilgisayar Mühendisliği*** Programı Yüksek Lisans öğrencilerinden ***17080200036*** numaralı ***Sakın CAN***'ın "Beykent Üniversitesi Eğitim-Öğretim Yönetmeliği"nin ilgili maddesine göre hazırlayarak Enstitümüze teslim ettiği ***"Lise Öğrencilerinin Üniversite Giriş Başarılarının Eğitsel Veri Madenciliği İle Tahmini"*** konulu tezini, Yönetim Kurulumuzun 30/03/2021 tarih ve 2021/10 sayılı toplantısında seçilen ve On-line toplanan biz jüri üyeleri huzurunda, Beykent Üniversitesi Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 29. maddesinin 3. fıkrası gereğince 45 dakika süre ile Microsoft Teams programı aracılığıyla on-line olarak aday tarafından savunulmuş ve sonuçta adayın tezi hakkında ***"OYBİRLİĞİ"*** ile ***"KABUL"*** kararı verilmiştir.

İşbu tutanak, 2 nüsha olarak hazırlanmış ve Enstitü Müdürlüğü'ne sunulmak üzere tarafımızdan düzenlenmiştir.

DANIŞMAN
Dr. Öğr. Üyesi Ze*** AL***
(Beykent Üniversitesi)

ÜYE
Dr. Öğr. Üyesi Ed** ŞA***
(Beykent Üniversitesi)

ÜYE
Dr. Öğr. Üyesi Se*** AY***
(Bahçeşehir Üniversitesi)

Adı ve Soyadı :Sakın CAN
Danışmanı : Dr. Öğr. Üyesi Zeynep ALTAN
Türü ve Tarihi : Yüksek Lisans, 2021
Alanı : Bilgisayar Mühendisliği
Anahtar Kelimeler : Eğitsel Veri Madenciliği, Üniversite Başarı Tahminleri, Makine Öğrenmesi, Python

ÖZ

LİSE ÖĞRENCİLERİNİN ÜNİVERSİTEYE GİRİŞ BAŞARILARININ EĞİTSEL VERİ MADENCİLİĞİ İLE TAHMİNİ

Eğitimin çok önemli olduğu günümüz Türkiye'sinde, birçok üniversite bölümü ile lisans eğitimi seçeneği sunulmaktadır. Üniversite sınavlarında başarılı olmak için birden fazla etken söz konusudur. Eğitim alanında devam eden araştırmalara ve çalışmalara katkı sağlamak amacıyla üniversite sınavına katılan kişilerden toplanan veriler ile veri madenciliği ve makine öğrenmesi algoritmaları kullanılarak sınavda başarı oranının tahmini yapılmıştır.

1979-2020 yılları arasında üniversite sınavına katılan 677 kişinin katıldığı anket verilerinden yola çıkılarak üniversiteye yerleştirilmeye hak kazanılmasının veriler aracılığı ile tahmin edilmesine yönelik veri madenciliği makine öğrenmesi algoritmalarından en verimli sonuçları veren karar ağaçları sınıflandırması, doğrusal regresyon, rastgele orman algoritması, destek vektör makineleri, K-en yakın komşu algoritması ve Gaussian NB algoritması kullanılmıştır

Veri setinde katılımcıların; cinsiyeti, okuduğu alanın ders notları, okuduğu okul türü, özel ders bilgisi, kaynaklarına ulaşım kolaylıkları, ders çalışma saatleri, sınıf mevcudu ve tüm değişkenler eğitimsel değişken olarak belirlenmiştir.

Seçilen değişkenlere göre veri setleri oluşturulup öğrenme ve tahmin yapılarak algoritmaların sonuç üzerine etkisi değerlendirildiğinde Gaussian NB sınıflandırma modeli %73.77 ‘lik bir doğruluk oranı ile en yüksek tahmini apmıştır. Aynı zamanda sınavı kazanma ile ilgili önemli değişkenler hakkında bilgi edinilmiştir.

Name and Surname : Sakın CAN
Supervisor : Dr. Lecturer Zeynep ALTAN
Degree and Date : Master, 2021
Major : Computer Engineering
Key Words : Educational Data Mining, University Success Predictions, Machine Learning, Python

ABSTRACT

ESTIMATION OF HIGH SCHOOL STUDENTS 'SUCCESS OF ENTERENCE TO UNIVERSITY WITH EDUCATIONAL DATA MINING

In today's Turkiye where education is essential, there are so many Universities provide undergraduate education. There are multiple factors to be successful in university exams. In order to contribute ongoing research and studies in the field of education, "the success rate in the exams" are estimated using data mining and machine learning algorithms with the data collected from the people who took the university exam.

The study is based on the survey data of 672 participants who participated in the university exam between 1979-2020. Based on the data the most effective data mining and machine learning algorithms such as "decision classification", "linear regression", "orbit forest algorithm", "support vector properties", "K-N Nearest neighbor algorithm" and "Gaussian NB algorithm" are used to predict the "the success rate in the exams".

In the data set, educational variables are set as the participants'; gender, course grades of the field of study, school type, getting private lesson, easy access to education resources, studying hours, class population.

After data sets were created according to the selected variables and the effect of algorithms on the results were evaluated by learning and predicting, "the Gaussian NB classification model" made the highest estimation with an accuracy rate of 73.77% At the same time, we obtained valuable information about the variables of "passing the exam".

İÇİNDEKİLER

Sayfa No.

TABLOLAR LİSTESİ

Sayfa No.

ŞEKİLLER LİSTESİ

Sayfa No.

KISALTMALAR

AUC **:** Area Under the Curve (ROC Eğrisi Altında Kalan Alan)

AYT **:** Alan Yeterlilik Testleri

CRISP-DM **:** The Cross-Industry Standard Process for Data Mining(Sektörler Arası Standart Süreç Modeli-Veri Madenciliği)

EVM **:** Eğitsel Veri Madenciliği

KDD **:** Knowledge and Data Discovery (Bilgi Keşfi)

KNN **:** K Nearest Neighbors **(**K-En Yakın Komşu Algoritması)

NB **:** Gaiussun Naive Bayes Algoritması

PISA **:** Programme for International Student Assessment (Uluslararası Öğrenci Değerlendirme Programı)

ROC : Receiver Operating Characteristic (Alıcı İşletim Karakteristiği)

STD **:** Standart

SVM **:** Support Vector Machines (Destek Vektör Makineleri)

Ort **:** Ortalama

ÖSYM **:** Öğrenci Seçme Yerleştirme Merkezi

VM **:** Veri madenciliği

TYT **:** Temel Yeterlilik Testi

YKS **:** Yükseköğretim Kurumları Sınavı

GİRİŞ

Günümüz dünyasında eğitim ve veri bilimi uzmanları eğitim verileriyle ilgili birçok araştırma yapmaktadır. Teknoloji bize çoğu zaman avantajlı durumlar sergilese de bazen de bizi dezavantajlı durumlarla karşı karşıya bırakmaktadır. Bireyler ve kurumlar doğru bilgiye kısa sürede kolay yoldan ulaşmak istemektedir. Verilerin hızla artışı, büyüme hacmi insanların çalışmalarında kendi başına altından kalkabileceği bir iş değildir. Bu işin altından kalkabilecek verileri işleyip yararlı hale getirebilecek disiplinin adı veri madenciliğidir. Veri madenciliği, anlamsız görünen bilgilerin anlamlı hale getirilmesidir. Doğru bilgiye ulaşma ve bu bilgilerden faydalanma ihtiyacı son yıllarda eğitim sektöründe de aranılan bir ihtiyaç haline geldi. Bu çalışmada, eğitim verileri kullanılarak veri madenciliği ile bilgilerin analizi yapılıp eğitsel veri madenciliği oluşturulmuştur. Eğitsel veri madenciliği son zamanlarda popüler olması nedeni ile hızla geliştirilmiş ve eğitimdeki verileri işleyip seçenek sunması izlenecek bir yol haritası çıkarması, büyük bir ihtiyaca çözüm oluşturmuştur. Amaç eğitimde başarı oranını arttırarak bilinçli bir eğitim farkındalığı oluşturmaktır. Bilinçli eğitimin getirisi eğitimde kaliteyi doğru oranda yükseltecektir. Eğitsel veri madenciliği ile ilgili bir metafor kullanmak gerekseydi buna rahatlıkla eğitimin navigasyonu denilebilirdi. Bu çalışmadaki amaç üniversite sınav başarısının, lise öğrenimi boyunca etkili olan etkenler göz önünde bulundurularak veri madenciliği ve makine öğrenmesi yöntemleri kullanılarak veriler üzerinde tahminleme yapılamasını sağlamak ve mevcut sistemin bazı sorunlarına çözüm yolu aramaktır. Makine öğrenmesi bilgisayar öğrenmesini ele alan bir disiplindir. Veri madenciliği ve derin öğrenme gibi disiplinlerden beslenmektedir. Eğitim verileri ile makine öğrenmesi yöntemlerinden faydalanılarak, oluşabilecek durumların tahmininin yapılması eğitimle teknolojinin entegre çalışmasını sağlayabilecektir.

Eğitim sistemimizde ortaokuldan itibaren sınavlarla üst kademedeki eğitime geçiş yapılabilmektedir. Liseyi bitiren öğrenci de bir üst kademedeki eğitimi kazanmak için üniversite sınavına giriyor. Baz alınan 2019 yılı verilerine Yükseköğretim Kurumları Sınavı (YKS) istatistikleri incelendiğinde sınavın ilk basamağı olan Temel Yeterlilik Testi (TYT) sınavına başvuru yapan 2 milyon 515 bin 12 öğrenci adayından 2 milyon 390 bin 491'i üniversite sınavına katıldığı görülmüştür.

%74,16'sının 150 ve üzeri puan alarak ön lisans tercih etme barajını geçtiği, %53,72'sinin 180 ve üzeri puan alıp 2. basamak olan Alan Yeterlilik Testleri (AYT) sınavlarına katılma barajını geçtiği gözlenmiştir. Bu değerler 2018 istatistikleri ile karşılaştırıldığında barajı aşan öğrenci sayısının düşmüş olmuştur. Dört yıl boyunca lise okuyan öğrencilerin sınavları başarısızlıkla sonuçlanabilmektedir. ÖSYM, sınavlardan sonra sınava girmiş öğrencilerin okul, bölüm, il ve cinsiyet bazlı kazanma oranlarını bizlere sunmaktadır. Bize sunulan bu bilgiler daha detaylı inceleme yapabilmek için yeterli değildir. Neden sonuç ilişkisini analiz etmememize yetmemektedir. Bu yüzden üniversite sınavını kazanmak için okul, bölüm, il ve cinsiyet birer faktör olsa da belirleyici değişkenler değildir. Sınavı kazanma başarısına etki eden sosyo-demografik faktörler de incelenmiştir. Bunlar, cinsiyet, bölüm, ders notları, okul türü, özel ders, kaynak erişim kolaylığı, ders çalışma saati, dijital öğrenme platformu kullanımı, yeterli hissetme, çalışma durumu, akran zorbalığı, danışmanlık bilgisi, doğru seçim inancı, sınıf mevcudu, diploma notu, sınava giriş tarihi, öğrenim ili, sosyal hesap kullanımı, sağlık problemi, aile birlikteliği, ailenin hayatta olup olmama durumu, aile fertlerinin çalışma durumu, ailenin gelir düzeyi, ailenin eğitim düzeyi, kardeş miktarı, yükseköğretimde kardeş olma durumu, lisedeki üniversite etkinliklerine katılma durumu, ön lisans geçiş hakkı, üniversite giriş sınavı baraj geçme durumu, tercih yapma hakkı ve varsa kazanılan üniversite bilgisi ele alınmıştır. . Çalışmada öğrencinin, eğitim hayatını etkileyen diğer pek çok faktör araştırılmıştır. Bu çalışmada gelişmemiş bölgelerde akademik başarıları etkileyen faktörlerin hem okul kaynaklı hem de aile kaynaklı olduğu görülmüştür. Genel olarak öğrenci başarısını etkileyen en önemli faktör olarak ise öğrencinin not durumu, çalışıp çalışmadığı, cinsiyeti, ailenin desteği, okul etkinlikleri, öğretmen tutumu, öğrencinin eğitim algısı ve arkadaş sayısı olarak değerlendirmiştir. Ayrıca sosyo-demografik (ailenin yapısı, büyüklüğü, eğitimi vb.) etkenler ile okul etkenleri (öğretmen-öğrenci mevcudu, sınıf hacmi vb.) dışında çevresi, gelenekleri, dini, okulun fiziki yapısı başarıya etki eden faktörler olarak değerlendirilmektedir.

Bu çalışmada gençlerin üniversite sınavlarındaki başarı değerlendirmelerini yapmak üzere geniş kapsamlı bir anket çalışması yapılmıştır. Sınavdaki başarı durumlarını değerlendirmek üzere öz niteliklerin anket verileriyle desteklendiği bir tahminleme sistemi oluşturulup, veriler analiz edilmiştir öz nitelikler arasında sosyo-

demografik faktörleri içeren değişkenler de yer almaktadır. Bu değişkenler oluşturulurken üniversite sınavı başarı oranını inceleyen eğitimle ilgili çalışmaların sonuçlarından, eğitmenlerin fikirlerinden yararlanılmış ve veri setleri oluşturulmuştur. Veri setleri 1979-2020 yılları arasında üniversite sınavına katılan 416'sı kadın, 261'i erkek olan 677 katılımcının olduğu anket verilerinden oluşturulmuştur. Anket Google anket araçları ile oluşturulup online soru- cevap şeklinde olmuştur. Verilerin analizi ve sınıflandırma süreçlerinde KDD (Knowledge and Data Discovery) süreç modeli kullanılmıştır. KDD, veriyi değerli ve bir amaca hizmet edecek hale getirme sürecidir diye tanımlanabilir. Üniversite sınavındaki başarı durumu tahminine yönelik kullanılan veri madenciliği ve makine öğrenmesi algoritmaları içinden, karar ağaçları sınıflandırması, lojistik regresyon, rastgele orman algoritması, destek vektör makineleri, Gaussian NB algoritması ve K-en yakın komşu algoritması yöntemleri kullanılmıştır. Bu yöntemlerin seçilmesinin nedeni denenen diğer yöntemlere göre daha yüksek sonuçlar elde edilmesidir.

Çalışmadaki anket verilerine ulaşmak için Google anket kullanılmıştır ve anket verileri csv dosyasına çevrilerek veri setleri oluşturulmuştur. Modelleme, eğitim, tahmin ve tahmin sonuçlarının birleştirmek için platform olarak Anaconda seçilmiş, Python yazılımı ve kütüphaneleri kullanılmıştır. Python açık kaynak kodlu, platform bağımsız olduğu için tercih edilmiştir. Veri seti %75'i eğitim, %25'i tahmin için belirlenmiştir. Veri setini ayırma ve seçme işlemleri dışarıda tutma (holdout) ve k-katlı çapraz doğrulama ile yapılmıştır. Öznelik seçiminde değişken seçimi (feature selection) filtreleme yöntemi, anova yöntemi ve ki-kare yönteminden faydalanılmıştır. Belirlenen algoritmalar ile oluşturulan modellerde veri setindeki sonuçlar için tahmin yapılmıştır. Tahmin hedefi üniversiteyi kazanma durumudur. Üniversite sınavını kazanmanın tahmin edilmesinde çıkarılan sonuçlar için algoritmaların başarı oranı incelenmiştir. Yapılan analizler neticesinde Gaussian NB modelinin üniversite sınavını kazanma tahminini yapmaya ilişkin tüm farklı sorguların en başarılı tahmini üreten modelleme olduğu görülmektedir.

Çalışmanın bundan sonraki kısmında tezin kavramsal temellerini oluşturan veri madenciliği ile ilgili bilgilendirmeler yapılmıştır. 2. Bölümde ise CRISP-DM süreçlerine ait adımlar takip edilerek üniversite sınavını kazanma tahmini çalışmaları

ayrıntılandırılmıştır. Bu bölümde sınıflandırma algoritmaları kıyaslanarak modeller arasındaki en iyi tahmini yapan algoritma modeli seçilmiştir. 3. Bölümde sonuçlar değerlendirilmiştir. Burada çalışmanın tamamını kapsayan bir değerlendirme yapılarak, gelecekte yapılabilecek çalışmalara örnek olabilmesi için önerilerde bulunulmuştur.

BİRİNCİ BÖLÜM

VERİ MADENCİLİĞİ ve MAKİNE ÖĞRENMESİNE GENEL BAKIŞ

1. VERİ MADENCİLİĞİ

Dünya nüfusu gittikçe artmakta ve teknoloji kullanımı da doğru oranda ilerlemektedir. Yapılan bir işlem, her sektörde bilgi niteliğindedir. Günümüzde verileri kullanışlı hale getirebilmek ekonomiye büyük katkı sağlamaktadır. Veri madenciliği(VM), anlamsız görülebilen bilgilerin istatiksel ve matematiksel yöntemler ile anlamlı hale getirilmesidir (Özdemir, 2016). Gerek kamu kuruluşları gerekse özel sektör verilerden bir anlam ilişkisi kurulmasını isteyerek veri madenciliğinin önemini kavramıştır. VM, veri depolama araçları ve teknolojilerine bağlı olarak gelişip yayılmaktadır. Bilinmeyen bir bilgiye veriler aracılığı ile ulaşılması veri madenciliğinin temelini oluşturmaktadır. VM, astrofizikten laboratuvarlara, risk analizlerinden sahtekarlığa e-ticaretten tedarik zincirine vb. farklı araştırma alanlarına yayılmıştır. Verilerin işlendiği kaynaklar göz önüne alındığında veri madenciliğinin çok geniş bir uygulama alanı olduğu söylenebilir.

Veri madenciliğinde çok sayıda yöntem ve algoritma kullanılır. Bu yöntemler sınıflandırma, kümeleme ve birliktelik kuralları olarak gruplandırılabilir. Sınıflandırma, veri madenciliğinde en sık kullanılan yöntemlerdendir. Veri kümesindeki tanımlı sınıflara veriyi dağıtır. Verinin sınıflara ayrılması için bir süreç takip edilir, bu süreçte dağıtılan veriler sayesinde eğitim kümesinde dağılım şekli öğrenilir ve yeni veri geldiğinde eğitim kümesinden öğrenilen ile doğru şekilde test kümesi sınıflandırılmaya çalışılır. Kümeleme, sınıf içi benzeşmenin en üst düzeyde, sınıflar arası benzeşmenin en alt düzeyde olması prensibine sahip verileri kümelemektedir (Han ve diğ.,2012). Birliktelik Kuralları, durumların birlikte gerçekleşme olasılığının ortaya çıkarılmasına dayanmaktadır (Özkan, 2016). Veri madenciliğinde kullanılan yöntemler, temel düzeydeki istatistikten, zor ve karışık yapıya doğru ilerlemektedir (Tuffery, 2011).

Tablo 1.Veri Madenciliği Teknikleri ve Algoritmaları

Sınıflandırma Teknikleri ve Algoritmaları	Karar Ağaçları	ID3 Algoritması C 4.5 ve C5 Algoritması CART Algoritması SLIQ Algoritması SPRINT Algoritması Değişken Merkezli Karar Ağacı Algoritması
	İstatistiğe Dayalı Algoritmalar	Bayesyen Sınıflandırma Regresyon CHAID Algoritması
	Mesafeye Dayalı Algoritmalar	K-En Yakın Komşu Algoritması En Küçük Mesafe Sınıflandırıcısı
	Yapay Sinir Ağları	İleri Sürümlü Yapay Sinir Ağları Hata Geriye Yayma Yöntemi
Birliktelik Kuralları ve İlişki Analizi	AIS Algoritması SETM Algoritması Apriori Algoritması AprioriTid Algoritması	
Kümeleme Analizi	Hiyerarşik Yöntemler	SLINK Algoritması ve Tek Bağlantı Tekniği CURE Algoritması CHAMELEON Algoritması BIRCH Hiyerarşik Yöntemle Kategorik Verilerin Kümelenmesi
	Bölünmeli Yöntemli	K-Ortalama (K-Means) Algoritması PAM Algoritması CLARA Algoritması CLARANS Algoritması
	Yoğunluğa Dayalı Algoritmalar	DBSCAN Algoritması OPTICS Algoritması DENCLUE Algoritması
	Grid Temelli Algoritmalar	STING Algoritması Dalga Kümeleme CLIQUE Algoritması
	Genetik Algoritmalar	

Kaynak: Silahtaroğlu, Gökhan. 2013. Veri Madenciliği Kavram ve Algoritmalar (3. Basım), İstanbul; Papatya Bilim, s. 5-7.

Tablo 1'de VM teknikleri ve algoritmaları yer almaktadır. Bölüm 2.5'te kullanılan modellerle ilgili detaylı bilgi verilmiştir.

Aşağıda son yıllarda yapılan eğitsel veri madenciliği, VM, makine öğrenmesi ve sınıflandırma algoritmalarıyla ilgili kavramsal araştırmalara örnekler verilmektedir.

1.1. Veri Madenciliği Çalışmaları

Veri madenciliği literatürde tanımsal ve tahmine dayalı veri analizleri olarak değerlendirilmektedir. Aşağıda bu konudaki çalışmalar özetlenmektedir.

VM 1950'li yıllarda bilgisayarların matematiksel sayımlarında kullanılmaya başlanmıştır. Veri miktarı çoğaldıkça veri analizi ihtiyaç haline gelmiştir. Yapılan çalışmalar veri madenciliğinin o yıllardan beri popüler olduğunu kanıtlamaktadır. Verinin önemli olduğu her alanda VM kullanımı söz konusudur. Ersöz 'ün yaptığı tez çalışmasında Uludağ Üniversitesi Eğitim Fakültesi öğrencilerinin olduğu grubun veri madenciliği sınıflandırma yöntemleri ile öğrenci profillerini çıkarmıştır. Çalışma sonunda öğrenci başarılarının önem sırası öğrencinin okuduğu bölüm, dönem, geliş şekli ve cinsiyeti gibi farklı durumlar olarak çıkartılmıştır (Ersöz, 2016). 2014 yılında yapılan başka bir çalışmada ise Gray ve diğerleri veri madenciliği sınıflandırma yöntemleri kullanarak öğrencinin ilk yıllarında başarısızlık durumu incelenmiştir. Çalışmada veriler anket yoluyla alınmıştır. Farklı sınıflandırma yöntemleri kullanılarak veriler analiz edilmiştir. Kullandığı yöntemler, Bayes sınıflandırması, karar ağaçları, Lojistik regresyon, SVM, yapay sinir ağları ve k-yakınsak komşu algoritmalarıdır. Çalışma sonunda 21 yaş üstü grupta Bayes ve Lojistik regresyon tahminlemesi diğer yöntemlere göre daha düşük kalmıştır. Çalışma sonunda 21 yaş üstü verilerin 21 yaş altına göre daha zorlu yapıya sahip olduğu çıkmıştır (Gray ve diğ. ,2014). Marquez-Vera ve arkadaşlarının olduğu 2013'te yapılan diğer bir eğitimsel veri madenciliği çalışmasında öğrencinin eğitimi boyunca başarısız olabilecek ve okulu bırakabilecek kişilerin tahmini araştırılmıştır. Çalışmaya 670 öğrenci katılmıştır. Bu öğrencilerin ortaöğretim bilgileri çalışmada yer almıştır. Veri madenciliği yöntemlerinden sınıflandırma yöntemlerini kullanmıştır. Akademik başarıyı etkileyen faktörlerin neler olduğu gözlemlenmiştir. Doğru sınıflandırma kuralları oluşturabilmek için genetik programlama önerilerek öğrencinin akademik durumu ve nihai performansının tahmini yapılmıştır(Marquez ve ark., 2013).

Larose, "*Discovering Knowledge in Data: An Introduction to Data Mining*" isimli kitabında veri madenciliğini 6 basamağa ayırır. Bu adımlar problemi tanımlamakla başlar; verileri tanıma, veri hazırlama, modelleme, değerlendirme ve

uygulama aşamaları olarak devam eder. Her aşama bir sonraki aşama için önemlidir (Larose, 2005). Aşamalar Bölüm 1.3'te detaylandırılacaktır. 2018'de Villegas ve arkadaşları çalışmalarında KDD metodolojisini eğitim alanında veri madenciliği yöntemleriyle uygularlar. Bu çalışma üniversite öğrencilerinin pazarlama yöntemlerini nasıl öğrendiklerini konu eder. Araştırmada KDD 'nin hedefe yönelik adımları kolaylaştırdığından bahsedilir. Çalışma sonucunda e-öğrenme platformlarının daha fazla tercih edildiği sonucu çıkarılır (Villegas ve ark., 2018).

1.2. Eğitsel Veri Madenciliği

Eğitsel veri madenciliği son zamanlarda popüler olmasına rağmen eğitimdeki verileri işleyip seçenek sunması, başarı oranını artırması için bir yol haritası çıkarması, büyük bir ihtiyaca çözüm bulmuştur. Kapsamlı olarak açıklanırsa, veri madenciliği ve gizli yapısal kurallı zengin, geniş ve kendine özgü eğitim verilerinin buluşu gibi makine öğrenmesi yöntemlerini, potansiyel olarak birçok yapısalcı öğrenim ortamının ürettiği verilere benzeyen çeşitli yöntemlerdir (Berland ve diğerleri, 2014). Eğitsel veri madenciliği elde edilen eğitsel bilgileri inceleyen bir disiplindir. Eğitsel veri madenciliği sayesinde öğrencinin de geri bildirimlerden yararlanması sağlanmaktadır.

Öğrenim modelleri günümüzde geniş bir araştırma konusudur. Magdin, veri madenciliği yöntemlerini kullanarak, öğrencinin çalışmaları sırasında kullandığı araç-gereci kendine ait hissetmesi üzerine bir çalışma yapmıştır. Bu çalışmaya göre her öğrenciye özel öğrenme modeli seçmiş ve öğrencilerin kullanmış olduğu araç-gereci kullanım kolaylığına göre kişiye özel ayarladığında öğrencilerin derste daha etkin olduğunu gözlemlemiştir (Magdin, 2015).

1.2.1. Eğitsel Veri Madenciliği Çalışmaları

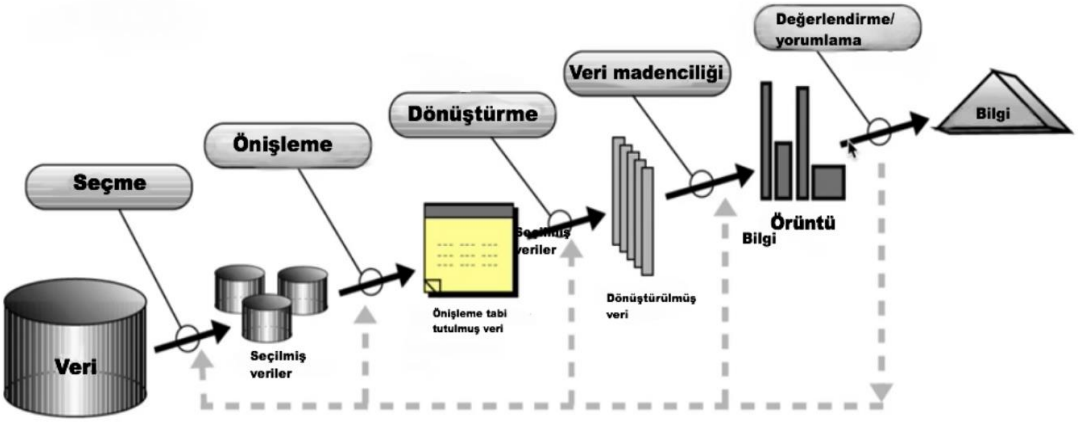
Eğitsel veri madenciliği (EVM) literatürde iki yönden ele alınır: Eğitim içeriği kısmı ve veri madenciliği kısmı. Eğitim kısmında eğitimde hangi konu ele alınacaksa o konu alınır, araştırılır ve veri madenciliği kısmına adım atılmış olur. Veri madenciliği kısmında uygun metotlar irdelenir ve eğitime katkı sağlayacak analizler yapılır.

EVM yeni yeni yaygınlaşmaya başlasa da popüler hale gelmiştir. Son sistematik inceleme çalışmaları, EVM ve bunun çeşitli eğitim alanlarındaki uygulamalarına ilişkin geniş ve büyüyen bir araştırma grubunu vurgulamıştır (Aldowah ve diğerleri,

2019; Baker ve diğerleri, 2017; Dutt ve diğerleri, 2016; Peña-Ayala, 2014). Eğitimde, araştırmacılar ve uygulayıcılar genellikle birincil amacın bir dizi yordayıcıdan bir sonuç (bağımlı) değişkeni çıkarmak olduğunu düşünürler. Yani problemi tahmin üzerinden çözmeye çalışırlar. (Berland ve diğerleri, 2014; Sinharay, 2016). EVM sürecinin birincil amacı, belirli bir priori hipotezi (yapılan bir araştırma çalışmasından önce üretilmiş olandır) olmaksızın verilerden yeni bilgiler bulmak ve çıkarmak ve keşfedilen bilgileri şu amaç için kullanmaktır: mümkünse, yeni bir teori inşa etmektir (Bulut, Yavuz 2019). EVM bağlamında, eğitim araştırmacılarının ilgi alanları temel olarak öğrenme, tahmine dayalı, davranışsal ve görsel analitik gibi çeşitli boyutlara odaklanmaktadır (Aldowah ve diğerleri, 2019). Gürdal ve Çakıcı "*Eğitsel veri madenciliği*" adlı makalesinde eğitsel veri madenciliği ile ilgili araştırmasında öğrencilerin yeteneklerini, ilgi alanlarını ve daha başarılı bir eğitimin nasıl olabileceği gibi problemleri analiz edip çözüm sunmada farklı bir perspektif getirdiğini söylemiştir (Gürdal ve ark., 2017). Can, yaptığı çalışmada temel eğitimden ortaöğretime geçiş sınavı kazanımlarını veri madenciliği yöntemleri ile değerlendirmiştir. Çalışma sonucunda Türkçe dersine hâkim olan öğrencilerin soruları anlamada daha başarılı olduğu çıkarımı yapılmıştır (Can, 2017). Bravo-Agapito ve diğerleri, yabancı dil öğrenimini iyileştirmek için eğitimsel veri madenciliği yöntemlerini kullanan araştırmaları analiz etti. Çalışmaların genel olarak öğrenci başarısını tahminlemek, öğrencilerin motivasyonunu dikkate almak ve öğretmenlere geri dönüş sağlamak için yapıldığı sonucuna vardılar (Bravo-Agapito ve diğerleri, 2019).

1.3. Veri Madenciliği Süreci

Veri madenciliği ile amaç tahminleme ya da tanımlama yapmaktır. Büyük verilerden oluşan veri setleri ile modelleme yapmadan önce verilerin ön işlemden geçmesi gerekir. Bilgiye ulaşma sürecinde KDD adımlarını izlemek, araştırmalarda tercih edilen bir yöntemdir. Aslında veriyi değerli ve bir amaca hizmet edecek hale getirme sürecidir. Genellikle ilk kez yapılan, araştırma amaçlı yapılan çalışmalar için kullanılmaktadır(Şeker,2018).



Şekil 1. Bilgi Keşfi-KDD süreçleri

Kaynak: Şeker, Şadi Evren, 2018. "Knime ile Uçtan Uca Veri Bilimi", https://sadievrenseker.com/wp-content/uploads/veribilimi_knime.pdf , 8-64, E.T: 16.11.2020

Veri işleme aşamaları Şekil 1.'de görüldüğü gibi veri temizleme, birleştirme, seçme, dönüştürme, veri madenciliği, desenler, bilgi sunumu aşamaları ile gerçekleşir.

- **Veri Temizleme:** Veriler içindeki eksik olan, uygun olmayan, hatalı girilmiş verileri temizleme adımıdır.
- **Veri Birleştirme:** Birkaç veri kaynağından alınan verilerin birleşme adımıdır.
- **Veri Seçme:** Veri kümelerinden alınan verilerden, modele kullanılması uygun verileri seçme adımıdır.
- **Veri Dönüştürme:** Veri uygun formlara dönüştürülür bu adımda
- **Veri Madenciliği Uygulaması:** Veri madenciliği algoritmaları uygulandığı adımdır.
- **Desenler:** Örüntü tanımlama adımıdır.
- **Bilgi Sunumu:** Elde edilmiş bilginin kullanıcıya sunulma adımıdır.

2. MAKİNE ÖĞRENMESİ

Teknolojinin ilerlemesiyle birlikte makine öğrenmesi hayatımıza giriş yaptı. Anlayabilen, öğrenebilen sistemler büyük bir talep oluşturmaktadır. Makine öğrenmesi verilerden öğrenilen bilen bir sistem, oluşturan tekniktir. Geçmişteki deneyimlerden öğrenme sağlanır ve karşılaşılabilecek durumlar için tahminde bulunur. Makine öğrenmesi; denetimli, denetimsiz ve pekiştirmeli öğrenme olmak üzere üç başlıkta incelenmektedir.

- **Denetimli Öğrenme:** Problemi sınıflandırma problemi içerisinde değerlendirir ve eğitim yapılan sistemde oluşturulan modelde test verileri ile tahmin ve tanıma yapar (Chao, 2011). Denetimli öğrenmenin problemlere yaklaşımlarını sınıflandırma ve regresyon olarak ikiye ayırabiliriz. SVM, Lojistik Regresyon, Gaussian NB, Yapay Sinir Ağları, Rastgele Orman, Karar Ağaçları, Karar Ormanı algoritmaları örnek verilebilir. Örneğin, regresyon algoritmaları ile bir kişinin resmine bakılarak cinsiyeti tahmin edilebilir; ya da sınıflandırma algoritmalarıyla da tümörlü bir hastanın tümörünün kötü huylu olup olmadığı tahmin edilebilir.
- **Denetimsiz Öğrenme:** Problemi kümeleme problemi içerisinde değerlendirir. Girilen örnekler arasındaki ilişkiyi bulmaya çalışır. Örnek algoritma olarak K-Ortalama verilebilir. Örneğin, kümeleme algoritmalarıyla müşterileri ilgi alanlarına göre gruplandırabiliriz.
- **Takviyeli Öğrenme:** Sisteme girilen bilgiler ve çıktıların doğru olup olmadığını kontrol eder. Kontrol sonucunda çıktılar doğruysa doğruluk oranını günceller. İnsanlar gibi öğrenme şekline yatkındır çünkü neden sonuç ilişkisi kurar. Takviyeli öğrenme yöntemleri ile örneğin bir robotun kolunu nasıl hareket ettirebileceği öğretilebilir.

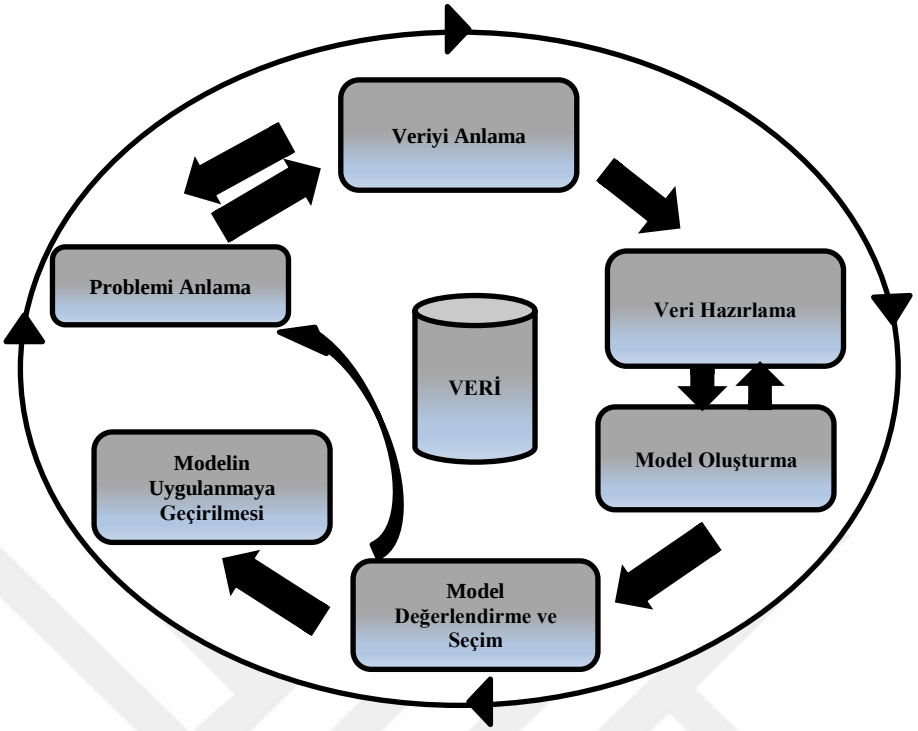
2.1. Makine Öğrenmesi Çalışmaları

Yıldız 2014'te makine öğrenmesi yöntemlerini kullanarak uzaktan eğitim alan öğrencilerin performanslarını doktora tezinde değerlendirmiştir. Çalışmada 218 öğrencinin 6 haftalık uzaktan öğretim sistemi verileri kullanılarak model oluşturulmuş ve 95 öğrenciye ait veriler test verisi olarak ayrılmıştır. Kullanılan algoritmalar

karşılaştırılıp performans ölçümleri yapılmıştır. Makine öğrenmesi yöntemleri eğitimin birçok alanında kullanılmıştır. Başka bir örnek vermek gerekirse 2017 yılında Gök tarafından yapılan tez çalışmasında makine öğrenmesi yöntemleri kullanılarak akademik başarı tahminlemesi yapılmıştır. 6,7 ve 8. Sınıfta okuyan öğrencilere 24 soruluk bir anket çalışması yapılmıştır. Elde edilen verilere regresyon ve sınıflandırma yöntemleri uygulanmıştır. Uygulanan modeller sonucunda Rastgele orman algoritmasının Türkçe ve genel başarı ortalamasını en iyi tahmin eden algoritma olduğu çıkarımı yapılmıştır. Portekiz de yapılan başka bir çalışmada (Cortez ve diğerleri, 2008) iki farklı ortaokulda okuyan öğrencilere anket çalışması yapılmıştır. Bu çalışmada öğrencilerin ders notları tahmini yapılmıştır. Tahminleme de Naive Bayes algoritması en yüksek başarı oranını vermiştir. 2010 yılında Türkiye'de yapılan başka bir çalışmada SVM kullanılarak 434 üniversite öğrencisinin üniversite sınavından aldığı puana göre matematik başarı oranı %86 doğru tahminlenmiştir. (Güner ve Çomak, 2011). Botelho ve diğerleri, makine öğrenmesi yöntemleri ile öğrencilerin davranış ve duygularını algılayan bir model oluşturmayı amaçladı. Özellik seçme yönteminin kullanımının yüksek başarı oranına sahip olduğu modeller geliştirmek için daha disiplinli bir seçenek olduğu çıkarımını yapmışlardır(Botelho ve diğerleri 2019).

2.2. CRISP-DM

CRISP-DM The Cross-Industry Standard Process for Data Mining (Sektörler Arası Standart Süreç Modeli-Veri Madenciliği) metodolojisi süreci güvenilir ve standart hale getirebilmek açısından basamak basamak prosedürler önerir. Bu basamaklar 6 adımdan oluşmaktadır. Problemi anlama/tanımlama, veriyi anlama, veriyi hazırlama, model oluşturma, modeli değerlendirme ve seçme, modelin uygulamaya geçmesini içeren döngüsel bir süreçtir. Yıldız, Börekçi, 2020'de yaptıkları çalışmada dokuzuncu sınıf öğrencilerinden eğitimsel veriler ile bir görü geliştirmeye çalışmışlardır. Sınav sonucuna göre başarılı olma, olmama durumu araştırmışlardır ve araştırma boyunca CRISP-DM süreçlerini takip etmişlerdir. Günümüzde pazarlama, dolandırıcılık, eğitim uygulamaları olmak üzere çeşitli çalışmalarda kullanılmıştır. Araştırma boyunca izlenen yol Şekil 2'de gösterilmiştir.



Şekil 2-CRISP-DM Döngü Şeması

Kaynak: Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R., 2000. CRISP-DM 1.0 Step-By-Step Data Mining Guide. SPSS. https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf E.T: 20.12.2020

Problemi Anlama: İlk adımda problemin çıktılarının ve gereksinimlerin ifade edilmesi durumudur. Bu durum net bir şekilde ortaya konmalıdır. Bu adımın doğru şekilde yapılması veri madenciliği kısmında bir yol haritası olmasını sağlayacaktır (Balaban, Kartal, 2015).

Veriyi Anlama: Problemi tanıdıktan sonraki önemli bir aşamadır. Probleme uygun verinin seçilmesi gerekir. Veri doğru şekilde tanınmazsa yapılan işlemler anlamsız hale dönüşebilir (Şeker, 2018).

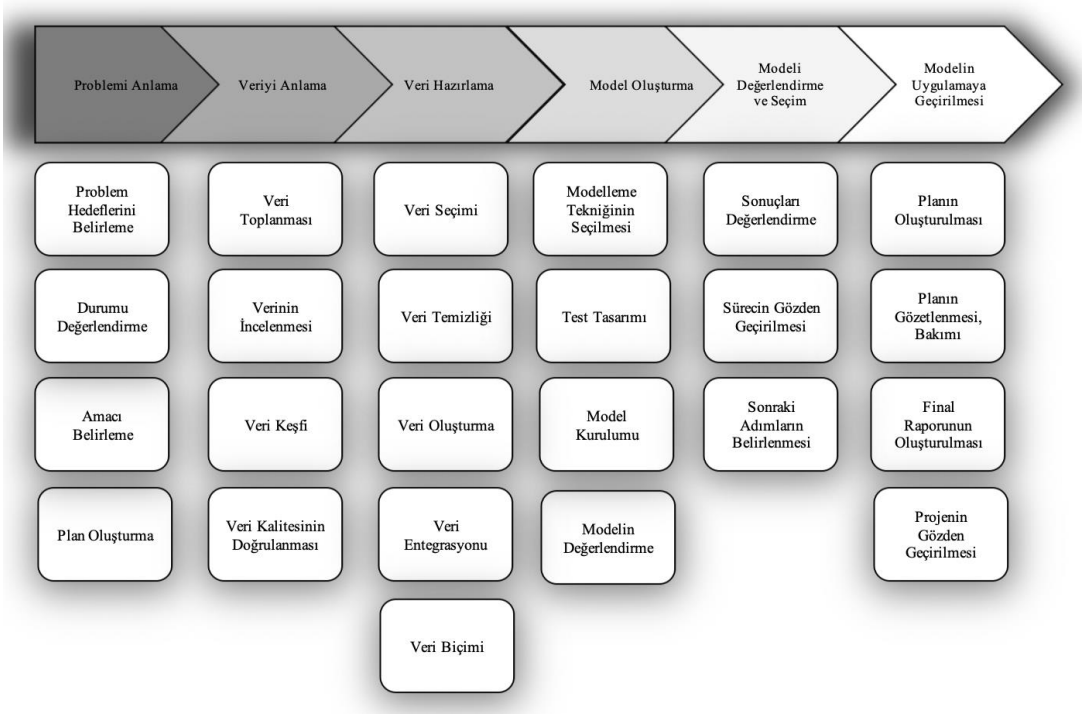
Veri Hazırlama: Verinin işleme aşamasıdır. Eksik veriler, gürültülü veriler, gereksiz veri çıkarımı, veri dönüşümleri gibi veriyle ilgili hangi işlemlerin yapılacağının kararlarının alındığı bir aşamadır (Şeker, 2017).

Model Oluşturma: Problemde kullanacağımız modeli eğitme aşamasıdır. Başarı oranı en yüksek modeli inşa etme aşamasıdır. En iyi modele ulaşmak için çoklu modellemeler yapılır (Amanet, 2020).

Modeli Değerlendirme ve Seçim: Modelin hedeflenen başarıya ulaşıp ulaşmadığı kontrol edilir ve teknik değerlendirmeler yapılır. Bu aşamada dışarıda tutma (holdout), k-katlı çaprazlama, doğruluk, duyarlılık gibi değerlendirme yöntemleri kullanılır. (Özdemir, 2016)

Modelin Uygulamaya Geçirilmesi: Bu aşamada artık model hedef sonuçlara yönelik karar üretebilir ya da değerler yeterli bulunmadığı takdirde baştaki adıma geri dönebilir. Problemi yeniden tanımlayabilir.

Şekil 3'te CRISP-DM aşamalarının adımları özet şeklinde verilmiştir.



Şekil 3. CRISP-DM Aşamaları

Kaynak: Özdemir, Şebnem. 2016. "Eğitimde Veri Madenciliği ve Öğrenci Akademik Başarı Öngörüsüne İlişkin Bir Uygulama" Doktora Tezi, İstanbul Üniversitesi Enformatik Ana Bilim Dalı.

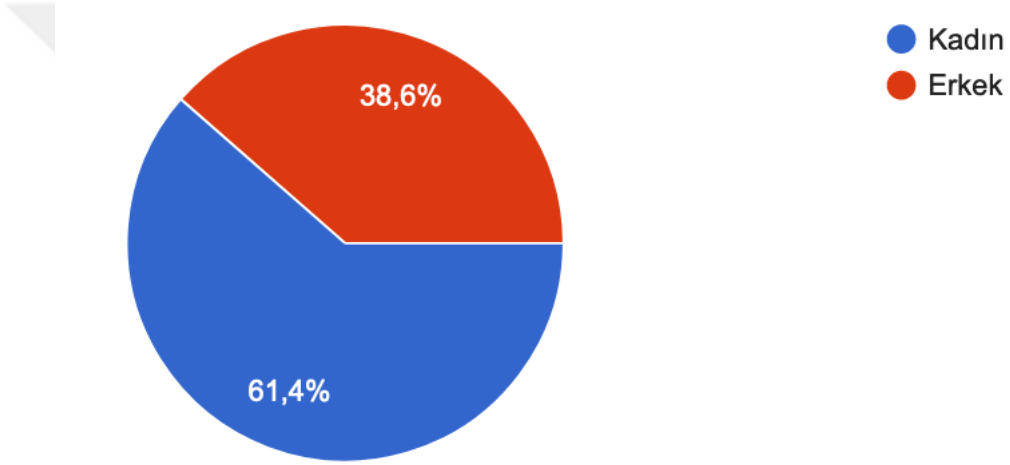
İKİNCİ BÖLÜM

MALZEME VE YÖNTEM

1. VERİNİN TOPLANMASI VE İŞLEME HAZIRLANMASI

1.1. Verinin İçeriği

Bu çalışmanın örneklem grubunu Türkiye'de üniversite sınavına girmiş olan kişiler oluşturmaktadır. Genel eğitim düzeyini temsil eden toplam 676 kişi yer almaktadır. Bu kişilerin %61,4 ü kadın, %38,6'ı erkektir.



Şekil 4. Örneklem Grubu Cinsiyet Dağılımı

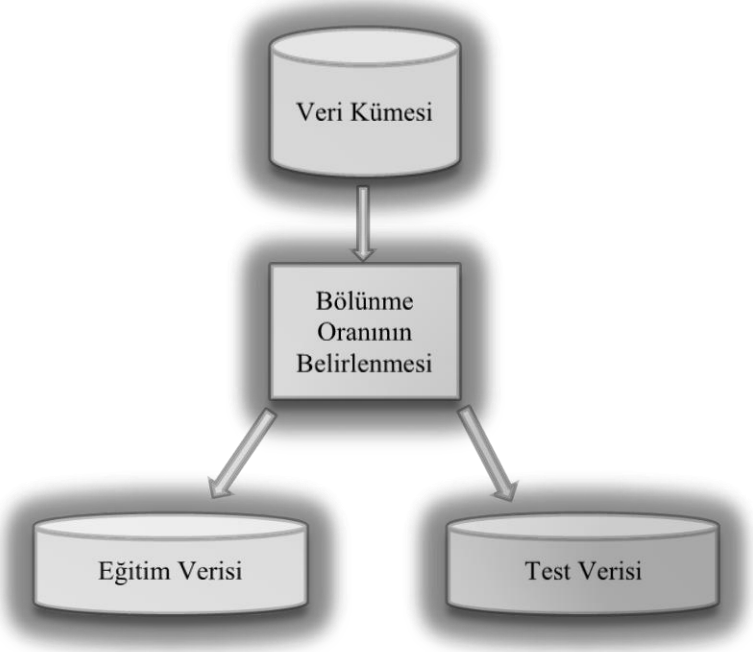
1.2. Veri Toplama Aracı

Bu çalışmada, araştırmacı tarafından oluşturulan "anket formu" akademik başarıyı etkileyen faktörlerin literatürdeki çalışmaları araştırılarak, sektördeki eğitimcilere danışılarak oluşturulmuştur. Anket için Google form kullanılmış ve anket online olarak yapılmıştır. Toplanan cevaplar csv dosya uzantısı ile veri analizi aracında incelenmiştir.

1.3. Veri Setleri ve Parametreler

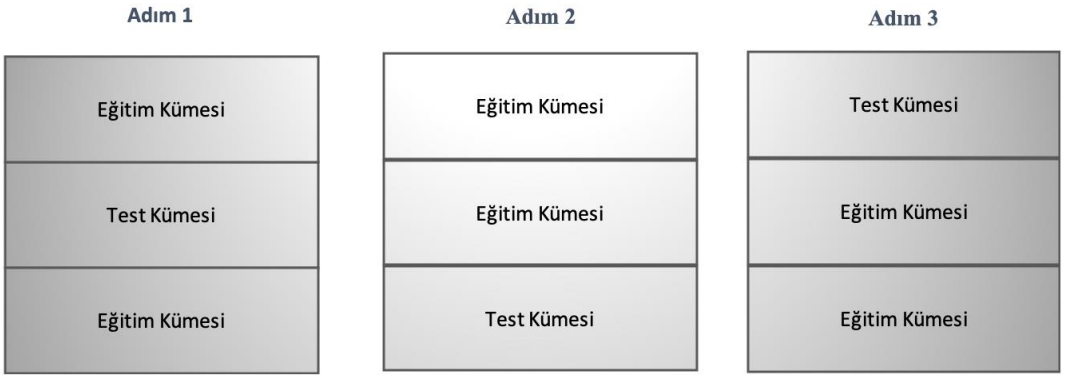
Veri madenciliği problemlerinde kullanılan veri setleri ve parametreler ile sonuca yönelik seçimler yapılmalıdır. Model uygulanmadan önce kayıp veriler yapılandırılmalı, verilerdeki gürültü ortadan kaldırılmalı, veriler az ise farklı veri setleriyle birlikte birleşme yapılmalı, özniteliklerin varyansları birbirine göre önemli ölçüde farklı ise dönüştürme yapılmalıdır. Parametreler seçilirken modeller göz önünde bulundurulmalıdır; rastgele orman algoritması kullanılırken ağaç sayısına, karar ağaçlarında düğüm, yaprak ve dal sayısına, lojistik regresyon kullanırken öğrenme oranına, SVM kullanırken k sayısına dikkat edilmelidir. Veri kümesinin eğitim ve test için ayırırken bu çalışmada kullanılan yöntemler: Dışarıda tutma (Holdout) ve k-katlı çapraz doğrulamadır. Öznitelik seçiminde bu çalışmada, değişken seçimi (feature selection) filtreleme seçeneklerinden korelasyon, anova ve ki-kare yöntemi kullanılmıştır

Dışarıda tutma (Holdout): Büyük boyutlu veri setlerinde tercih edilir. Veri kümesinin belli bir orandaki kısmını test etmek ve eğitmek için kullanır. Genel olarak rastgele bir seçim yapılır. Örneğin %70 eğitim %30 test olarak ayrılır (Özkan, 2016). Şekil 5'te dışarıda tutma (holdout) yapısını görebilirsiniz.



Şekil 5. Holdout Yapısı

K-Katlı Çapraz Doğrulama (K-folds validation): Veri seti k sayısında eşit alt kümelere bölünür sonrasında k-1 tanesi eğitim için kullanılır. Geriye kalan kısım da test için kullanılır (Özkan, 2016). Şekil 6'da k=3 için örnek katlı çapraz doğrulama yapısı verilmiştir.



Şekil 6. K Katlı Çapraz Doğrulama Yapısı

Kaynak: Özkan, Y. 2016, Veri Madenciliği Yöntemleri (3.Basım). İstanbul; Papatya Bilim, 217

Pearson Korelasyon (Correlation): Özellik seçim yöntemlerinden filtreleme seçeneklerinden biridir. Özniteliklerin doğrusal olarak ilişkili olduğu veri setlerinde ilişkinin değişimini ölçmeye yarar. Aldığı değerler -1 ile 1 arasındadır. İlişki 1 değerine yaklaştıkça mükemmelleşir. Veri kümesindeki öznitelikler arasındaki korelasyon modelin başarını etkilemektedir (Akben ve Alkan 2015).

Ki-Kare Testi (Chi2): Kategorik değişkenler ve hedef değişken arasındaki ilişki derecesini ölçmeye yarar.

Anova Testi: Bağımlı değişkenin, bağımsız değişkenler üzerine etkisini araştıran varyans analizidir.

2. VERİ İŞLEME ARAÇLARI

Bu çalışmada veri analizi ve modelleme için dil olarak Python arayüz olarak Anaconda kullanılmıştır. Detaylı bilgiler bir sonraki bölümde anlatılmıştır.

2.1. Python

Python; programlama dilleri arasında veri madenciliği ve derin öğrenme gibi geniş kütüphaneler sunan açık kaynak kodlu ücretsiz, yazılımı ve anlaşılması kolay, okunabilirliği yüksek bir dildir. Platform bağımsızdır, interaktif modda kullanıma izin verir. Google, Wikipedia, Yahoo, NASA gibi dünya devleri pythonu tercih ediyor. Kolaylıkla verileri analiz edebilmesi ve basit ara yüzü sayesinde kullanırlığı artmıştır.

Veri madenciliğinde en çok kullanılan 2.dildir. Uzun satırlar yazmamıza gerek kalmadan birkaç satırda halletmemizi sağlar. İstatistiksel analizleri çözmede daha kabiliyetli olması en çok tercih edilen dillerden biri olmasını sağlamıştır. Bunun yanı sıra zengin bir kullanıcı topluluğuna sahip olduğundan yeni başlayanla uzmanına kadar birçok çözüm için kaynaklara ulaşım rahatlığı sunar. Orta ölçekli işler için en uygun dildir (Cerebro. 2018. "Python Neden Bu Kadar Popüler"https://medium.com/kodcular/python-neden-bu-kadar-popüler-d7f0f6819de5 Erişim Tarihi :30.11.2020).

2.2.Anaconda

Python ve R dilini kullanabileceğimiz açık kaynak kodlu tümleşik dağıtımıdır. Veri bilimi ve benzeri bilimsel çalışmalar için kolaylık sunan bir dağıtımdır. Anaconda Navigatör ile gelen IDE'ler veri madenciliği, derin öğrenme gibi çalışmalarda kullanabileceğimiz çok çeşitli kütüphaneleri paket halinde, tekrardan indirmeye gerek kalmadan sunar. Anaconda.org adresinden işletim sistemine göre kolayca kurulum yapılabilir.

2.3.Pandas Kütüphanesi

Veri yapılarına sütun ekleme silme, veri kümelerini birleştirme, şekillendirme, verileri düzenleme, eksik verilerle işlem yapma, verileri sistematik olarak artan azalan

sırada düzenleme, verilerde yenileme ve filtre uygulama gibi özelliklere sahiptir. Kurulum yaptığımız Pandas Kütüphanesi;

import pandas as pd yazarak kullanılmaya başlanabilir (Murat Gülcan. 2018. "Python Pandas Kütüphanesi"https://medium.com/@wmuratgulcan/python-pandas-kütüphanesi-597209068238 Erişim Tarihi: 30.11.2020).

2.4.Numpy Kütüphanesi

Matematik kütüphanesidir, bilimsel hesaplamaların hızlı yapılmasını sağlar. Numpy dizelerini kullanır. Bu daha hızlı olmasını sağlar. Kurulum yaptığımız Numpy Kütüphanesi;

import numpy as nd yazarak kullanılmaya başlanabilir (Merve Durna. 2019. "Veri Bilimi için Temel Python Kütüphaneleri-1: Numpy"https://medium.com/bilişim-hareketi/veri-bilimi-için-temel-python-kütüphaneleri-1-numpy-750429a0d8e5 Erişim Tarihi:30.11.2020).

2.5. SeaBorn Kütüphanesi

Verileri görselleştirmeyi sağlar, matplotlip kütüphanesine arayüz sağlar. İstatiksel grafikleri renk seçenekleri sunarak ilgi çekici hale getirir. Kurulum yaptığımız Numpy Kütüphanesi;

import seaborn as sns yazarak kullanılmaya başlanabilir (Alperen Balık. 2018. "Seaborn ile Veri Görselleştirilmesi"https://www.veribilimiokulu.com/blog/seaborn-ile-veri-gorsellestirmesi/ Erişim Tarihi:30.11.2020).

2.6. Scikit Learn Kütüphanesi

Veri madenciliği, yapay zekâ gibi konularda en çok kullanılan kütüphanelerdendir. Veri analiziyle ilgili uygulamaları başından sonuna kadar kullanılmasını sağlar. Öğrenme modelleri için temel ihtiyaçların çoğunu sunar (Eksik verileri doldurma, çapraz doğrulama, sonuç değerlendirme vb.). Kurulum yaptığımız ScikitLearn Kütüphanesi;

from sklearn import datasets yazarak kullanılmaya başlanabilir (Birol Yüceoğlu. 2017. "Scikit-Learn ile Veri Analitiğine Giriş"

http://www.veridefteri.com/2017/11/23/scikit-learn-ile-veri-analitigine-giris/ Erişim Tarihi:30.11.2020).

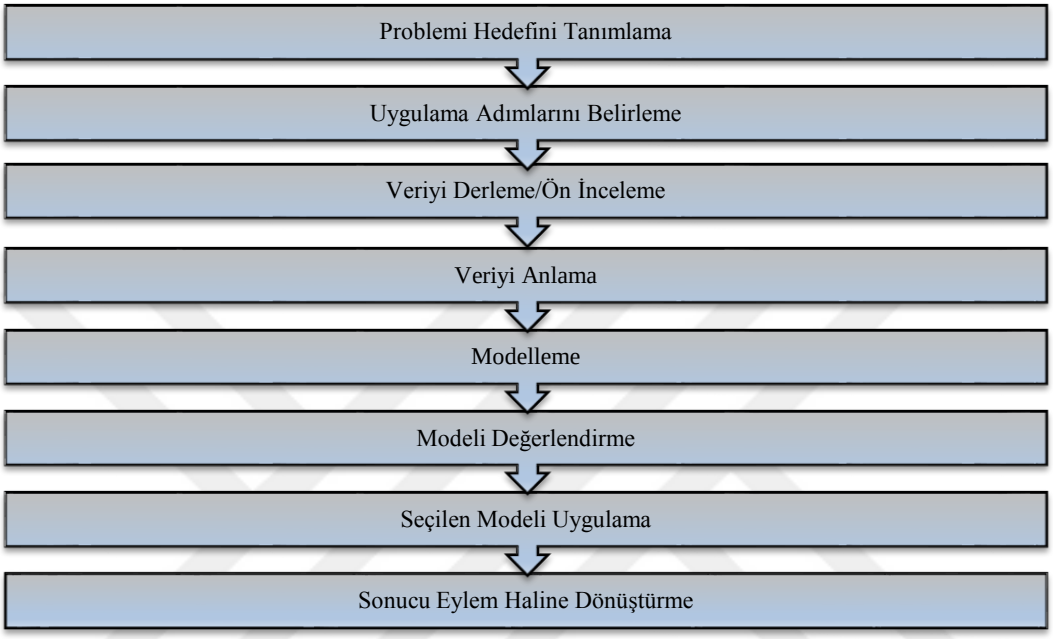
2.7. Matplotlip Kütüphanesi

Veri görselleştirme için kullanılan temel seviye bir kütüphanedir. İki ve üç boyutlu grafikler çizilmesini sağlar. Kurulum yaptığımız Matplotlip Kütüphanesi;

import matplotlip.pyplot as plt yazarak kullanılmaya başlanabilir (Mert Alabaş. 2019. "Python ile Veri Görselleştirme: Matplotlib Kütüphanesi" https://medium.com/datarunner/matplotlibkutuphanesi-1-99087692102b Erişim Tarihi: 30.11.2020).

3. CRISP-DM YÖNTEMİ SÜREÇLERİ

Bu bölümde çalışmanın planlanmasından verilerin derlenmesine ve sonuçların üretilmesine kadar CRISP-DM adımları sunulmaktadır.



Şekil 7. Çalışma Modeli

3.1. Problemin Tanımlanması ya da Hedeflerin Belirlenmesi

Günümüzün eğitim sistemi 4+4+4 olmak üzere toplamda 12 yıl zorunlu üç kademeye ayrılmıştır. Birinci kademe ilkokul, ikinci kademe ortaokul, üçüncü kademe ortaöğretim olarak düzenlenmiştir. İlk 8 yıldan sonra her eğitim kademesine geçişte okullaşma sınav ile olmaktadır. Her üst kademeye geçişte okullaşma oranı azalmaktadır. Tablo 2 Millî Eğitim Bakanlığının ve Yüksek Öğretim Kurumunun 2020'de yayınladığı raporlardan alınan sonuçlarını görebilirsiniz.

15 yaş grubu öğrencilerden elde edilen raporları sunan diğer ülkelerle kıyaslayan üçer yıllık aralarla yapılan 79 ülkenin katıldığı PISA 2018 sonuçlarına göre Türkiye, okuma becerilerinde 40. sırada, matematik okur yazarlığında 42. sırada fen okur yazarlığında 39. sırada yer almaktadır. 2015 PISA sonuçlarına göre sıralamalarda artış görünmektedir. Okul türlerine göre yapılan analizde fen lisesi öğrencilerinin en başarılı grupta olduğu görülmektedir. Okuma becerileri ve fen alanında kız öğrencilerin başarı oranı daha yüksek açıklanmıştır. Bölgelere göre değerlendirilme

yapıldığında Batı Marmara, Doğu Marmara ve Batı Anadolu'dan katılan öğrencilerin daha başarılı olduğu belirtilmiştir (Milli Eğitim Bakanlığı. 2018. "PISA 2018 Türkiye Ön Raporu" http://pisa.meb.gov.tr/wp-content/uploads/2020/01/PISA_2018_Turkiye_On_Raporu.pdf Erişim Tarihi 09.12.2020).

Bu çalışmada eğitimsel veri madenciliği tekniklerinden faydalanarak "lise öğrencilerinin üniversite giriş başarılarının eğitsel veri madenciliği ile tahmin edilmesi" problemi ele alınmıştır. Problem ele alınırken Bölüm 5.1.1. Akademik Başarıyı Etkileyen Faktörler incelenmiştir.

3.1.1. Akademik Başarıyı Etkileyen Faktörler

Gelişmiş ülkelerdeki eğitim politikalarının sisteme olumlu sonuçlar getirdiğini gören ülkemiz, eğitimle ilgili birçok önlem almaya çalışıp kalkınma programları oluşturmuştur. On birinci kalkınma planında (T.C Kalkınma Bakanlığı. 2013. "Onuncu Kalkınma Planı (2019-2023)." Ankara. https://www.sbb.gov.tr/wp-content/uploads/2019/07/OnbirinciKalkinmaPlani.pdf Erişim Tarihi: 07.12.2020) amaç, hayat boyu öğrenme imkanları için ulaşım sağlayarak girişimci, etiğe önem veren, yenilikçi, teknolojiye ayak uyduran bireyler yetiştirmektir. Bu hedefler doğrultusunda okullaşma oranları arttırılacaktır. Öğrencilerin ruhsal ve bedensel gelişimlerine göre eğitim ortamları hazırlanacaktır. Eğitim kurumlarının veri analiz imkanları yükseltilecek, gelen veriler bazında planlama eğitim sistemi geliştirilecektir (T.C Kalkınma Bakanlığı, 2019:554.2). Eğitim politikaları veriye dayalı belirlenecek; uygulamalar veri üzerinden analiz edilecektir (T.C Kalkınma Bakanlığı, 2019: 554). Veri tabanları birleştirilip eğitsel veri ambarı oluşumu sağlanacak ve bu verilerden yapay zekâ teknolojileriyle faydalanılacaktır (T.C Kalkınma Bakanlığı, 2019: 554.1). Bu hedefler doğrultusunda veri madenciliği akademik başarıyı etkileyen faktörlerin çıkarımı için büyük önem arz etmektedir. Türkiye'de eğitim başarı odaklı olup en önemli hedef üniversiteye girebilmek olmuştur (Yıldırım,2006).

Tablo 2'de ülkemizdeki okullaşma sayıları belirtilmiştir. Bu tablo incelendiğinde eğitim kademesi yükseldikçe ortaokuldan sonra okullaşma sayılarında azalma görülmeye başlanmıştır. Kadın okullaşma sayısı erkek okullaşma sayısından düşüktür (Milli Eğitim Bakanlığı. 2020." Örgün Eğitim İstatistikleri" https://istatistik.yok.gov.tr/ , https://sgb.meb.gov.tr/www/icerik_goruntule.php?KNO=396 Erişim Tarihi: 12.11.2020).

Tablo 2. Türkiye Geneli 2020 Okullaşma Bilgileri

GRUP	TOPLAM ÖĞRENCİ SAYISI	KADIN ÖĞRENCİ SAYISI	ERKEK ÖĞRENCİ SAYISI
Doktora	101 242	46 943	54 299
Yüksek Lisans	297 001	138 888	158 113
Lisans	4 538 926	2 119 610	2 419 316
Ön Lisans	3 002 964	1 526 121	1 476 843
Ortaöğretim	5 630 652	2 645 534	2 985 118
Ortaokul	5 701 564	2 816 120	2 885 444
İlkokul	5 279 945	2 561 756	2 718 189
Anaokulu	1 629 720	783 471	846 249

Akademik başarıyı etkileyen faktörler üzerine Yıldız ve Börekçi'nin 2020'de yaptığı çalışmaya bakıldığında öğrencilerin ve ailelelerin demografik bilgileri, çalışma şekilleri, öğrenme etkinliklerine katılımları, ders tutumları ve bilimsel-epistemolojik inançlarının akademik başarıya olumlu yönde bir etkinin gözlemlenmiştir. Üniversite sınavına hazırlanıp başarı odaklı yerleşen öğrencilerin yıl sonunda ders notlarında başarısızlık söz konusu olabiliyor. Kurt ve Erdem bu konuda bir çalışma yapar ve çalışmada başarılı olan, olmayan öğrencilerin bilgilerinden yola çıkarak veri madenciliği yöntemleri ile analiz eder. Araştırma sonucunda öğrencilerin çalışma durumu, araştırmacı nitelikleri, doğru bölüm seçimi, rehberlik faaliyetleri, diploma notu ve cinsiyetin başarı üzerindeki olumlu etkisini saptamışlardır. Demir (2009), gelişmemiş bölgelerde yaptığı araştırmasında akademik başarıyı etkileyen faktörleri okul kaynaklı ve aile kaynaklı olarak bir arada ele almıştır. Başarıyı etkileyen en önemli faktörleri ise öğrencinin not durumu, çalışıp çalışmadığı, cinsiyeti, algısı, arkadaş sayısı, ailenin öğrenciye desteği, okul etkinlikleri ve öğretmen tutumu şeklinde sıralanmıştır. Güneş'in 2012 de yayınlanan "ÖSYS başarısını etkileyen faktörler analizi" adlı makalesinde sosyo-demografik (ailenin yapısı büyüklüğü, eğitimi vb.) etkenler ve okul etkenleri (öğretmen-öğrenci mevcudu, sınıf hacmi vb.) dışında çevresi, gelenekleri, dini, okulun fiziki yapısı başarıya etki eden faktörler olarak değerlendirmektedir(Güneş, 2012).

3.2. Uygulama Adımlarını Belirleme

Bu çalışmanın ilk aşamasında veri toplama aracı etik kurula sunulmuş ve etik kurul raporu alınmıştır. Tek seferde yanıtlanacak şekilde duygu içermeyen genel sorular sorulmuştur. Anket sorularına Ek 1'de erişilebilir.

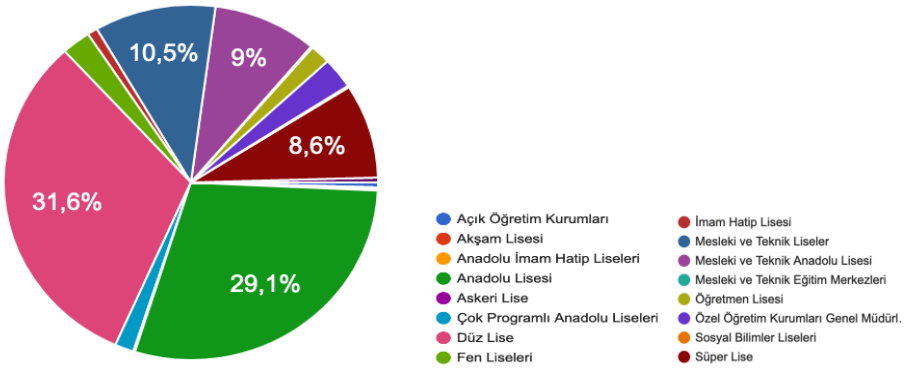
3.3. Veriyi Derleme ve Ön İnceleme

Bu araştırma için oluşturulan veri seti, veri madenciliği yöntemleri uygulanmadan önce ön işlemden geçirilmiştir.

- Eksik veriler belirlenmiş ve elenmiştir.
- Gürültüler azaltılmıştır.
- Csv formatına dönüşüm yapılmıştır

3.4. Veriyi Anlama

Toplanan veri setinde 50 tane öz nitelik ve 676 tane kişi verisi bulunmaktadır. Tablo 3'te veri setine ait tüm öz nitelikler ve türleri bulunmaktadır. Verilerin ön işlem sonrası anketi yarım bırakan 4 kişi verisi silindikten sonra 672 veri ile analizler yapılacaktır. Veri setinde kişilerin girdiği ortak olmayan bilgiler düzeltilmiştir. Örneğin diploma notu alanını beşlik, onluk ve yüzlük istemde cevap verenlerin diploma notu yüzlük sisteme dönüştürülmüştür. Şekil 8'de veri setindeki lise türlerine göre dağılım yer almaktadır. En çok katılım %31,6 ile düz lise mezunu tarafından olmuştur.



Şekil 8. Lise Türlerine Göre Veri Dağılımı

Tablo 3. Veri Seti Değişken İçerikleri ve Türleri

Açıklama	Değişken Gösterimi	Tür
Cinsiyeti	CINSIYET	object
Okuduğu Alan	ALAN	object
Eşit Ağırlık Matematik Notu	MATNOT	int32
Eşit Ağırlık Türk Dili ve Edebiyatı Notu	TDNOT	int32
Eşit Ağırlık Tarih Notu	TARNOT	int32
Eşit Ağırlık Coğrafya Notu	COGNOT	int32
Yabancı Dil İngilizce Notu	INGNOT	int32
Yabancı Dil Matematik Notu	MAT2NOT	int32
Yabancı Dil Türk Dili ve Edebiyatı Notu	TD2NOT	int32
Yabancı Dil Tarih Notu	TARNOT1	int32
Yabancı Dil Coğrafya Notu	COGNOT1	int32
Sözel Türk Dili ve Edebiyatı Notu	TDNOT1	int32
Sözel Tarih Notu	TARNOT2	int32
Sözel Coğrafya Notu	COGNOT2	int32
Sözel Felsefe Notu	FELSEFENOT	int32
Sayısal Matematik Notu	MATNOT2	int32
Sayısal Fizik Notu	FIZKNOT	int32
Sayısal Kimya Notu	KIMYANOT	int32
Sayısal Biyoloji Notu	BIONOT	int32
Okul Türü	LISETUR	object
Özel Ders Alma Durumu	DERSHANEFLAG	object
Ders İçeriklerine Ulaşım Kolaylığı	DERSICERIK	object
Çalışma Saati	CALSAAT	float64
Dijital Öğrenim Platformlarından Yararlanma	DOP	object
Okunan Lise Yeterliliği	LISEIYETER	object
Okurken Çalışma Durumu	CALISMA	object
Yaşanan Akran Zorbalığı	AKRANZORBA	object
Rehberlik Servisi	DANISMANLIK	object
Alan Seçiminin Doğruluğu	DOGRUYER	object
Sınıf Mevcudu	MEVCUT	int64
Diploma Notu	DNOT	float64
Sınav Giriş Tarihi	TARIH	float64
Öğrenim Görülen İl	IL	object
Sosyal Medya Hesabı Varlığı	SOSYALMEDYA	object
Sağlık Sorunu	SAGLIK	object
Annenin Yaşama Durumu	ANNEDURUM	object
Babanın Yaşama Durumu	BABADURUM	object
Ailenin Gelir Durumu	GELIR	object
Annenin Eğitim Durumu	ANNEEGITIM	object
Babanın Eğitim Durumu	BABAEGITIM	object

Tablo 3. Devamı

Ailenin Medeni Durumu	AILEMEDENI	object
Kardeş Sayısı	KARDES	int64
Üniversitede Okuyan Kardeş	UNIKARDES	object
Üniversiteye Ziyaret	UNIZIYARET	object
Ön lisans Durumu	ONLISANS	object
Barajı Geçme Durumu	BARAJ	object
Sınavı Kazanma Durumu	UNIHAK	object
Üniversiteye Yerleşme Durumu	UNI	object
Annenin Çalışma Durumu	ANNECDURUM	object
Babanın Çalışma Durumu	BABACDURUM	object

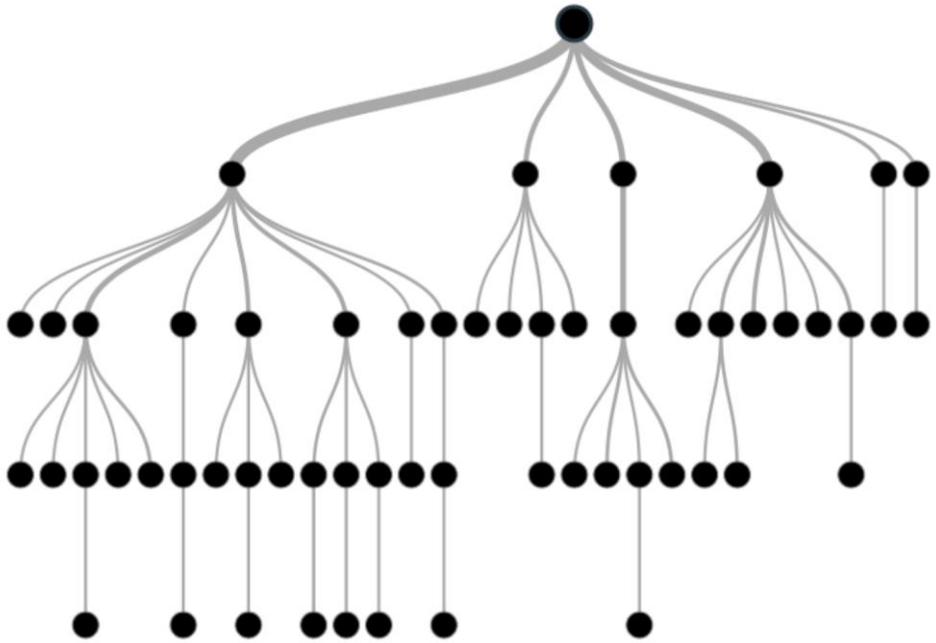
4. VERİ MADENCİLİĞİ TEKNİKLERİ ALGORİTMALARI

Bu tez çalışmasında modelleme yapılırken kullanılan algoritmaların detayları bu bölümde açıklanacaktır. Veri madenciliği teknikleri kullandığı algoritmalarla özelliklerine göre dörde ayrılır ve bunlar: karar ağaçları sınıflandırması (ID3, C4.5 ve C5, CART, SLIQ, SPRINT, Değişken Merkez, Karar Ağacı,), istatistiğe dayalı algoritmalar (NB, Regresyon, CHAID), mesafeye dayalı sınıflandırma (K-En Yakın Komşu Algoritması, En Küçük Mesafe Sınıflandırıcısı) ve yapay sinir ağlarıdır (İleri Sürümlü Yapay Sinir Ağları, Hata Geriye Yayma Yöntemi) (Silahtaroğlu, 2013). Algoritma geliştirme alanında birçok yeni algoritmalar üzerinde çalışılmaktadır. Kullanılan algoritmaların performanslarını ölçmek için dışarıda tutma (holdout) yöntemi ve k-katlı çapraz doğrulama yöntemleri seçilmiştir.

K-kat ile k=10'a kadar çapraz doğrulama yapılmıştır. Dışarıda tutma (holdout) ile Eğitim seti %60, %70, %75,%80 oranlarında olacak şekilde ayrılmıştır. Veri setleri her modele girişinde rastgele belirli oranlarla test edilmiştir.

4.1. Karar Ağaçları Sınıflandırması

Karar ağaçları, sınıflandırma problemlerinde en çok tercih edilen algoritmadır. Denetimli bir öğrenme yöntemidir. Amacı girilen verilerden özellik çıkarımı yapıp karar kurallarını öğrenip tahminleme yapmaktır. Ağacın derinliği arttıkça karmaşık hale gelir ve model için uygun bir hal alır (Scikit-Learn Developers "Decision Trees"https://scikit-learn.org/stable/modules/tree.html?highlight=id3 , Erişim Tarihi: 26.11.20). Geçmiş verilerden tanımlanmış bir hedef değişkene sahiptir. Verileri kurallarla küçük gruplara bölerek, en tepeden en aşağı inen bir çözümü vardır (Kantardzic, 2011). Şekil 9'da ağaç yapısının şekli görünmektedir.



Şekil 9. Karar Ağacı Yapısı

Karar ağaçlarını anlamak ve yorumlamak kolaydır, karar ağaçlarıyla görselleştirme yapılabilir. Karar ağaçları, büyük veri hazırlıklarına ihtiyaç duymaz. Modelin doğrulanması mümkündür. Maliyeti azdır. Nominal ve nominal olmayan değerleri işleyebilir. Doğrusal olmayan değerleri işleyebilir. Yalnız şu açıdan dikkat etmek gerekir ki karar ağaçları dengeli olmayan veri kümesiyle problemlidir, bu yüzden karar ağacını oluşturmadan önce veri kümesini dengelemek iyi olacaktır.

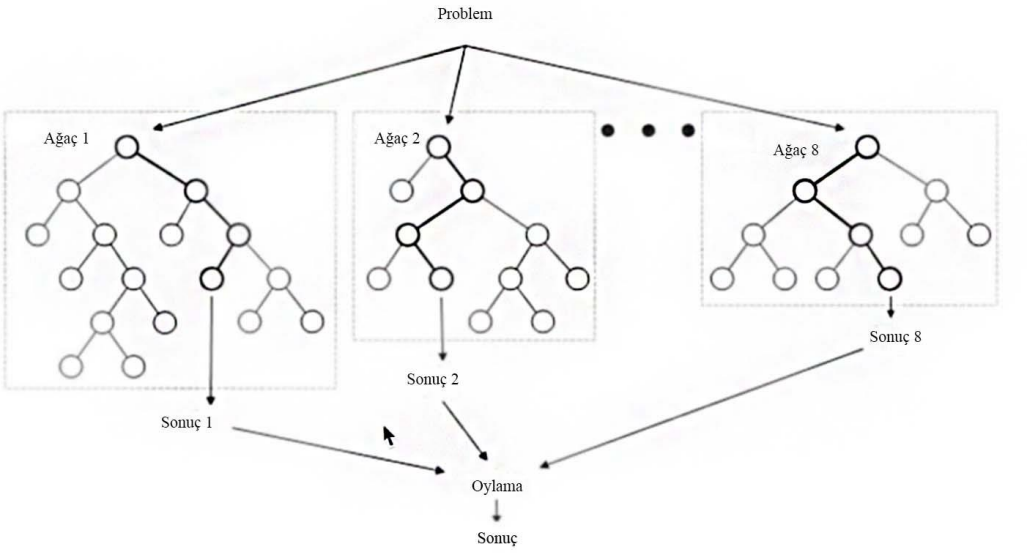
Karar ağaçları pazarlama piyasası, müşteri kazanımı, riskli kredi uygulamaları, potansiyel müşteri olup olmama gibi uygulamalarda tahminleme için kullanılabilir (Avinash Navlani. 2018. "Understanding Random Forest Classifiers in Python" https://www.datacamp.com/community/tutorials/random-forests-classifier-python Erişim Tarihi: 28.11.20) .

4.1.1. CART

Diğer sınıflandırma algoritmalarına çok benzerdi ancak sayısal hedef özellikleri desteklemesi ve kural setlerini hesaplaması açısından diğerlerinden farklı yönleri de vardır. Her düğümde, bilgi kazancı sağlayan özelliği ve eşini kullanarak onları ikili alt gruplara ayırır ve homojen yapı oluşturur.. Sınıflandırma ve regresyon amacıyla kullanılabilir. Dallanmayı Twoing ve Gini algoritmalarıyla yapar (Gedleç, Ş. , Yılmaz, H. B. 2020 "Karar Ağaçlarında Algoritma Seçimi"https://www.datasciencearth.com/karar-agaclarinda-algoritma-secimi/ Erişim Tarihi 28.11.20).

4.2. Rastgele Orman Algoritması

Rastgele orman algoritması karar ağaçlarını temel alır hem sınıflandır hem de regresyon yöntemidir. Problemdeki her ağaç için sınıflandırma yapar ve oylama ile sınıfını belirler. Karar ağaçlarındaki aşırı öğrenme problemi rastgele orman algoritmasında yoktur (Breiman 2001). Bunun nedeni rastgele ormanda problemi çözmek için rastgele seçilen veri setleri ve öznitelik seti üzerinde eğitim gerçekleştirilir. Bu şekilde birçok karar ağacı oluşmuş olur ve her biri tahmin için kullanılır. Algoritma hem regresyon hem de sınıflandırma problemleri için kullanılabilir. Karar ağaçlarına göre yorumlaması zordur. Şekil 10'da Rastgele orman yapısı görünmektedir.

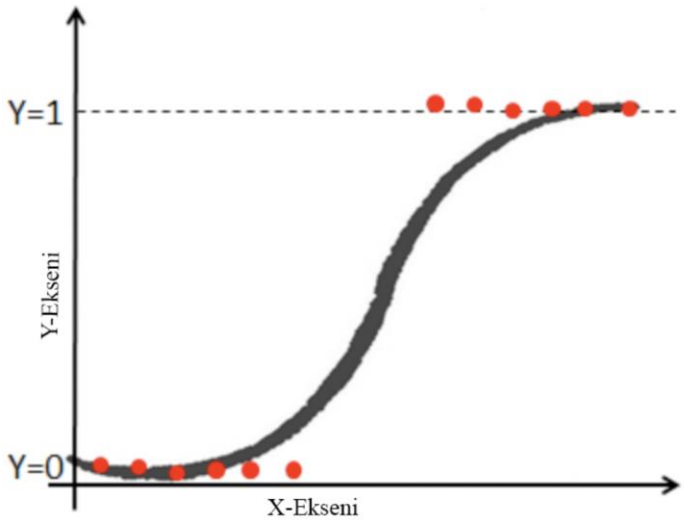


Şekil 10. Rastgele Orman Yapısı

Rastgele ormanlar, özellik seçimi öneri motorları görüntü sınıflandırması gibi çeşitli uygulamalara sahiptir. Sadık kredi başvuru sahiplerini sınıflandırmak, dolandırıcılık faaliyetlerini belirlemek ve hastalıkları tahminleme için kullanılabilir.

4.3. Lojistik Regresyon

Lojistik regresyon, spam mail hastada diyabet tahmini belirli ürün alımları gibi çeşitli problem tahminlemesi için kullanılabilir. İstatistik bir yöntemdir. Bağımlı değişken Bernoulli (ikili) dağılımını takip eder. Sonuç hedef veya hedef değişken ikiye ayrılmıştır. Tahminleme en yüksek olasılığa göre yapılır. Yapısı nedeniyle uygulaması kolaydır. Yüksek hesaplama gücü gerektirmez. Özellikleri ölçümlemeye gerek duymaz. Aşırı uyuma karşı dirençsizdir. Çok fazla kategorik özelliği işlemede sıkıntı yaşayabilir. Benzer değişken özelliklerinde sıkıntı yaşayabilir. Şekil 11'de Lojistik Regresyon yapısı görünmektedir.

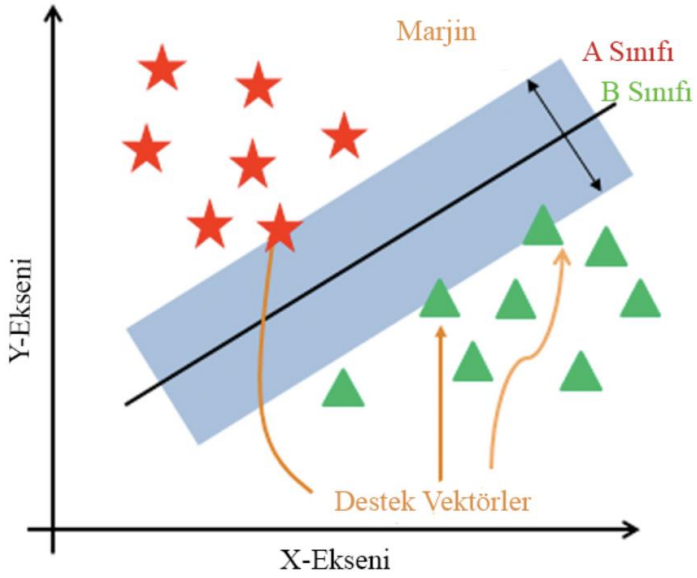


Şekil 11. Lojistik Regresyon Yapısı

Lojistik regresyon spam maillerde, hastanın diyabet olma tahmininde, belirli ürün alımlarında gibi çeşitli problemlerde tahminleme için kullanılabilir (Avinash Navlani. 2019. "Understanding Logistic Regression in Python" https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python Erişim Tarihi: 28.11.20).

4.4. Destek Vektör Makineleri (SVM)

Sınıflandırma grubunda yer alır ama hem sınıflandırma hem de regresyon problemlerinde kullanılır. SVM öğrenme yöntemlerinin kesiştiği bir yapıdadır. İstatistikteki yapısal risk minimasyonu gibi çalışmaktadır (Abdalla ve Erdoğmuş 2014). Destek vektörleri, hiper düzleme en yakın olan veri noktalarıdır. SVM'nin amacı, veri kümesini sınıflara, en iyi şekilde bölen bir maksimum marjinal hiper düzlem bulmaktır. Genel anlamda hızlı bir tahmin sunar. Net bir ayırma marjı ve yüksek boyutsal boşlukla iyi çalışır. Daha az bellek kullanır. Yüksek eğitim süresinden dolayı büyük veriler için çok kullanışlı değildir. Şekil 12' de Destek Vektör Makineleri yapısı görünmektedir.

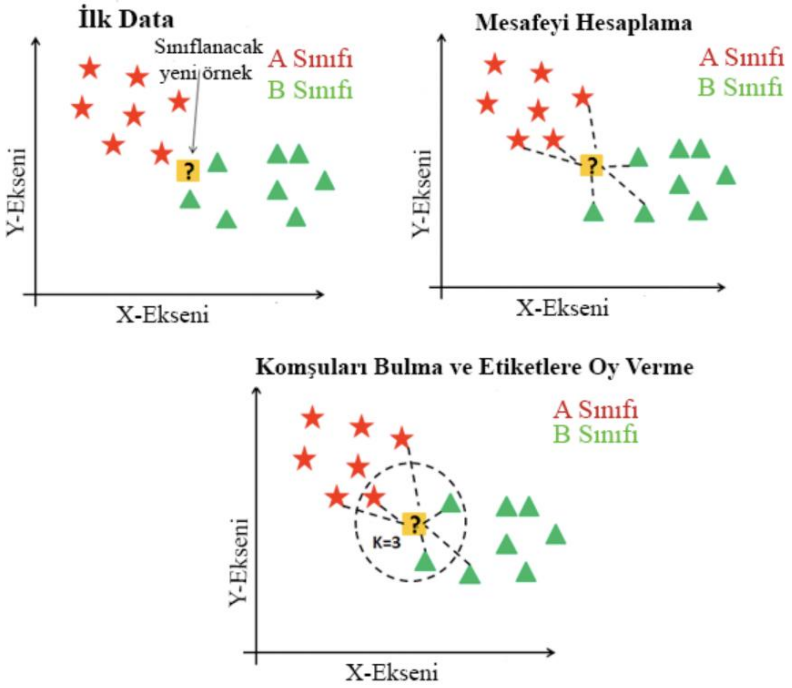


Şekil 12. Destek Vektör Makineleri Yapısı

Yüz tanıma, parmak okuma, maille gelen haberin sınıflandırılması, el yazısı tanıma gibi çeşitli problemlerde ayırt edici sınıflandırıcı olarak kullanılabilir (Avinash Navlani. 2019."Support Vector Machines with Scikit-Learn"https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python Erişim Tarihi: 28.11.2020).

4.5. K-En Yakın Komşu Algoritması (KNN)

K en yakın komşuların sayısıdır ve k sayısı için belirli bir sayı seçme mecburiyeti yoktur. Sezgilerinize güvenerek seçebilirsiniz. K sayısının komşularına uzaklığına göre belirlediği bir karar aşamasıdır. K sayısının dışında uzaklık miktarı ve öznitelik sayısına bağlıdır. KNN'de eğitim kısmı yoktur (Bhatia ve Vandana, 2010). Çalışma mantığı önce mesafeyi hesaplar, sonra en yakın komşuları bulur ve etiketlere oy verir. Az miktarda komşu düşük önyargılı olur ve esnek uyumlu yüksek varyansa sahip olur. Çok sayıda komşu ise karar sınırı daha yumuşak, ön yargısı yüksek, varyansı düşük olur. Eğitime gerek olmadığı için eğitim aşaması hızlıdır, regresyon problemlerinde kullanılabilir, doğrusal verileri iyi analiz edebilir. Test aşaması yavaş ve pahalıdır. Büyük boyutlu veriler için uygun değildir. Şekil 13' te K-En Yakın Komşu yapısı görünmektedir.

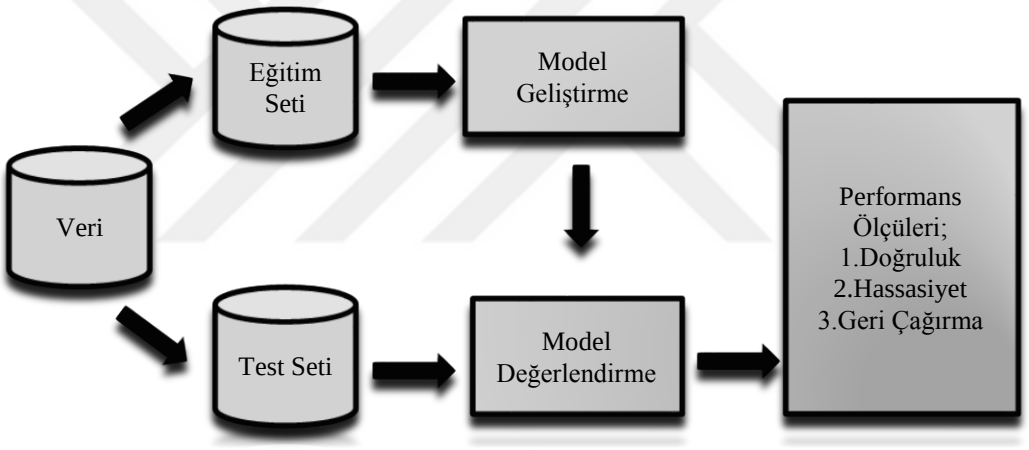


Şekil 13. K-En Yakın Komşu Yapısı

Potansiyel seçmeni bulma, kredi riski hesaplama, video tanıma, el yazısı tanıma gibi çeşitli problemlerde tahminleme için kullanılabilir (Avinash Navlani. 2019." KNN Classification using Scikit- Learn" https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn Erişim Tarihi: 28.11.2020).

4.6. Gaussion NB Algoritması

NB sınıflandırması istatistiğe dayalı bir algoritmadır, Bayes kuralına dayalıdır. NB, elde olan ve sınıflanmış verileri kullanır ve yeni gelen verinin hangi sınıfa ait olduğunun olasılığını hesaplar (Silahtaroğlu, 2013). Aynı zamanda denetimli öğrenme algoritmalarındandır. Tek bir özellik için olasılığı hesaplama şekli şu şekildedir: verilmiş olan sınıf etiketleri için önceki olasılığı hesaplar, sonrasında her bir etiketin her sınıfı için olma olasılığını hesaplar, daha sonrasında Bayes kuralına göre bu olasılıkları değerlendirir, son olarak hangi sınıfın daha yüksek olasılıkta olduğunu gösterir. Hesaplama maliyeti düşüktür. Çoklu sınıf tahminlerinde iyidir. Ancak belirli bir sınıfın eğitimi yoksa tahmin yapamayacaktır. Şekil 14'te Gaussian NB Yapısı görünmektedir.



Şekil 14. Gaussian NB Yapısı

Tavsiye sistemleri, müşteri yorumlarının iyi olup olmadığını belirleme, yazım denetlemesi, spam denetimi gibi çeşitli problemlerde tahminleme için kullanılabilir (Avinash Navlani. 2019."Naive Bayes Classification Using Scikit-Learn"https://www.datacamp.com/community/tutorials/naive-bayes-scikit-learn Erişim Tarihi:28.11.2020).

5. MODELİ DEĞERLENDİRME VE SEÇME

Kullanılan modeller bu aşamada performans değerlendirme metrikleri ile değerlendirilmiştir. Bölüm 5.1.'da kullanılan metriklerin detayları yer almaktadır.

5.1. Performans Değerlendirme Metrikleri

Veri madenciliğinde problemin çözümü için kurulan modellerin çalışma sonucunu değerlendirmek için çeşitli metrikler vardır. Bu çalışmada kullanılan metrikler aşağıda açıklanmıştır. Bölüm 3'te kullanılan performans kıyaslama ölçüleri ve elde edilen sonuçlar detaylı şekilde yer almaktadır

Karışıklık Matrisi (Confusion Matrix): Gerçek veri ile tahminlenen verinin karşılaştırılması yapılır. ROC ve AUC eğrisini, hassasiyeti, özgünlüğü, doğruluğu, geri çağırmayı ölçmek için kullanışlıdır.

Tablo 4. Karışıklık Matrisi

		Gerçek Değer	
		Pozitif (1)	Negatif (0)
Tahmini Değer	Pozitif (1)	TP	FP
	Negatif (0)	FN	TN

TP (True Positive): Doğru ve pozitif bir tahminleme yapıldığını gösterir

FP (False Positive): Yanlış ve pozitif bir tahminleme yapıldığını gösterir

FN (False Negative): Yanlış ve negatif bir tahminleme yapıldığını gösterir

TN (True Negative): Doğru ve negatif bir tahminleme yapıldığını gösterir

Doğruluk Oranı = (TP+TN) / (TP + TN + FP + FN)

Hassasiyet(Sensivity): Tahmin edilen pozitif örneklerden kaçının doğru olduğunu verir.

Hassasiyet=(TP)/(TP+FP)

Geri Çağırma(Recall): Tüm pozitif sınıfların dışındaki doğru tanımlanan örnek sayısıdır

Geri Çağırma = TP/(TP+FN)

Tahmin(Precision): Doğru var olarak tahminlerin , toplam var tahminlerine oranıdır.

Tahmin=TP/(FP+TP)

F Puanı(F Score): Sınıflandırma işleminin iyi olup olmadığını gösteren performans ölçüsüdür. Sıfır ile bir arasında değer alır. Bire yaklaştıkça doğruluk oranı artar (Şengül Karaderili. 2018. "Hata Matrisini Anlamak" https://medium.com/@sengul_krdrl/hata-matrisini-anlamak-7035b7921c0f Erişim Tarihi: 30.11.2020).

F Puanı = (2*Hassaiyet*Geri Çağırma) /Hassasiyet+Geri Çağırma

ROC (Receiver Operating Characteristic): Farklı sınıflar için olasılık eğrisi sunar. Modelde doğru ve olumlu tahminlemelenen (TP) verilerin ve yanlış olumlu olan (FP) verilerin grafik çizimi ile gösterimidir (Narkhede 2018).

AUC (Area Under the Curve):ROC eğrisi altında kalan alandır. X ekseninde FP, y ekseninde TP ve eğrinin altındaki alanın ölçümüdür. Alan ne kadar büyük olursa modelin başarısı o kadar yüksektir (Berna Taş.2019. "ROC Eğrisi ve Eğri Altında Kalan Alan(AUC)" https://medium.com/@bernatas/roc-e%C4%9Frisi-ve-e%C4%9Fri-alt%C4%B1nda-kalan-alan-auc-97b058e8e0cf Erişim Tarihi: 30.11.2020).

Ortalama Kare Hata: Regresyon eğsinin bir dizi noktaya yakınlık mesafesini açıklar. Sonuçlar pozitif değer alır. Sıfıra yaklaştıkça performans artar. Ortalama kare hata formülü aşağıdadır.

$$= \frac{1}{n}\sum_{j=1}^{n} e_j^2$$

Kök Ortalama Kare Hata: Modelin tahminlediği değerlerin gerçek olan değerlere olan uzaklığıdır. Negatif puanlar daha iyi performans sergiler. Kök ortalama kara hata formülü aşağıdadır.

$$= \sqrt{\frac{\frac{1}{n}\sum_{j=1}^{n} e_j^2}{n}}$$

Ortalama Mutlak Hata: İki sürekli değişken arasındaki farktır. Sıfıra yaklaşan sonuçlar daha iyi performans gösterir. Ortalama mutlak hata formülü aşağıdadır (Anonim. 2020. "MSE,RMSE, MAE, MAPE ve Diğer Metrikler" https://veribilimcisi.com/2017/07/14/mse-rmse-mae-mape-metrikleri-nedir/ Erişim Tarihi: 30.11.2010).

$$= \frac{1}{n}\sum_{j=1}^{n}|e_j|$$

5.2. Seçilen Modeli Uygulama

Modellerin kıyaslanması sonucu model seçilir. Seçilen performans, yüksek model ve karşılaştırmaları Değerlendirmeler Bölümünde detaylandırılmıştır.

5.3. Sonucu Eylem Haline Dönüştürme

CRISP-DM adımlarının sonuncu olan bu aşamada elde edilen sonuçların analizleri gerçekleştirilir. Analizler sonucunda yeni bir inceleme alanı açılıp açılmayacağına bu adımda karar verilir. Bu yapının nasıl kurgulandığı Bölüm 3'te detaylandırılmıştır. Yeni bir alanın incelenmesi yönünden görüşler, sonuç bölümünde değerlendirilecektir.

ÜÇÜNCÜ BÖLÜM
ÜNİVERSİTE ADAYLARINA YÖNELİK BULGULARIN DEĞERLENDİRİLMESİ

Çalışmanın bu bölümünde karar ağacı, lojistik regresyon, Gaussion NB, rastgele orman, K-en yakın komşu, destek vektör makineleri algoritmalarından ve öznitelik çıkarım yöntemlerinden elde edilen sonuçların değerlendirilmeleri yer almaktadır. Her analiz ve yöntemde kullanılan veri kümesi, kullanılan öznitelikleri, performans değerlendirme yöntemlerine ait temel bilgeler açıklanmıştır.

Veri kümesi üzerinde yapılan çalışmalarda, Bölüm 3.5' de açıklanmış olan algoritmalar kullanılmıştır. Kullanılan her algoritmanın başarısını etkileyen dışa ayırma (hold out), k-kat çapraz geçerleme yöntemleri kullanılmıştır. Modellerin performansını arttırmak için anova, korelasyon ve ki-kare öznitelik çıkarım yöntemleri uygulanmıştır ve performans sonuçları hesaplanmıştır. Bu sonuçlar Bölüm 3.6.1'de anlatılmış olan performans değerlendirme metrikleri ile değerlendirilmiştir.

1. VERİ SETİ İLE İLGİLİ DEĞERLENDİRMELER

Veri madenciliği teknikleri uygulanmadan önce veri setinin istatiksel değerlerine bakılmalıdır. Veri setine ait bu değerler Anaconda Navigatör 'de temel istatiksel hesaplamalarla ve Google Form grafiklerle elde edilmiştir. Elde edilen değerler ders notlarına uygulanan normalizasyon sonucudur. Ders notları üzerinde minimum ve maksimum değerleri ele alınarak normalleştirme yapılmıştır. Matematiksel ifadesi aşağıda yer almaktadır. Verinin değeri

$$\text{düşükDeğer} + \frac{(\text{yüksekDeğer} - \text{düşükDeğer}) * (\text{verininDeğeri} - \text{minDersNotu})}{\text{maxDersNotu} - \text{minDersNotu}}$$

denklemi ile hesaplanmıştır.

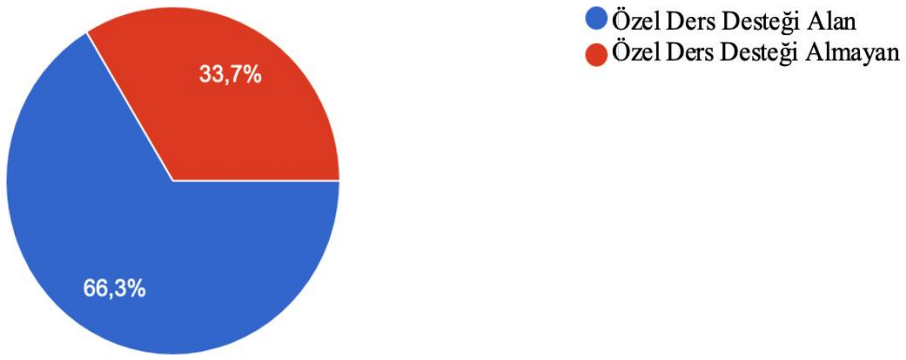
Tablo 5. Veri Setinin Sayısal Veri Değerleri

	Miktar	Ortalama	Standart Sapma	Min.	25%	50%	75%	Max.
MATNOT	615	2.965.608	1.245.206	1	2	3	4	5
TDNOT	615	2.767.196	1.654.119	0	1	3	4	5
TARNOT	615	3.005.291	1.292.695	1	2	3	4	5
COGNOT	615	2.603.175	1.706.965	0	1	3	4	5
INGNOT	615	3.571.429	0.551274	3	3	4	4	5
MAT2NOT	615	1.928.571	1.390.639	0	1	2	3	5
TD2NOT	615	2.994.709	0.861787	2	2	3	4	5
TARNOT1	615	3.047.619	0.872727	2	2	3	4	5
COGNOT1	615	3.082.011	0.877750	2	2	3	4	5
TDNOT1	615	2.288.360	1.542.990	0	1	2	4	5
TARNOT2	615	2.753.968	1.188.047	1	2	3	4	5
COGNOT2	615	2.412.698	1.525.746	0	1	3	4	5
FELSEFENOT	615	2.343.915	1.556.445	0	1	2	4	5
MATNOT2	615	2.902.116	1.645.973	0	2	3	4	5
FIZKNOT	615	2.727.513	1.609.763	0	1	3	4	5
KIMYANOT	615	2.791.005	1.638.795	0	1	3	4	5
BIONOT	615	2.796.296	1.566.658	0	2	3	4	5
CALSAAT	615	3.621.693	2.436.679	0	2	3	5	16
MEVCUT	615	31.658.730	8.774.659	10	25	30	38	50
DNOT	615	81.993.210	11.594.832	50	73.083	83.3	91.6	100
TARIH	615	2.001.859.788	6.930.417	1979	1999	2003	2006	2019
KARDES	615	2.944.444	1.635.022	0	2	2.5	3	13

Tablo 5 Veri setinin sayısal veri değerleri tablosunda en temel istatistik hesaplamalar yer almaktadır. Tablo 5 incelendiğinde kategorik veriler sayısal veri olarak algılanmamıştır. Eğer kategorik veriler yer almış olsaydı modelleme yapılmadan önce veriler sayısal değerlere dönüştürülürdü. Bu işlem CRISP-DM sürecinin Veri Derleme bölümünde yapılmaktadır.

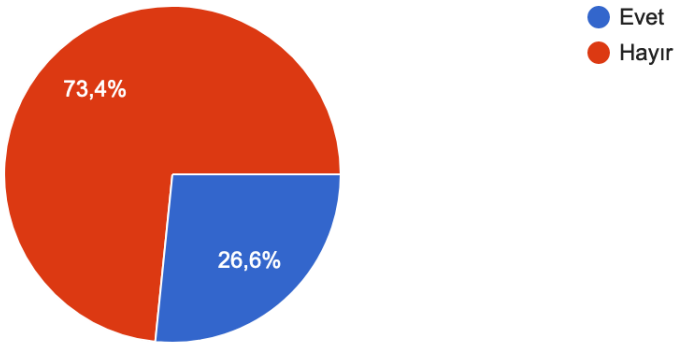
Verilerin grafik şeklinde incelenmesi, analiz edilme aşamasında gözden kaçırılmaması gereken noktaları ele alınmasını sağlayacaktır. Veri setine baktığımızda Matematik not ortalaması 2.965608, Türk Dili Edebiyatı not ortalaması 2.767196 görünmektedir.

Sınav için ayrılan çalışma süresini incelediğimizde günlük çalışma ortalaması 3.62 saat olarak yer almaktadır.



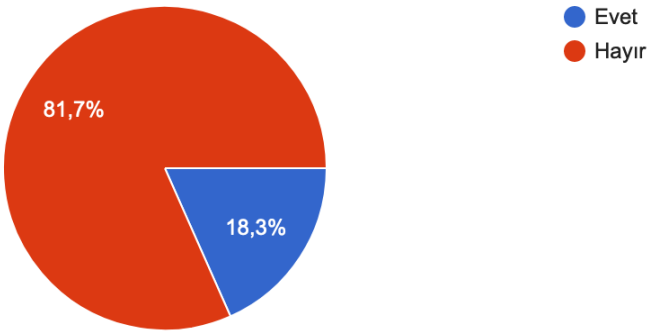
Şekil 15. Özel Ders Alma Durumu

Üniversite sınavına hazırlanırken özel ders alma, dershaneye gitme faktörleri incelediğinde katılımcıların %66,3'ü özel ders almış ya da dershaneye gitmiştir (Şekil 15).



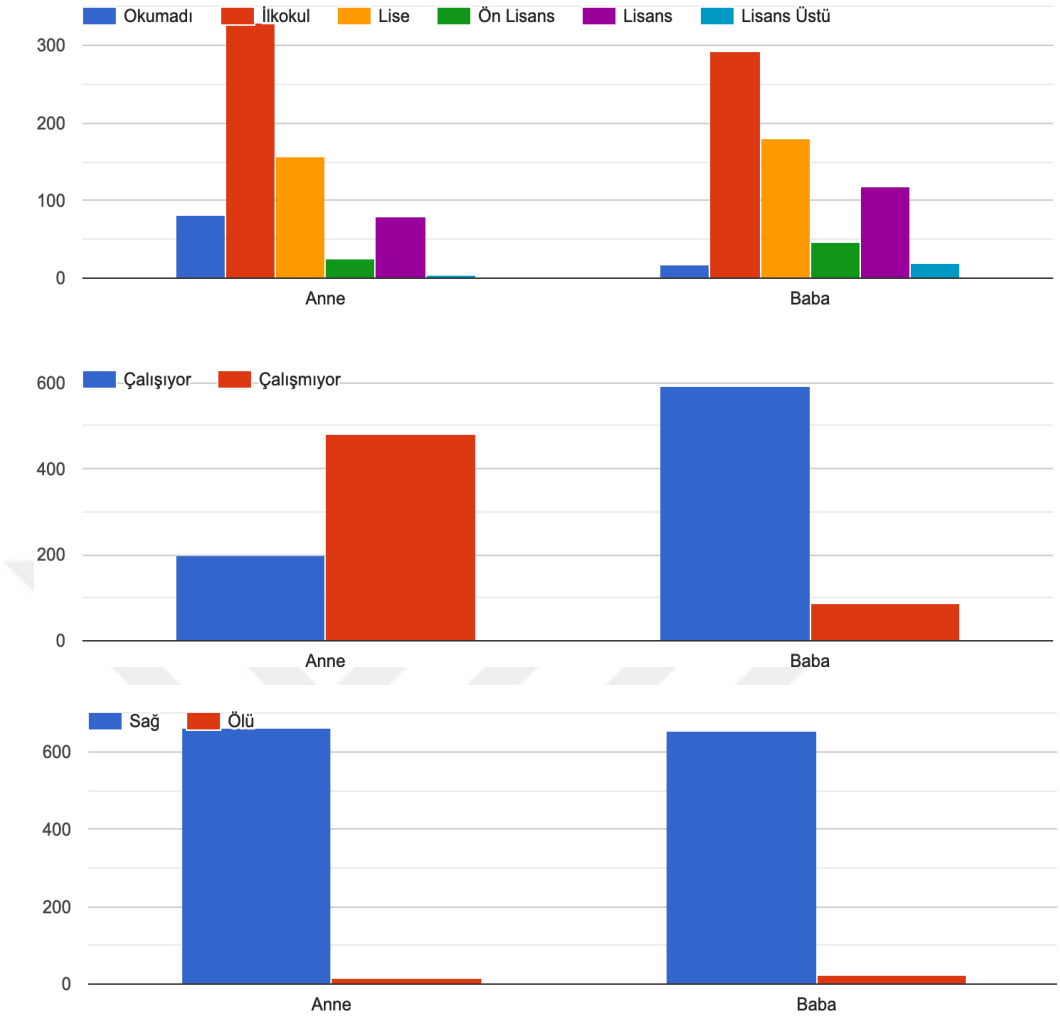
Şekil 16. Lisede Aldığı Eğitimi Sınav İçin Yeterli Bulma Durumu

Katılımcıların %26,6'sı lisede aldığı eğitimi yeterli bulurken %73,4'ü yetersiz bulmuştur (Şekil 16).



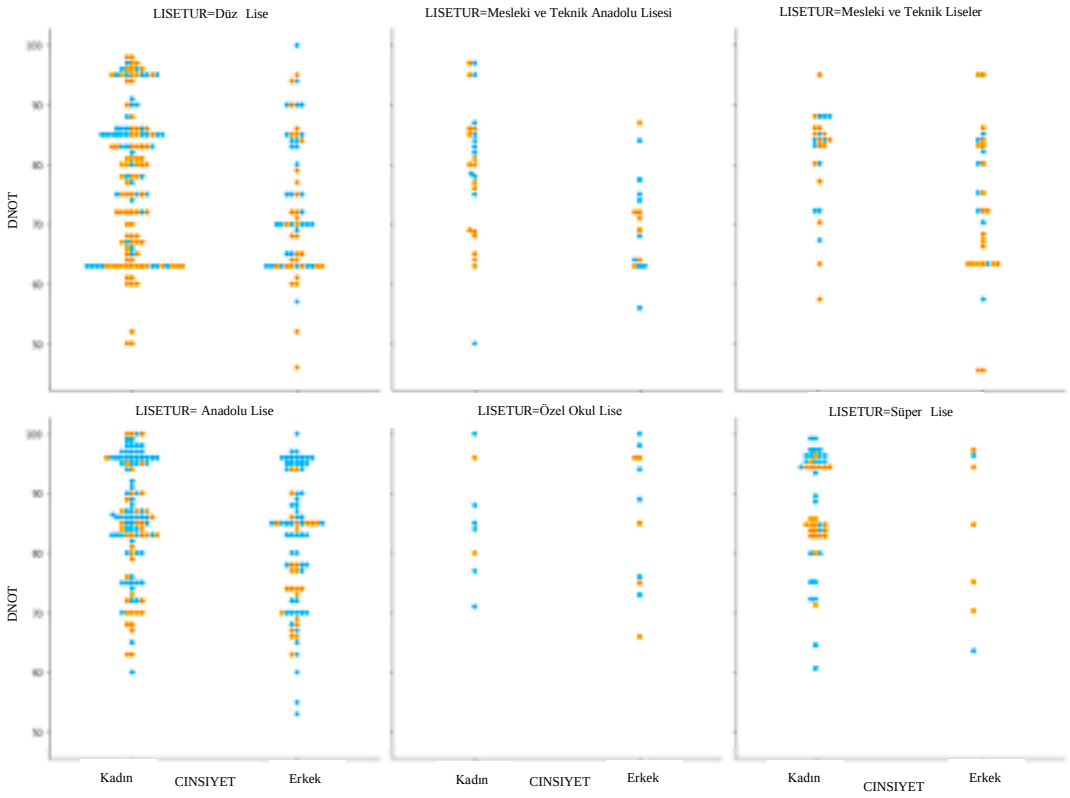
Şekil 17. İş yerinde çalışma durumu

Katılımcıların %18,3'ü üniversite sınavına hazırlanırken bir iş yerinde çalışmıştır (Şekil 17). Katılımcıların ailelerinin eğitim, çalışma ve yaşama durumları Şekil 18- Anne-Baba Eğitim, Çalışma, Yaşama Durumu 'da verilmiştir.



Şekil 18. Anne-Baba Eğitim, Çalışma, Yaşama Durumu

Şekil 18'de üniversiteye hak kazanma durumunu etkileyecek şekilde homojen bir dağılıma sahip olunmadığı için anlamlı bir fark, her durum için gerçekleşmemiştir. Ek 3'te ailenin eğitim, çalışma ve yaşama durumuna göre üniversiteyi hak kazanma durumu istatistiksel grafikleri verilmiştir. Bu grafikler incelendiğinde katılımcının annesinin eğitim seviyesi arttıkça üniversiteyi kazanma başarı oranı yükselmiştir. Annenin eğitim durumu anlamlı bir fark oluştururken babanın eğitim seviyesi anlamlı bir fark oluşturmamıştır. Anne ve babanın yaşam durumu veri dağılımının miktarı nedeniyle etkili bir fark oluşturmamıştır. Babanın çalışma durumu çok etkilemiyorken annenin çalışıyor olması başarı oranını pozitif yönde etkilemiştir.



Şekil 19. Lise Türü, Cinsiyeti ve Diploma Notuna Göre Üniversite Tercihi Yapmaya Hak Kazanma Durumu

Şekil 19. Lise Türü, Cinsiyeti ve Diploma Notuna Göre Üniversite Tercihi Yapmaya Hak Kazanma Durumu incelendiğinde, üniversiteyi kazanma oranının en yüksek olduğu görülen lise türü, Anadolu liseleridir. Şekil 19'da 15 lise türünden anlamlı sonuç ifade eden 6 lise türü yer almıştır. Kalan lise türlerinin sayılarındaki azlık nedeni ile anlamlı sonuç ifade etmemektedir. Diploma notu yükseldikçe kazanma olasılığının yükseldiği görülmektedir. Cinsiyetin Erkek olmasının pozitif bir katkısı olmuştur. Şekil 19**.** Lise Türüne Göre Tercih Edilen Üniversiteye Yerleşmeye Hak Kazanma Durumu tablosunda hangi lise türünden kaç tane tercih yapmaya hak kazanan olduğunun detayları yer almaktadır. Veri setindeki bilgilerin detaylı tabloları ve grafikleri Ek 2. Veri Setinden Elde Edilen Bilgilerin Dağılımı 'da yer almaktadır.

1.1. Öznitelik Analizi

Üniversite sınavında başarılı olmayı etkileyen bir çok faktör vardır. Bu faktörler Bölüm 2'de, Problemi tanımlama adımında Bölüm 2'de Akademik Başarıyı Etkileyen Faktörler başlığı altında incelenmiştir.

Bunların belli başlıcaları kişinin lisede okuduğu alan bilgisi, ders notları, okuduğu lise türü, günlük çalışma saati, rehberlik servisi , dershane , aile gelir seviyesi, ailesinin yaşama ve medeni durumu gibi demografik özellikler yer almaktadır. Veri setinde bulunan öznitelik bilgileri Tablo 6. Öznitelik Listesi'nde yer almaktadır.

Tablo 6. Öznitelik Listesi

Öznitelik Adı	Açıklama
CINSIYET	Veri kümesindeki adayların cinsiyetini ifade eder.
ALAN	Veri kümesindeki adayların liseden mezun olduğu alanı ifade eder. (sözel, eşit ağırlık, sözel, yabancı dil)
MATNOT	Veri kümesindeki adayların eşit ağırlık matematik notunu ifade eder.
TDNOT	Veri kümesindeki adayların eşit ağırlık Türk dili ve edebiyatı notunu ifade eder.
TARNOT	Veri kümesindeki adayların eşit ağırlık tarih notunu ifade eder.
COGNOT	Veri kümesindeki adayların eşit ağırlık coğrafya notunu ifade eder.
INGNOT	Veri kümesindeki adayların yabancı dil İngilizce notunu ifade eder.
MAT2NOT	Veri kümesindeki adayların yabancı dil matematik notunu ifade eder.
TD2NOT	Veri kümesindeki adayların yabancı dil Türk dili ve edebiyatı notunu ifade eder.
TARNOT1	Veri kümesindeki adayların yabancı dil tarih notunu ifade eder.
COGNOT1	Veri kümesindeki adayların yabancı dil coğrafya notunu ifade eder.
TDNOT1	Veri kümesindeki adayların sözel Türk dili ve edebiyatı notunu ifade eder.
TARNOT2	Veri kümesindeki adayların sözel tarih notunu ifade eder.
COGNOT2	Veri kümesindeki adayların sözel coğrafya notunu ifade eder.
FELSEFENOT	Veri kümesindeki adayların sözel felsefe notunu ifade eder.
MATNOT2	Veri kümesindeki adayların sayısal matematik notunu ifade eder.
FIZKNOT	Veri kümesindeki adayların sayısal fizik notunu ifade eder.

KIMYANOT	Veri kümesindeki adayların sayısal kimya notunu ifade eder.
BIONOT	Veri kümesindeki adayların sayısal biyoloji notunu ifade eder.
LISETUR	Veri kümesindeki adayların hangi tür liseden mezun olduğunu ifade eder.
DERSHANEFLAG	Veri kümesindeki adayların lisede özel ders alma durumunu ifade eder.
DERSICERIK	Veri kümesindeki adayların üniversite sınavına hazırlanırken ders içeriklerine ulaşım kolaylığını ifade eder.
CALSAAT	Veri kümesindeki adayların günlük ders çalışma saatini ifade eder.
DOP	Veri kümesindeki adayların üniversite sınavına hazırlanırken dijital öğrenim platformlarından yararlanma durumunu ifade eder.
LISEIYETER	Veri kümesindeki adayların okuduğu liseyi yeterli bulma durumunu ifade eder.
CALISMA	Veri kümesindeki adayların okurken çalışma durumunu ifade eder.
AKRANZORBA	Veri kümesindeki adayların lise döneminde yaşanan akran zorbalığı durumunu ifade eder.
DANISMANLIK	Veri kümesindeki adayların okudukları lisenin rehberlik servisinin olma durumunu ifade eder.
DOGRUYER	Veri kümesindeki adayların lisedeki alan seçiminin doğruluğunu ifade eder.
MEVCUT	Veri kümesindeki adayların sınıf mevcudunu ifade eder.
DNOT	Veri kümesindeki adayların lise diploma notunu ifade eder.
TARIH	Veri kümesindeki adayların üniversite sınavına giriş tarihini ifade eder.
IL	Veri kümesindeki adayların lisede öğrenim gördüğü ili ifade eder.
SOSYALMEDYA	Veri kümesindeki adayların sosyal medya hesabı varlığını ifade eder.

SAGLIK	Veri kümesindeki adayların lise döneminde sağlık sorunu olma durumunu ifade eder.
ANNEDURUM	Veri kümesindeki adayların sınav öncesi annenin yaşama durumunu ifade eder.
BABADURUM	Veri kümesindeki adayların sınav öncesi babanın yaşama durumunu ifade eder.
GELIR	Veri kümesindeki adayların sınav öncesi ailenin gelir durumunu ifade eder.
ANNEEGITIM	Veri kümesindeki adayların sınav öncesi annenin eğitim durumunu ifade eder.
BABAEGITIM	Veri kümesindeki adayların sınav öncesi babanın eğitim durumunu ifade eder.
AILEMEDENI	Veri kümesindeki adayların sınav öncesi ailenin medeni durumunu ifade eder.
KARDES	Veri kümesindeki adayların sınav öncesi kardeş sayısı ifade eder.
UNIKARDES	Veri kümesindeki adayların sınav öncesi üniversitede okuyan kardeş durumunu ifade eder.
UNIZIYARET	Veri kümesindeki adayların sınav öncesi üniversiteye ziyaret etme durumunu ifade eder.
ONLISANS	Veri kümesindeki adayların sınavsız ön lisans geçiş durumunu ifade eder.
BARAJ	Veri kümesindeki adayların sınav barajını geçme durumu bu ifade eder.
UNIHAK	Veri kümesindeki adayların ilk tercihte istediği üniversiteye yerleşme hakkı kazanma durumunu ifade eder.
UNI	Veri kümesindeki adayların hangi üniversiteyi kazandığını ifade eder.
ANNECDURUM	Veri kümesindeki adayların sınav öncesi annenin çalışma durumunu ifade eder.
BABACDURUM	Veri kümesindeki adayların sınav öncesi babanın çalışma durumunu ifade eder.

Veri kümesinin oluşturulma aşamasında eğitimin eksik veriden etkilenmemesi için eksik olan verilerin bulunduğu satırlar silinmiştir. Toplamda 676 katılımcının olduğu veri kümesinden eksik veriler silinerek 615 veriye ulaşılmıştır. Veride gürültüyü azaltmak amacıyla düzeltmeler yapılması gereken öznitelikler seçilmiştir. Bu öznitelikler arasında diploma notu bilgisini içeren DNOT yüzlük sistemde, sınava giriş tarihini belirten öznitelik olan TARIH dört basamaklı yıl formatında ve eksik olan ders notları normalizasyon formülüne uygun şekilde düzeltilmiştir.

Ayrıca kategorik bilgiler içeren öznitelik verileri label encoder ve one hot encoder yöntemi ile veri madenciliği algoritmalarına uygun bir forma dönüştürülmüştür. Bu yöntemler verilerin sayısallaştırılmasına yarar. Veriler "one hot encoder" ile mevcut olan değer "1" , olmayan değer "0" alır. "label encoder" ile kategorik her veri sayısal bir değer alır. "Label Encoder" ile tekil değeri ikiden küçük veya eşit olan kolonlar "0" ve "1" ile doldurulmuştur.

Mevcut veri kümesi 50 adet özniteliğe sahiptir. Aralarındaki kategorik veriler "get dummies" yöntemi ile 144 özniteliğe dönüştürülmüştür. Bu yöntem ile uniq değeri 2'den fazla olan öznitelik değerleri sütuna alınıp satırda "1" ve "0" olarak yazdırılmıştır. Bu sayede veride object yapılı değer kalmaması sağlanmıştır.

1.2. Aşırı Uyum ve Sonrası Algoritmaların Değerlendirilmesi

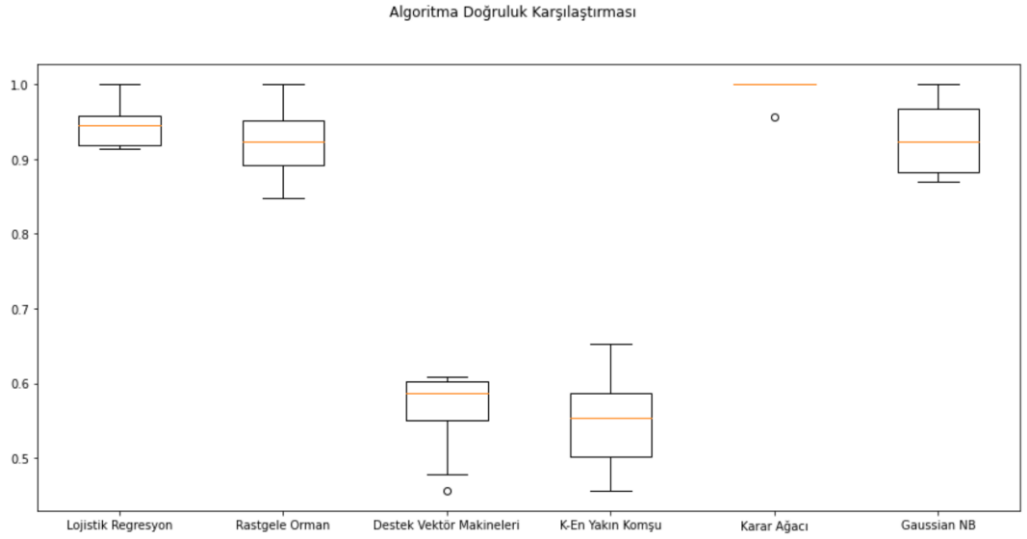
Veri kümesi ile yapılacak olan eğitim için araştırmalar ve testler yapılmıştır. Veri madenciliği yönteminde başarı oranının artması için veri seti ve algoritmalar test edilmiştir. Uygun olmayan özniteliklerle aşırı öğrenme (overfitting) gerçekleşmiştir. Başarı oranı oldukça yüksek çıkmıştır. Sebeplerinden biri algoritma eğitim verisinin en alt kırılımına kadar çalışıp mevcut veri seti üzerinde ezber gerçekleştirmesidir. Yapılan algoritmaların amacı her şeyi tahminlemesi değil genel bir doğru bulmasıdır. Diğer sebebi ise tercih edilen üniversiteyi kazanma durumunun tahminlemesinin yapılması istenen modelin, yerleşmiş olduğu üniversite bilgisine sahip olması yani 'UNI' özniteliğinin mevcut olmasıdır. Daha basit bir dille günlük hayattan örnek vermek gerekirse, marketteki alışveriş durumuna göre sepet analizi yapıp müşterinin soğan alıp almadığını tahmin etmek isteniyor. Modele soğan alıp almadığını tahmin etmesini isterken bilgi olarak taze, kuru gibi soğanla ilişkili olan bilgeler verilirse tahmin zor olmayacaktır ve aşırı öğrenme kaçınılmaz bir son olacaktır. Model

oluşturulurken 144 öznitelik kullanılmıştır. 615 verinin tamamı kullanılmıştır. Hedef alınan öznitelik olarak üniversite tercih etme hakkına sahip olma bilgisini taşıyan 'UNIHAK' özniteliği belirlenmiştir. Dışarı tutma (hold out) yöntemi ile 0.25 oranındaki veri test, kalan kısmı ise eğitim için ayrılmıştır. Modele 10 katlı çapraz doğrulama yapılmıştır. Sonuçlar incelendiğinde modelin aşırı öğrenme gerçekleştirdiği tespit edilmiştir.

Tablo 7. Aşırı Öğrenme Gösteren Algoritma Değerleri

Algoritma	ROC AUC Ort	ROC AUC STD	Doğruluk Ort	Doğruluk STD
Karar Ağacı	99.60	1.20	99.57	1.30
Lojistik Regresyon	97.83	2.17	96.10	3.04
Rastgele Orman	96.93	2.98	92.62	4.58
Gaussian NB	96.39	2.84	92.85	4.63
Destek Vektör Makineleri	60.02	10.86	58.11	8.79
K-En Yakın Komşu	57.13	11.77	55.26	11.40

Tablo 7. 'Aşırı Öğrenme Gösteren Algoritma Değerleri' adlı tablo incelendiğinde 'UNI' özniteliğinin sağladığı 99.60 'lık başarı oranı ile karar ağaçları algoritması aşırı öğrenme gösteren birinci model olmuştur. Şekil 20'de aşırı uyum gösteren algoritmaların doğruluk karşılaştırmaları yer almaktadır. Doğruluk karşılaştırmalarında sonuç 1'e yaklaştıkça başarı oranı artmıştır. Karar ağacı, lojistik regresyon, rastgele orman ve Gaussian NB algoritmalarındaki aşırı uyum şekilde 20'de belirgin bir şekilde görünmektedir. Destek vektör makineleri ve K-en yakın komşu algoritmalarında aşırı uyum görünmemiştir. Aşırı öğrenmeye bağlı ortalama hata değerleri, karışıklık matrisi ve sınıflandırma raporu Ek 4'te yer almaktadır.



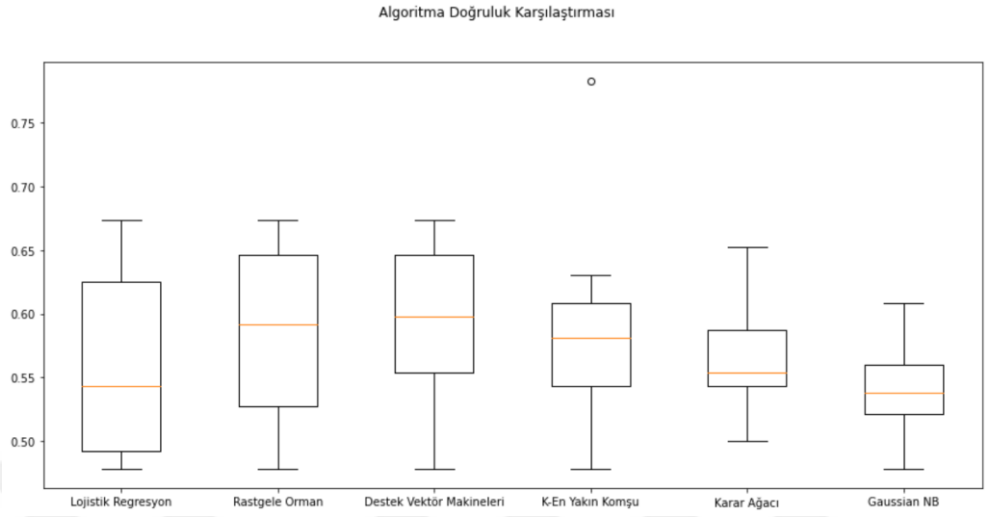
Şekil 20. Aşırı Uyumlu Algoritma Doğruluk Karşılaştırması

Modelin aşırı öğrenme gösterme nedeni yukarıda bahsedildiği gibi tercih edilen üniversite bilgisinin veri kümesi içinde yer almasından kaynaklanmaktadır. Veri kümesinden tercih edilen üniversite bilgisi olan 'UNI' özniteliği silindiğinde en yüksek değeri gösteren 61.78 ile Rastgele Orman algoritması olmuştur (Tablo 8). Aşırı öğrenme gösteren karar ağacı algoritması 'UNI' silindiğinde 56.86 değerine düşmüştür. Şekil 21'de algoritmaların doğruluk karşılaştırmaları yer almaktadır.

Tablo 8. Algoritma Değerleri

Algoritma	ROC AUC Ort	ROC AUC STD	Doğruluk Ort	Doğruluk STD
Rastgele Orman	61.78	11.16	58.14	6.93
K-En Yakın Komşu	60.09	8.77	58.58	8.12
Destek Vektör Makineleri	59.06	7.40	59.02	6.45
Lojistik Regresyon	58.05	11.13	56.20	7.53
Karar Ağacı	56.84	5.32	56.63	4.63
Gaussian NB	54.47	8.90	54.01	3.71

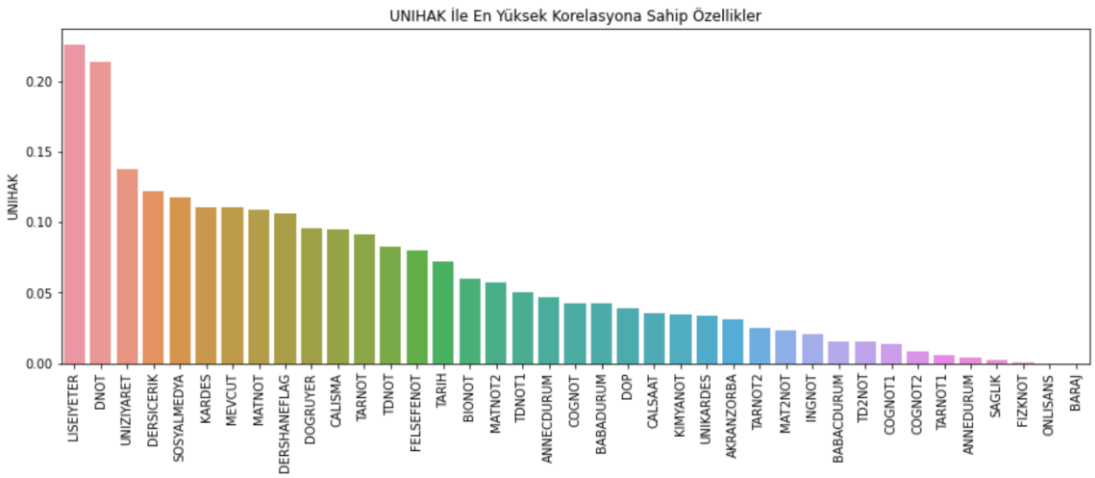
Şekil 20 ile Şekil 21 arasındaki tek fark 'UNI' özniteliğinin olmamasıdır. Aynı algoritmalarla öznitelik farklılığı ile elde edilen sonuçlar düşüş göstermiştir. Algoritmaların ortalama hata değerleri, karışıklık matrisi ve sınıflandırma raporu Ek 5'te yer almaktadır.



Şekil 21. Algoritma Doğruluk Karşılaştırması

2. FARKLI YÖNTEMLERLE ÖZNİTELİK SEÇİMİNİN MODELE ETKİSİ

Özniteliklerin seçimindeki amaç en etkin öznitelikleri seçerek yüksek performans elde etmektir. Birçok öznitelik çıkarım yöntemi araştırılmış ve test edilmiştir. Öznitelik seçim yöntemlerine ait algoritma sonuçlarına Ek 6'dan ulaşabilir. Şekil 22'de yer alan hedef alınan öznitelik UNIHAK ile en yüksek korelasyona sahip öznitelikler yer almaktadır. Öznitelik yöntemleri uygulanırken 20, 30, 40, 50 öznitelik, k-kat ile k=2,4,6,8,10'a kadar çapraz doğrulama yapılmıştır. Dışarıda tutma (holdout) ile Eğitim seti %60, %70, %75,%80 oranlarında olacak şekilde ayrılmıştır. Veri setleri her modele girişinde, rastgele belirli oranlarla karıştırılarak test edilmiştir.



Şekil 22. UNIHAK ile En Yüksek Korelasyona Sahip Öznitelikler

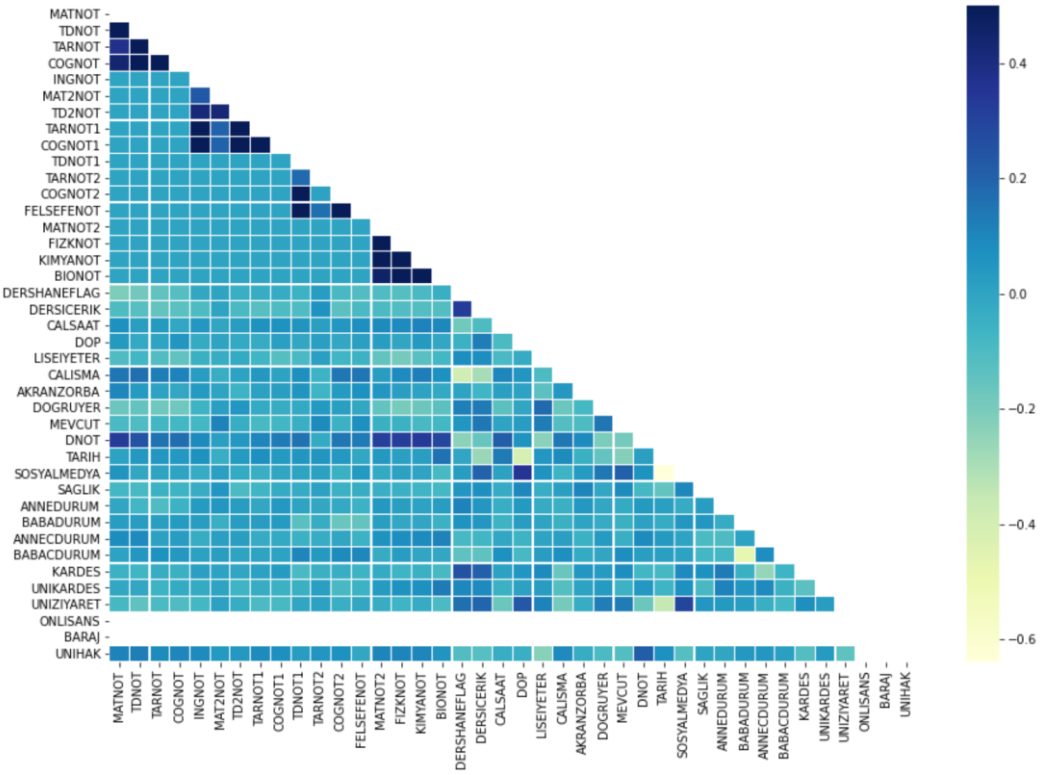
Şekil 22'de hedef öznitelik olan UNIHAK ile en yüksek korelasyon dağılımını gözlemleyebilirken Tablo 9'da birbiri ile yüksek korelasyon sağlayan, aralarında güçlü bir ilişki olan öznitelikleri gözlemleyebiliyoruz. Güçlü ilişkisi olan öznitelikler elde edilen modelin başarısını arttırır.

Tablo 9. Birbiri ile En Yüksek Korelasyona Sahip Öznitelikler

Öznitelik 1	Öznitelik 2	Korelasyon Kat Sayısı
TARNOT1	COGNOT1	0.861017
INGNOT	TARNOT1	0.834105
TARNOT	COGNOT	0.756666
INGNOT	COGNOT1	0.747938
LISETUR_Mesleki ve Teknik Anadolu Liseleri	IL_Bilecik	0.706531
MATNOT2	ALAN_Sayısal	0.688062
AILEMEDENI_Boşanmış	AILEMEDENI_Evli	0.685773
BIONOT	ALAN_Sayısal	0.662117
TARIH	SOSYALMEDYA	0.638505

2.1. Pearson Korelasyon Yöntemi ile Öznitelik Çıkarımı

Pearson korelasyon yöntemi, değişken seçimi yönteminin filtreleme seçeneklerinden biridir. Özniteliklerin ilişkili olduğu veri setlerinde ilişkinin değişimini ölçmeye yarar. Aldığı değerler -1 ile 1 arasındadır. İlişki 1 değerine yaklaştıkça mükemmelleşir. Veri kümesindeki öznitelikler arasındaki korelasyon modelin başarını etkilemektedir. Korelasyon değerleri incelendiğinde hedef alınan öznitelik ile yüksek korelasyona sahip öznitelikler Şekil 22'de yer almaktadır. Tablo 9'de ise birbiri ile en yüksek korelasyona sahip olan öznitelikler yer almaktadır. Genel anlamda korelasyon dağılımı tüm öznitelikler için Şekil 23. korelasyon grafiğinde yer almaktadır. Bu grafiği incelediğimizde renkler koyulaştıkça aradaki korelasyonda artmaktadır. Negatif değer alan öznitelikler arasında negatif bir korelasyon vardır.



Şekil 23. Korelasyon Grafiği

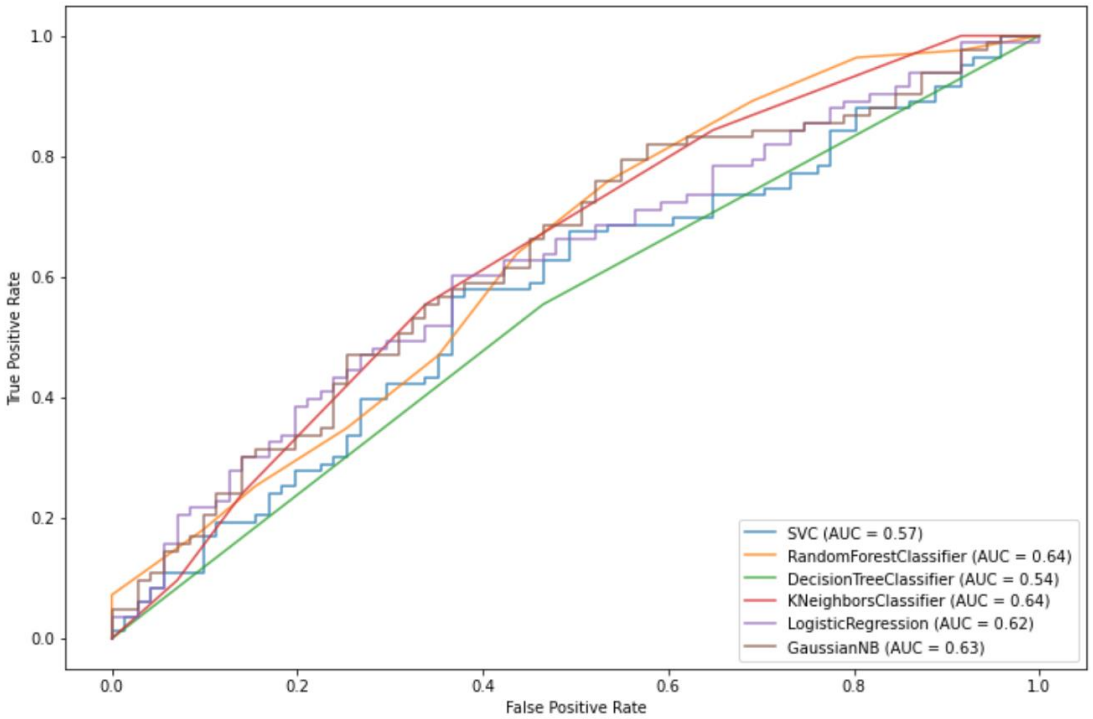
Korelasyon değerlerine bakılarak model yeni yönteme göre oluşturulmuştur. Yeni yöntemde yüksek korelasyona sahip özniteliklerle model tekrardan oluşturtulup, korelasyonu düşük olan öznitelikler modelden çıkartılmıştır. Sırasıyla korelasyon değeri 0.5, 0,6, 0,7 olan 20, 30, 40, 50, 60, 70, 80 tane öznitelikle 0.75 eğitim verisi ayrılarak, 10 katlı çapraz doğrulama yapılarak test edilmiştir. Elde edilen sonuçlara göre 0.7 ve altı korelasyon değerine sahip olan 30 öznitelik ile Rastgele Orman algoritması bu yöntem için en başarılı sonucu vermiştir.

Rastgele Orman analizi incelendiğinde ortalama ROC AUC değeri 71.09 , ortalama F skoru 0.64 ve ortalama karesel hata değeri 0.370130 bulunmaktadır. Her öznitelik çıkarım yöntemi ve öznitelik sayısı için lojistik regresyon, karar ağacı, Gaussian NB, rastgele orman, destek vektör makineleri, K-en yakın komşu algoritmaları test edilmiş ve karşılaştırmalar yapılmıştır. Sonuca en yüksek etkisi olan öznitelik "LISEIYETER" olarak değerlendirilmiştir. Tablo 10'da testler sonucunda en yüksek değeri veren algoritmaların değerleri yer almaktadır. Ek 5'te performans tabloları ve kullanılan öznitelik bilgileri ayrıntılı olarak sunulmuştur.

Tablo 10. Korelasyon Yöntemi Öznitelik Çıkarım Sonucu En Yüksek Algoritma Değerleri

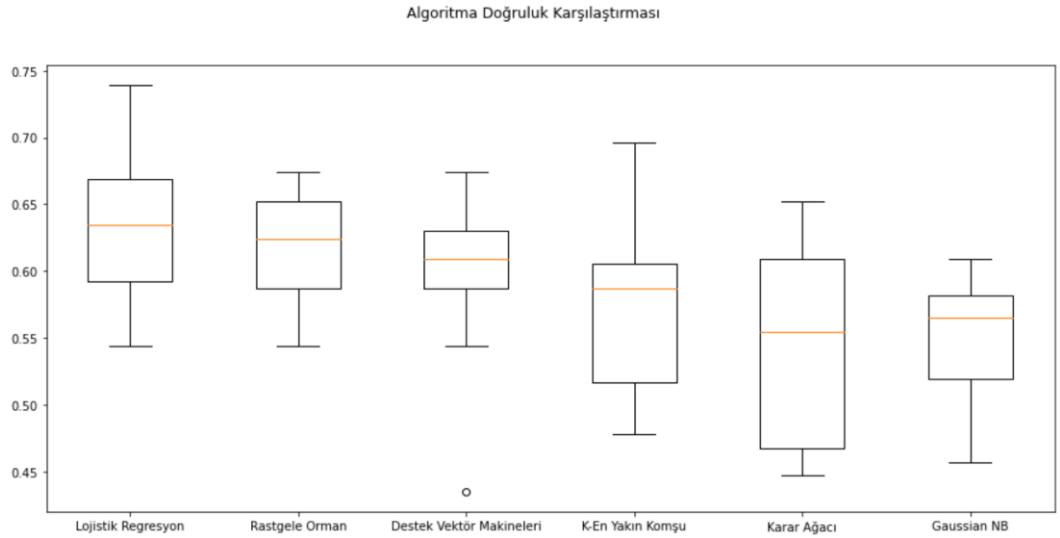
Algoritma	ROC AUC Ort.	ROC AUC Std.	Doğruluk Ort.	Doğruluk Std.	Öznitelik Sayısı	Korelasyon Değeri
Rastgele Orman	71.09	8.47	66.16	3.90	30	0.7
Gaussian NB	70.36	6.87	53.58	6.63	40	0.5
Lojistik Regresyon	67.90	6.24	62.27	5.47	40	0.5
Destek Vektör Makineleri	65.91	9.12	61.81	6.35	80	0.6
Karar Ağacı	62.68	6.46	62.69	6.27	20	0.5
K-En Yakın Komşu	59.55	8.85	58.38	9.33	30	0.5

ROC eğrisi gerçek pozitif ve yanlış pozitife denk gelen noktaların birleşimidir. Şekil 24'te korelasyon yöntemi öznitelik seçimi sonucu oluşan ROC eğrileri incelendiğinde, köşegen çizme eğilimi olan karar ağacı algoritması en düşük değerlerden birini vermiştir.



Şekil 24. Korelasyon Yöntemi Öznitelik Çıkarımı Sonucu ROC Eğrisi

Kullanılan algoritmaların analizinden elde edilen Tablo 10'daki doğruluk değerlerinin standarttı ve ortalaması yer alırken Şekil 25'te doğruluk değerinin görselleştirilmiş şekli yer almaktadır.



Şekil 25. Korelasyon Yöntemi Öznitelik Çıkarımı Sonucu Algoritmaların Doğruluk Karşılaştırması

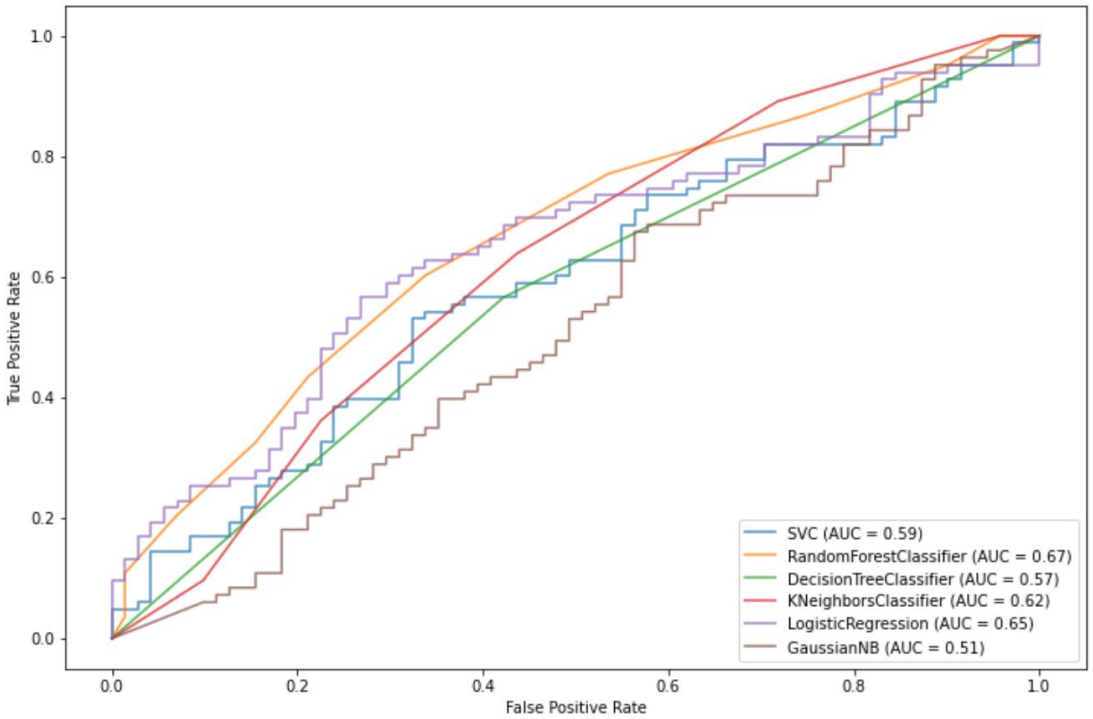
2.2. Anova Yöntemi ile Öznitelik Çıkarımı

Bağımlı değişkenin, bağımsız değişkenler üzerine etkisini araştıran istatistiksel tabanlı varyansların analizidir. Anova yöntemi uygulanırken 20, 30, 40, 50, 60, 70, 80 öznitelik, k-kat ile k=2,4,6,8,10'a kadar çapraz doğrulama yapılmıştır. Dışarıda tutma (holdout) ile Eğitim seti %60, %70, %75,%80 oranlarında olacak şekilde ayrılmıştır. Veri setleri her modele girişinde rastgele belirli oranlarla karıştırılarak test edilmiştir. Elde edilen sonuçlara göre 80 öznitelik ile %75 eğitim verisi ayrılan ve k=10 k katlı çapraz doğrulama ile Rastgele Orman algoritması bu yöntem için en başarılı sonucu vermiştir.

Rastgele Orman analizi incelendiğinde ortalama ROC AUC değeri 72.96 , ortalama F skoru 0.63 ve ortalama karesel hata değeri 0.409091 bulunmaktadır. Her öznitelik çıkarım yöntemi ve öznitelik sayısı için lojistik regresyon, karar ağacı, Gaussian NB, rastgele orman, destek vektör makineleri, K-en yakın komşu algoritmaları test edilmiş ve karşılaştırmalar yapılmıştır. Tablo 15'te testler sonucunda en yüksek değeri veren algoritmaların değerleri yer almaktadır. Ek 5'te performans tabloları ve kullanılan öznitelik bilgileri ayrıntılı olarak sunulmuştur.

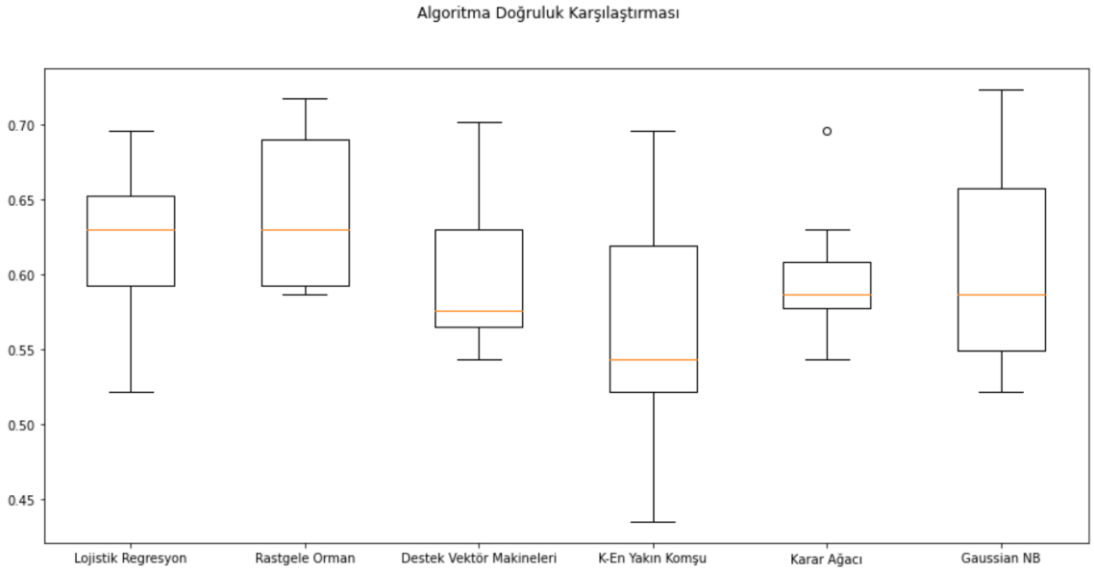
Tablo 11. Anova Yöntemi Öznitelik Çıkarım Sonucu En Yüksek Algoritma Değerleri

Algoritma	Yüzde	K Değeri	ROC AUC Ort	ROC AUC STD	Doğruluk Ort.	Doğruluk STD	Öznitelik Sayısı
Rastgele Orman	25_75	10	72.96	5.27	63.52	4.96	80
Gaussian NB	25_75	10	71.76	8.37	55.72	6.91	40
Lojistik Regresyon	25_75	10	69.35	6.28	63.11	5.89	20
Destek Vektör Makineleri	25_75	2	64.30	5.44	60.95	0.08	80
Karar Ağacı	25_75	10	63.18	6.00	63.13	5.81	70
K-En Yakın Komşu	25_75	2	59.42	0.36	59.43	7.66	30



Şekil 26. Anova Yöntemi Öznitelik Çıkarımı Sonucu ROC Eğrisi

Şekil 26'daki ROC eğrisi incelendiğinde iyi sonuçlar veren eğimler çizilmiştir.



Şekil 27. Anova Yöntemi Öznitelik Çıkarımı Sonucu Algoritmaların Doğruluk Karşılaştırması

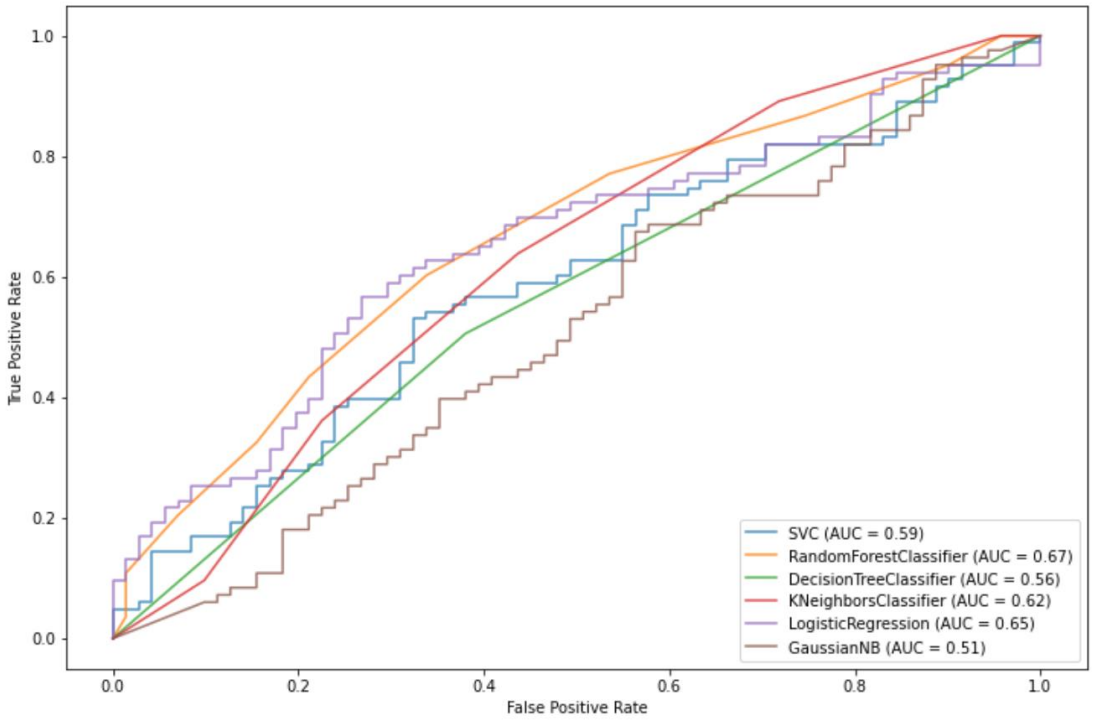
2.3. Ki-Kare Yöntemi ile Öznitelik Çıkarımı

Ki-kare yöntemi ile öznitelik çıkarımı, iki kategorik öznitelik arasında istatiksel ilişkiyi ölçer ve sonucunda özellikleri seçer. Amaç en etkili olan öznitelikleri seçmektir. Seçilen öznitelikler ile alt küme oluşturulur ve modelde uygulanır. Ki-kare yöntemi uygulanırken 20, 30, 40, 50, 60, 70, 80 öznitelik, k-kat ile k=2,4,6,8,10'a kadar çapraz doğrulama yapılmıştır. Dışarıda tutma (holdout) ile Eğitim seti %60, %70, %75,%80 oranlarında olacak şekilde ayrılmıştır. Veri setleri her modele girişinde rastgele belirli oranlarla karıştırılarak test edilmiştir. Elde edilen sonuçlara göre 40 öznitelik ile %70 eğitim verisi ayrılan ve k=2 k katlı çapraz doğrulama ile Gaussian NB algoritması bu yöntem için en başarılı sonucu vermiştir.

Gaussian NB analizi incelendiğinde ortalama ROC AUC değeri 73.77 ortalama F skoru 0.62 ve ortalama karesel hata değeri 0.467532 bulunmaktadır. Her öznitelik çıkarım yöntemi ve öznitelik sayısı için lojistik regresyon, karar ağacı, Gaussian NB, rastgele orman, destek vektör makineleri, K-en yakın komşu algoritmaları test edilmiş ve karşılaştırmalar yapılmıştır. Tablo 18'de testler sonucunda en yüksek değeri veren algoritmaların değerleri yer almaktadır. Ek 5'te performans tabloları ve kullanılan öznitelik bilgileri ayrıntılı olarak sunulmuştur.

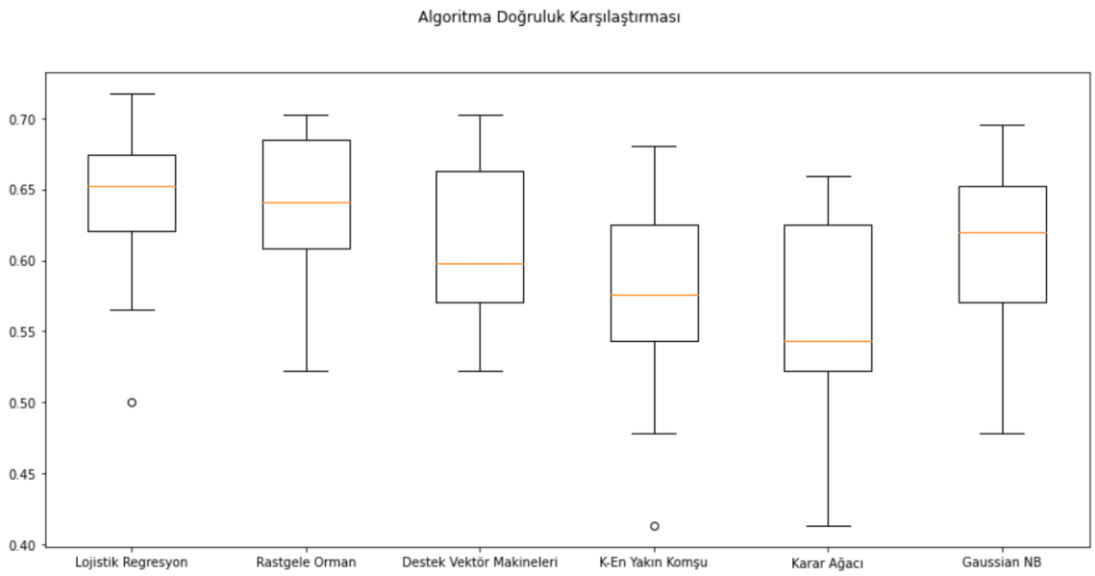
Tablo 12. Ki-Kare Yöntemi Öznitelik Çıkarım Sonucu En Yüksek Algoritma Değerleri

Algoritma	Yüzde	K Değeri	ROC AUC Ort.	ROC AUC STD	Doğruluk Ort.	Doğruluk STD	Öznitelik Sayısı
Gaussian NB	30_70	2	73.77	6.06	60.94	5.97	40
Rastgele Orman	40_60	6	73.15	7.00	66.15	6.12	80
Lojistik Regresyon	40_60	4	72.16	6.27	65.07	4.34	30
Destek Vektör Makineleri	40_60	2	64.62	5.67	60.72	5.97	50
Karar Ağacı	25_75	10	62.05	6.00	62.05	5.81	80
K-En Yakın Komşu	40_60	4	60.94	8.17	58.32	7.68	70



Şekil 28. Ki-Kare Yöntemi Öznitelik Çıkarımı Sonucu ROC Eğrisi

Şekil 28'de yer alan ROC eğrileri incelendiğinde karar ağacı algoritması düşük performans göstererek köşegen çizme eğilimindedir. Rastgele orman algoritması daha fazla gerçek pozitif değerler elde etmiştir.



Şekil 29. Ki-Kare Yöntemi Öznitelik Çıkarımı Sonucu Algoritmaların Doğruluk Karşılaştırması

3. FARKLI ALGORİTMALARIN PERFORMANS KARŞILAŞTIRMALARI

Veri kümesi üzerinde uygulanan yöntemlerin performans kriterleri açısından kıyaslanması sonucunda en iyi performansı gösteren model belirlenecektir. Çalışma kapsamında yapılan 1833 testten elde edilen performans detaylarının ilk sonuçları Ek 6'da Tablo 24' de sunulmaktadır. Tablo 24'te elde edilen her değer için 20, 30, 40, 50, 60, 70, 80 öznitelik sayısı için Karar ağacı, Gaussian NB, lojistik regresyon, rastgele orman, destek vektör makineleri ve K-en yakın komşu algoritmaları kullanılmış ve sonuçlar karşılaştırılmıştır.

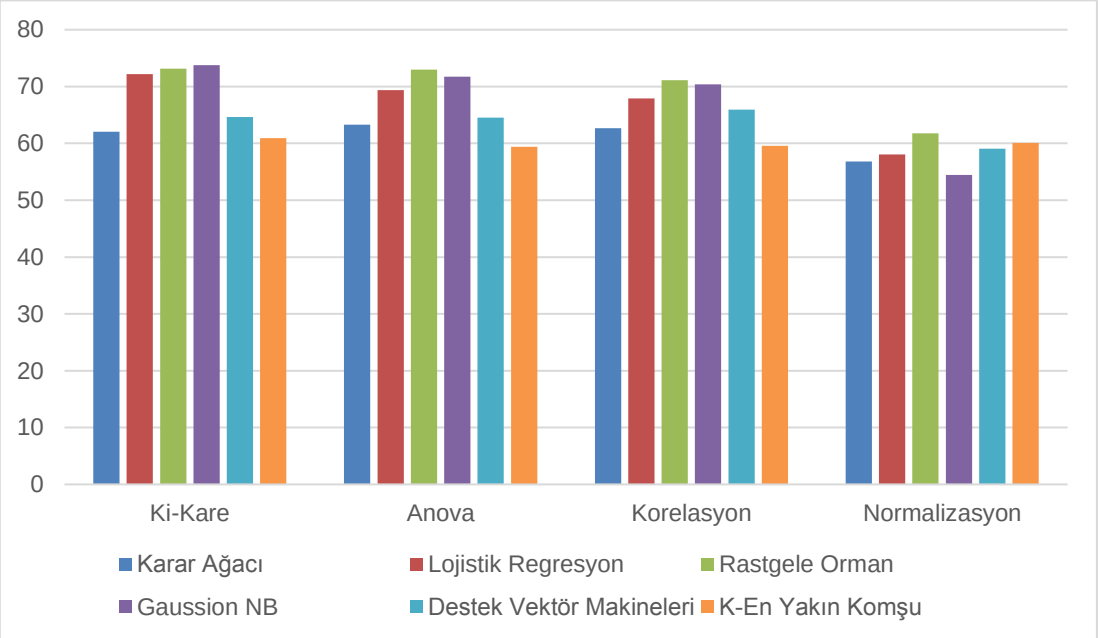
Tablo 13.En İyi Sonucu Veren Algoritmaların ve Yöntemlerin Karşılaştırılması

Algoritma	Yüzde	K	ROC AUC Ort.	ROC AUC STD	Doğruluk Ort.	Doğruluk STD	Yöntem	Öznitelik Sayısı
Gaussian NB	30_70	2	73.77	44353	60.94	35551	Ki-Kare	40
Rastgele Orman	40_60	6	73.15	33695	66.15	27851	Ki-Kare	80
Lojistik Regresyon	40_60	4	72.16	34486	65.07	28307	Ki-Kare	30
Destek Vektör Makineleri	30_70	10	65.91	9.12	61.81	6.35	Pearson	80
Karar Ağacı	25_75	10	63.18	6.00	63.13	5.81	Anova	20
K-En Yakın Komşu	40_60	4	60.94	8.00	58.32	43647	Ki-Kare	70

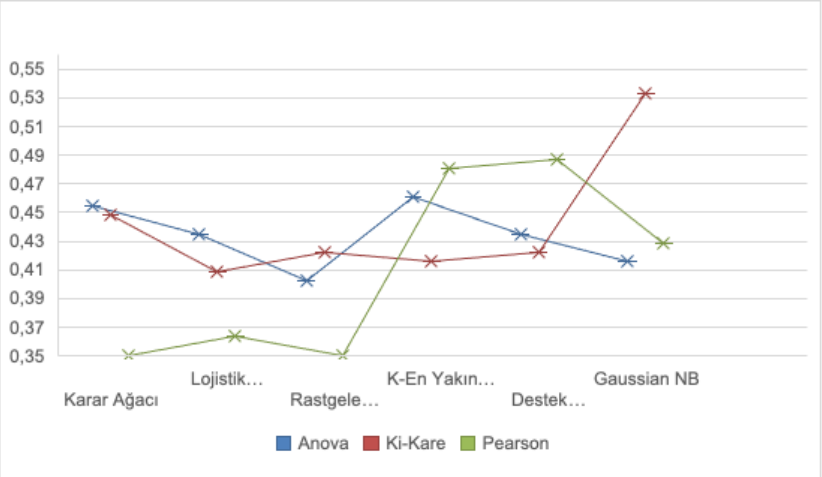
Ki-kare öznitelik çıkarım yöntemi ile 20, 30, 40, 50, 60, 70, 80 öznitelik, k-kat ile k=2,4,6,8,10'a kadar çapraz doğrulama yapılmıştır. Dışarıda tutma (holdout) ile Eğitim seti %60, %70, %75,%80 oranlarında ayrılmıştır. Elde edilen sonuçlara göre ki-kare yönteminde 40 öznitelik ile yapılan testte Gaussian NB algoritması en başarılı sonucu oluşturmuştur. Öznitelik sayısı 80 olduğunda rastgele orman algoritması önceki sonucuna göre ki- kare yöntemi ile daha başarılı sonuca ulaşılmıştır. Öznitelik sayısı 30 olduğunda lojistik regresyon algoritması önceki sonucuna göre ki- kare

yöntemi ile daha başarılı sonuç oluşturmuştur. Destek vektör makineleri algoritması ise en başarılı sonucunu Pearson korelasyon yöntemi ile seçilen 80 öznitelik ile almıştır. İlk sonucuna göre daha yüksek başarı elde etmiştir. Karar ağacı algoritması anova yöntemi ve 20 öznitelikle önceki sonucuna göre daha başarılı olmuştur.

Tüm öznitelik çıkarım yöntemleri kapsamında yapılan testlerde en iyi sonuç ki-kare yöntemi ile seçilmiş 40 özniteliğin olduğu Gaussian NB algoritması ile alınmış ve diğer yöntemlere göre daha başarılı olmuştur. Yapılan test sonucuna göre ROC AUC ortalaması 73.77'dir. Ortalama karesel hata değeri 0.532468'dir. F1 puanı 0.62'dir. Diğer algoritmaların da yer aldığı karışıklık matrisi ve sınıflandırma raporuna tablo 22'den ulaşılabilir. Bu algoritmaların doğruluk karşılaştırmalarına ise Şekil 30'tan ulaşılabilir.



Şekil 30. En İyi Sonucu veren Algoritmaların Doğruluk Değerleri Karşılaştırması



Şekil 31. Algoritmaların Yöntemlere Göre Ortalama Mutlak Hata Değeri Değişimi

3.1. Karar Ağacı Algoritmasından Elde Edilen Kurallar

Veri setine uygulana karar ağacı algoritmasıyla ilgili karşılaştırmalara bir önceki incelemeden ulaşabilir. Bu bölümde Karar ağacından elde edilen birtakım kurallara değinilecektir. Toplam 107 yaprak sayısına ve 15 derinliğe sahip ağaç yapısının, 5 derinlikte 24 yaprağının incelenmesi yapılacaktır.

```
|--- LISEIYETER <= 0.50
|   |--- DNOT <= 66.00
|   |   |--- MATNOT <= 4.31
|   |   |   |--- class: 0.0
|   |   |--- MATNOT >  4.31
|   |   |   |--- class: 1.0
|   |--- DNOT >  66.00
|   |   |--- CALISMA <= 0.50
|   |   |   |--- AKRANZORBA <= 0.50
|   |   |   |   |--- class: 1.0
|   |   |   |--- AKRANZORBA >  0.50
|   |   |   |   |--- DNOT <= 78.50
|   |   |   |   |   |--- class: 1.0
|   |   |   |   |--- DNOT >  78.50
|   |   |   |   |   |--- class: 0.0
|   |   |--- CALISMA >  0.50
|   |   |   |--- DANISMANLIK_Hayır <= 0.50
|   |   |   |   |--- BABAEGITIM_Okumadı <= 0.50
|   |   |   |   |   |--- class: 1.0
|   |   |   |   |--- BABAEGITIM_Okumadı >  0.50
```

```
|   |   |   |   |   |--- class: 0.0
|   |   |   |--- DANISMANLIK_Hayır >  0.50
|   |   |   |   |--- UNIZIYARET <= 0.50
|   |   |   |   |   |--- class: 1.0
|   |   |   |   |--- UNIZIYARET >  0.50
|   |   |   |   |   |--- class: 1.0
|--- LISEIYETER >  0.50
|   |--- MATNOT2 <= 4.11
|   |   |--- DANISMANLIK_Hayır <= 0.50
|   |   |   |--- CALISMA <= 0.50
|   |   |   |   |--- class: 0.0
|   |   |   |--- CALISMA >  0.50
|   |   |   |   |--- MEVCUT <= 41.00
|   |   |   |   |   |--- class: 1.0
|   |   |   |   |--- MEVCUT >  41.00
|   |   |   |   |   |--- class: 0.0
|   |   |--- DANISMANLIK_Hayır >  0.50
|   |   |   |--- IL_Manisa <= 0.50
|   |   |   |   |--- MATNOT2 <= 1.50
|   |   |   |   |   |--- class: 1.0
|   |   |   |   |--- MATNOT2 >  1.50
|   |   |   |   |   |--- class: 0.0
|   |   |   |--- IL_Manisa >  0.50
|   |   |   |   |--- KIMYANOT <= 3.50
|   |   |   |   |   |--- class: 0.0
|   |   |   |   |--- KIMYANOT >  3.50
|   |   |   |   |   |--- class: 1.0
|   |--- MATNOT2 >  4.11
|   |   |--- DNOT <= 71.50
|   |   |   |--- IL_İstanbul <= 0.50
|   |   |   |   |--- CINSIYET_Kadın <= 0.50
|   |   |   |   |   |--- class: 0.0
|   |   |   |   |--- CINSIYET_Kadın >  0.50
|   |   |   |   |   |--- class: 0.0
|   |   |   |--- IL_İstanbul >  0.50
|   |   |   |   |--- TDNOT <= 4.56
|   |   |   |   |   |--- class: 1.0
|   |   |   |   |--- TDNOT >  4.56
|   |   |   |   |   |--- class: 0.0
|   |   |--- DNOT >  71.50
|   |   |   |--- TARNOT2 <= 4.26
|   |   |   |   |--- CINSIYET_Kadın <= 0.50
|   |   |   |   |   |--- class: 1.0
|   |   |   |   |--- CINSIYET_Kadın >  0.50
|   |   |   |   |   |--- class: 0.0
|   |   |   |--- TARNOT2 >  4.26
|   |   |   |   |--- UNIKARDES <= 0.50
|   |   |   |   |   |--- class: 0.0
|   |   |   |   |--- UNIKARDES >  0.50
|   |   |   |   |   |--- class: 1.0

Ağacın Yaprak Sayısı : 107
Ağacın Derinliği : 15
```

Karar ağacının düğümleri, niteliğin kazandırdığı bilgi kazancı (information gain) ile oluşmaktadır. Yukarıda ağaç yapısının program kural çıktısı incelendiğinde dallanmanın liseyi yeterli bulma durumunu ifade eden LISEIYETER ile başladığı görülmektedir. Bu durum bu özniteliğin en yüksek bilgi kazancı sağlayan düğüm olduğunu göstermektedir. Şekil 22 'te veri setinin en yüksek bilgi kazançlarına sahip öznitelikleri sırayla yer almaktadır.

Karar Ağacı yapısı analiz edilirken elde edilen veri setine ait birtakım kurallar: **KURAL 1: EĞER** öğrenci mezun olduğu liseyi yeterli buluyor ya da bulmuyorsa **(LISEIYETER <= 0.5)** ve diploma notu 66'a eşit ya da küçükse **(DNOT <= 66.00)** ve matematik notu 4.31'den küçük ya da eşitse **(MATNOT <= 4.31)** üniversiteye yerleştirilmeye hak kazanmaz **(class: 0.0).**

KURAL 2 : EĞER öğrenci mezun olduğu liseyi yeterli buluyor ya da bulmuyorsa **(LISEIYETER <= 0.5)** ve diploma notu 66'a eşit ya da küçükse **(DNOT <= 66.00)** ve matematik notu 4.31'den büyük ise **(MATNOT > 4.31)** üniversiteye yerleştirilmeye hak kazanır **(class: 1.0).**

KURAL 3 : EĞER öğrenci mezun olduğu liseyi yeterli buluyor ya da bulmuyorsa **(LISEIYETER <= 0.5)** ve diploma notu 66'dan büyükse **(DNOT > 66.00)** ve bir iş yerinde çalışmıyorsa **(CALISMA <= 0.50)** akran zorbalığı yaşamadıysa **(AKRANZORBA <= 0.50)** üniversiteye yerleştirilmeye hak kazanır **(class: 1.0)**.

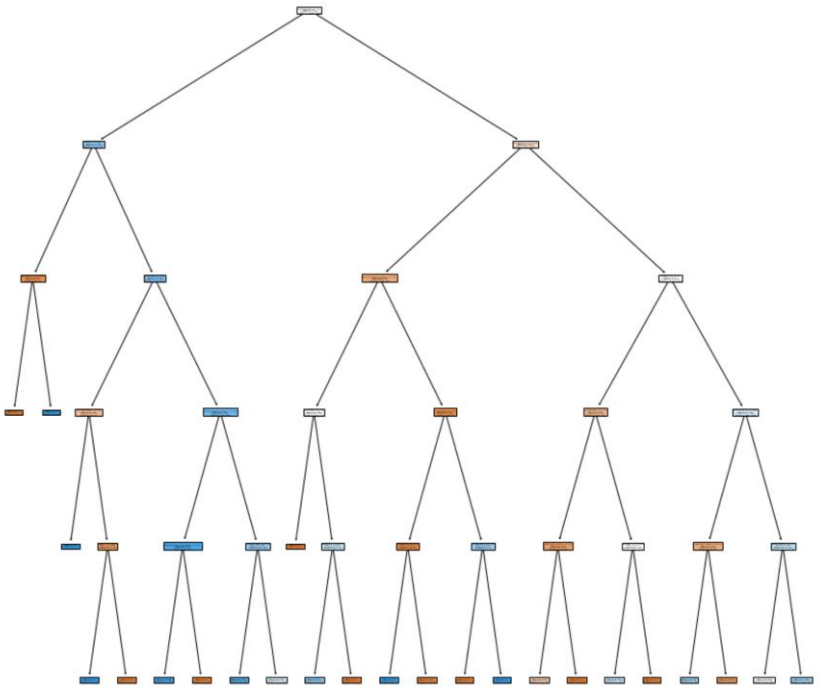
KURAL 4 : EĞER öğrenci mezun olduğu liseyi yeterli buluyor ya da bulmuyorsa **(LISEIYETER <= 0.5)** ve diploma notu 66'dan büyükse **(DNOT > 66.00)** ve bir iş yerinde çalışmıyorsa **(CALISMA <= 0.50)** akran zorbalığı yaşadıysa **(AKRANZORBA > 0.50)** ve diploma notu 78.50'a eşit ya da küçükse **(DNOT <= 78.50)** üniversiteye yerleştirilmeye hak kazanır **(class: 1.0)**.

KURAL 5 : EĞER öğrenci mezun olduğu liseyi yeterli buluyor ya da bulmuyorsa **(LISEIYETER <= 0.5)** ve diploma notu 66'dan büyükse **(DNOT > 66.00)** ve bir iş yerinde çalıştıysa **(CALISMA >0.50)** ve lisede danışmanlık aldıysa ya da almadıysa **(DANISMANLIK_Hayır <= 0.50)** ve baba eğitimli ya da eğitimsiz ise

(BABAEGITIM_Okumadı <= 0.50) üniversiteye yerleştirilmeye hak kazanır **(class: 1.0)**.

KURAL 6 : EĞER öğrenci mezun olduğu liseyi yeterli buluyor ya da bulmuyorsa **(LISEIYETER <= 0.5)** ve diploma notu 66'dan büyükse **(DNOT > 66.00)** ve bir iş yerinde çalıştıysa **(CALISMA >0.50)** ve lisede danışmanlık aldıysa ya da almadıysa **(DANISMANLIK_Hayır <=0.50)** ve baba eğitimsiz ise **(BABAEGITIM_Okumadı > 0.50)** üniversiteye yerleştirilmemeye hak kazanmaz **(class: 0.0)**.

KURAL 7 : EĞER öğrenci mezun olduğu liseyi yeterli buluyor ya da bulmuyorsa **(LISEIYETER <= 0.5)** ve diploma notu 66'dan büyükse **(DNOT > 66.00)** ve bir iş yerinde çalıştıysa **(CALISMA >0.50)** ve lisede danışmanlık almadıysa **(DANISMANLIK_Hayır >0.50)** ve lisede üniversite ziyaretlerine katıldı ise **(UNIZIYARET > 0.50)** üniversiteye yerleştirilmemeye hak kazanır **(class: 1.0)**.



Şekil 32. Kurallara Ait Ağaç Yapısı

SONUÇ

Öğrencilerin istediği düzeydeki başarıya ulaşabilmeleri ve hedeflerindeki yerlere gelebilmeleri için bir sonraki aşamaya geçmeleri önemli bir etken ise, bu etkenin araştırılması, başarı ya da başarısızlık kalıplarını ortaya çıkaran ilişkiler üzerinden keşfedilmesi, sürecin ortaklarına özellikle eğitimcilere çok önemli bir bilgi sağlayacaktır. Bu çalışmada lise mezunu bireylerin sosyo-demografik özellikleri, akademik başarılarını etkileyen faktörleri baz alınarak üniversiteye yerleşmeyi hak kazanma durumunu tahminlemek için veri madenciliği yöntemleri uygulanmıştır. Bu özellikler oluşturulurken akademik başarıyı etkileyen faktörlerle ilgili literatür taramaları yapılmıştır. Veri seti Türkiye sınırları kapsamındaki çeşitli illerden ankete katılan sosyo-kültürel farklılıklar gösteren kişilerden, homojen bir dağılım olmayacak şekilde ülkemizi temsil edebilecek nitelikte bilgiler elde edilmiştir. Bu çalışma, eğitsel veri madenciliği konusunda yapılan diğer çalışmalardan şu açılardan farklılıklar göstermektedir:

- Literatürde, çoğu çeşitli düzeylerde istatistiksel analize dayanan birçok başarılı çalışma vardır. Bu çalışmada ise, analiz için istatistiksel ve matematiksel yöntemleri birleştirerek veri madenciliği yöntem ve teknikleri kullanılmıştır.
- Araştırılan çalışmalara göre modellere ait performans kriterleri daha geniş aralıkta dışarı tutma (hold out) ve k-katlı çapraz geçerleme ile elde edilen sonuçlara göre öğrenci başarıları açıklanmıştır.
- Modellerin sonuçları karşılaştırılırken, doğruluk ve hata metrikleri dışında f puanı ve sınıflandırma raporları dikkate alınmıştır. Araştırılan çalışmalarda bu şekilde ayrıntılı incelemelere rastlanmamıştır.
- Çalışmalarda bağımlı ve bağımsız değişken arasındaki ilişki dışında hedef özelliği etkileyen , niteliğin değeri tahmin edilebilir durumdadır.
- Daha önce yapılmış ve literatürde yer alan diğer çalışmaların aksine davranışsal, sosyo-demografik özellikler, sosyo-ekonomik ve kültürel özellikler bir bütün olarak ele alınmıştır.

Çalışma kapsamında Karar Ağacı (CART), Gaussion NB, Lojistik Regresyon, K-En Yakın Komşu, Rastgele Orman ve Destek Vektör Makineleri Algoritmalarından faydalanılmıştır. En yüksek başarı gösteren algoritma Gaussion NB ve en iyi sonucu veren yöntem Ki-Kare öznitelik çıkarımı yöntemi ile ROC AUC oranı 73.77 elde edilmiştir . Bu sonuç 40 öznitelik ile elde edilmiştir. Özniteliklerin detaylı listesi Ek 8'de yer almaktadır. Öznitelik sayısı 80 olduğunda rastgele orman algoritması önceki sonucuna göre Ki- kare yöntemi ile daha başarılı sonuca ulaşılmıştır. Öznitelik sayısı 30 olduğunda lojistik regresyon algoritması önceki sonucuna göre Ki- kare yöntemi ile daha başarılı sonuç oluşturmuştur. Destek vektör makineleri algoritması ise en başarılı sonucunu Pearson korelasyon yöntemi ile seçilen 80 öznitelik ile almıştır. İlk sonucuna göre daha yüksek başarı elde etmiştir. Karar ağacı algoritması Anova yöntemi ve 20 öznitelikle önceki sonucuna göre daha başarılı olmuştur. Kullanılan algoritmaların ortalama mutlak hataları incelendiğinde 0.35-0.53 arasında farklılıklar gösterdiği görülmektedir. Hata değeri sıfıra yaklaştıkça sonuç başarılı olarak nitelendirilmiştir. Çalışma süresince Ek 9'da bazı örnekleri verilen 1833 test yapılmıştır. Yapılan testler sonucunda uygulanan öznitelik seçimi yöntemlerinden en başarılı sonucu veren yöntem Ki-kare olmuştur.

Çalışmanın analizleri Python programlama dili kullanılarak Jupiter Notebook programlama ortamında oluşturulmuştur. Python veri madenciliğinde en çok kullanılan 2.dildir ve uzun kod satırları yazılmasına gerek kalmadan birkaç satırda çözüm sağlar. İstatistiksel analizleri çözmede daha kabiliyetli olması çok tercih edilen dillerden biri olmasını sağlamıştır. Bunun yanı sıra zengin bir kullanıcı topluluğuna sahip olması, yeni başlayanından uzmanına kadar birçok çözüm için kaynaklara ulaşım kolaylığı sunar ve orta ölçekli işler için en uygun dildir.

Veri kümesi 676 mezun öğrenci bilgisini bulundurmakta olup eksik veriler silinerek 615 veri üzerinde gözlem yapılmıştır. Veriyi her test edişte k=2,4,6,8,10 kat çapraz geçerleme ve dışarı bırakma (hold out) yöntemi ile %80, %75, %70, %60 eğitim verisi ayrılarak veri kümesi oluşturulmuştur.

Bu çalışmadaki örnek grubunda ve kullanılan yöntemlerde bazı kısıtlamalar vardır. Bunlar:

- Örnek grubu olarak Türkiye sınırları içindeki okullardan mezun olan kişilerin bilgileri alınarak sınırlandırılmıştır.
- Sınıflandırma tabanlı veri madenciliği yöntemlerinden Karar ağacı, rastgele orman, Gaussian NB, lojistik regresyon, destek vektör makineleri ve K-en yakın komşu algoritmaları kullanılmıştır. Bu nedenle çalışma, belirlenen algoritmalarla sınırlıdır.

Çalışmanın anket ortamı üzerinden elde edilip derlenen verilere, veri madenciliği yöntemleri uygulanması ile üniversiteye geçiş başarısını tahmin edebilmesi öne çıkan özelliğidir. Gelen verilerden çıkarılan sonuçların eğitime katkı sağlayacağı öne çıkan diğer özelliklerdendir. Veri kümesi üzerinde yapılan incelemeler neticesinde, üniversiteye yerleşmeye hak kazanma başarısı üzerindeki etkenler şu şekilde açıklanabilir:

- Üniversiteye yerleşmeye hak kazanmayı en çok etkileyen faktör "liseyi yeterli bulma" durumudur. Önceki bölümde böyle bir sonuç elde edildiği söylenmişti. Oluşturulan karar ağaç modeli bilgi kazancını en yüksek bu öznitelikten almaktadır. Eğer mezun olunan lise öğrenci tarafından yeterli bulunuyorsa bu durum üniversite sınavındaki başarısını olumlu yönde etkilemektedir.
- Bir sonraki önemli faktör ise "diploma notu "'dur. Diploma notu 80'ni geçtikçe başarı oranı yükselmektedir. Diploma notu ve alan dersi notları birbirini etkileyen faktörlerdir
- Öğrencilik döneminde "üniversite ziyareti " yapılması sonucu olumlu etkileyen bir faktör olduğu çıkartılmıştır.
- Öğrencilik döneminde herhangi bir iş yerinde "çalışma" durumu sonucu olumsuz yönde etkileyen faktörlerdendir.
- Öğrencinin seçtiği alanı doğru bulması ve okumuş olduğu lise türü sonucu etkileyen diğer faktörlerden. Anadolu lisesi bitirmiş olmak sonucu olumlu şekilde etkilerken Meslek lisesinde okumuş olmak sonuçta olumsuz etki oluşturmaktadır.

- Ayrıca anne, baba eğitimi, öğrencinin okul rehberlik servisinden danışmanlık almış olması belirleyici faktörlerdendir.

Yukarıdaki bildirimler çalışmanın sonucuna daha yüksek etki eden faktörlerdir. Bu sonuçları elde edebilmek için veri madenciliği tekniklerinin yanında öznitelik seçim yöntemlerinin çalışmaya büyük bir katkısı olmuştur. Öznitelik seçim yöntemleri algoritmaların performanslarını yükseltmiştir. Öznitelik seçim yöntemleri uygulanmasaydı en yüksek performans 61, 78 olarak kalacaktı. Bu yöntemler ile performans 73.77'e yükseltilmiştir. Ayrıca seçim yöntemleri ile elde edilen özniteliklerin üniversiteyi tercih etme hakkında ne kadar etkili olduğu da ortaya çıkmıştır.

Bilinmeyen bir bilgiye veriler aracılığı ile ulaşılması ve sonucunda stratejiler belirlenebilmesi veri madenciliğinin en temel felsefesini oluşturmaktadır. Verilerin işlendiği kaynaklar göz önünde bulundurulduğunda veri madenciliğinin çok geniş bir uygulama alanı olduğu görülmektedir. Ama, eğitim alanında yapılan çalışmaların sınırlı sayıda olması bu alanda önemli bir açık oluşturmaktadır. Bu eksikliğin giderilmesine katkıda bulunacak ve yapılan çalışmanın devamı niteliğinde olabilecek araştırmalar aşağıda belirtilmektedir.

- Bu çalışmada elde edilen bulgular üniversite adayının üniversiteye yerleşmeye hak kazanma başarısına dayalı olarak oluşturulmuştur. Ders kriterleri baz alınarak başarıyı etkileyen faktörlerin algoritmalar kullanılarak karşılaştırmalar yapılabilir.
- Lise mezunu olma durumu ile sınırlandırılmış bu çalışma farklı eğitim aşamalarında da tekrarlanarak, ilgili bakanlık tarafından incelenip, önlemler alınması ve tedbirler oluşturabilecek bir sistem tasarlanabilir.
- Bu çalışma kapsamında kullanılmayan diğer yöntemlerin de kullanarak en uygun model araştırılabilir.
- Akademik başarıyı etkileyen faktörlerin ele alındığı literatür taramasından elde edilen özelliklerin daha spesifik bir şekilde ele alınıp sosyo-ekonomik, sosyo-kültürel ve bölgesel bazda başarı

tahminlenmesi yapılabilir. Bunun için örnek büyüklüğünün arttırılması önemlidir.

- Bu çalışmadan elde edile bilgiler ışığında bir web projesi oluşturularak sonuçların daha fazla kişiye ulaşmasının sağlanması ve farkındalık yaratılması önemlidir.

Bilginin büyük bir güç olduğu düşünülen günümüz dünyasında, asıl gücün bilgiyi kullanma becerisi olduğu veri madenciliğinin eğitim alanına uygulanarak eğitsel veri madenciliğini oluşturması, eğitimi klasik anlayışdan daha ileri bir seviyeye taşımıştır. Veriler üzerinde uygulanan yöntemler öğrencinin akademik başarısını tahminleyerek destekleyici adımlar sağlamıştır. Kullanılmış olan tüm yöntemlerle eğitimin daha başarılı, önlem alınabilir, izlenebilir hale getirilmesi sağlanarak katkıda bulunmak hedeflenmiştir. Yapılan değerlendirmeler sonucunda öğrenciye önerilebilecek davranışlar tespit edilmiştir. Veri madenciliğinin eğitim ve eğitimle dolaylı yoldan ilgili olan alt dallarıyla birlikte uygulanması toplumumuza ve insanlığa yararlı olacaktır.

KAYNAKÇA

Abdalla, S. ve Erdoğmuş, Ş. 2014. "Destek Vektör Makineleriyle Sınıflandırma Problemlerinin Çözümü için Çekirdek Fonksiyonu Seçimi." *Eskişehir Osmangazi Üniversitesi İİBF Dergisi,* 9(1), 175-198.

Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. 2019. Educational Data Mining and Learning Analytics for 21st century higher education: A Review and Synthesis. Telematics and Informatics, 37, 13-49.

Akben, S. B., Alkan, A. 2015. "Öznitelikler Arası Korelasyonun Düşük Olduğu Veri Kümelerinde Sınıflandırma Başarısını Artırmak İçin Yoğunluk Temelli Öznitelik Oluşturma". *Gazi Üniv. Müh. Mim. Fak. Der. Cilt 30, No 4*.

Berland , Matthew, Baker , Ryan S., Blikstein, Paulo, 2014, Educational Data Mining and Learning Analytics: Applications to Constructionist Research, 206

Bhatia, N. and Vandana., 2010. "Surveyof nearest neighbor techniques".*International Journal of Computer Science and Information Security*.8(2):302-305.

Botelho, A. F., Baker, R. S., & Heffernan, N. T. 2019. Machine-learned or expert-engineered features? Exploring feature engineering methods in detectors of student behavior and affect. In The twelfth international conference on educational data mining, Montréal, Canada.

Bravo-Agapito, J., Frances, C., & Seaone, I. 2019. Data mining in foreign language learning. WIREs: Data Mining and Knowledge Discovery, 10(1), e1287.

Can, Ertan. 2017. "Temel Eğitimden Ortaöğretime Geçiş Sınavı Kazanımlarının Veri Madenciliği Yöntemleri ile Değerlendirilmesi" Yüksek Lisans Tezi, Afyon Kocatepe Üniversitesi, Fen Bilimleri Enstitüsü, İnternet ve Bilişim Teknolojileri Yönetim Anabilim Dalı

Cortez, P., Silva, A.M.G. 2008, “Using Data Mining To Predict Secondary School Student Performance”, *Proceddings of 5th Annual Future Business Technology Conference*, Porto, 5-12.

Daniel, T. Larose. 2005. Discovering Knowledge in Data: An Introduction to Data Mining. New Jersey: Elsevier; John Wiley & Sons:43-53

Demir, C.E. 2009. “Factors Influensing The Academic Achievement of the Turkish Urban Poor”, *International Journ al of Educational Development* 29, 17-29.

Ersöz, Abdullah Ragıp. 2017. “Eğitsel Veri Madenciliği ile Öğrenci Profillerinin Belirlenmesi” Yüksek Lisans Tezi, Uludağ Üniversitesi, Eğitim Bilimleri Enstitüsü, Bilgisayar ve Öğretim Teknolojileri Eğitimi Ana Bilim Dalı

Gök, Murat. 2017 "Makine Öğrenmesi Yöntemleri ile Akademik Başarının Tahmin Edilmesi". *Gazi Üniversitesi Fen Bilimleri Dergisi Part C: Tasarım ve Teknoloji 5*, 139-148

Güner, N. ve Çomak, E. 2011. “Mühendislik Öğrencilerinin Matematik 1 Derslerindeki Başarısının Destek Vektör Makineleri Kullanılarak Tahmin Edilmesi”, *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi,* 17(2), 87-96.

Güneş, S, Görmüş, Ş, Yeşilyurt, F, Tuzcu, G. 2012. "ÖSYS Başarısını Etkileyen Faktörlerin Analizi". *Pamukkale Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*: 71-81.

Han, J., Kamber, M., & Pei, J. 2012. Data Mining Concepts and Techniques (3nd Edition b.). Amsterdam: Elsevier; Morgan Kauffman Publishers.

Kantardzic, M. 2011. Data Mining: Concepts, Models, Methods, and Algorithms. (3nd Edition b.) New Jersey: Elsevier; John Wiley & Sons:10-33

Kurt, Çağdaş, Erdem, O.Ayhan. 2012. "Öğrenci Başarılarını Etkileyen Faktörlerin Veri Madenciliği Yöntemleri ile İncelenmesi" *Journal of Polytecnic Dergisi* Cil15 Sayı :2 s. 111-116

Magdin, M., 2015. "Personalization of Student in Course Management Systems on the Basis Using Method of Data Mining", *The Turkish Online Journal of Educational Technology*, 14 1

Özdemir, Şebnem. 2016. "Eğitimde Veri Madenciliği ve Öğrenci Akademik Başarı Öngörüsüne İlişkin Bir Uygulama" Doktora Tezi, İstanbul Üniversitesi Enformatik Ana Bilim Dalı.

Özkan, Y. 2016, Veri Madenciliği Yöntemleri (3.Basım). İstanbul; Papatya Bilim, 217

Silahtaroğlu, Gökhan. 2013. Veri Madenciliği Kavram ve Algoritmalar (3. Basım), İstanbul; Papatya Bilim, 67.

Şeker, Ş. E. Eşmekaya, E. 2017. Eksik Verilerin Tamamlanması (Imputation), YBS Ansiklopedi, v. 4, is. 3, pp. 10 –17

Tufferry, S., 2011. Data Mining and Statistics for Desicion Making. New Jersey: John Wiley & Sons.

Yıldırım, İ. 2006. "Akademik Başarının Yordayıcısı Olarak Gündelik Sıkıntılar ve Sosyal Destek". *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi,* 30, 258-267

Yıldız, Muhammed Berke., Börekçi, Caner. 2020. "Predicting Academic Achievement with Machine Learning Algorithms." *Journal of Educational Technology & Online Learning*, 3(3), 372-392

Yıldız, Osman. 2014. "Öğrenmesi ile Uzaktan Eğitim Öğrencilerinin Performanslarının Değerlendirilmesi "Doktora Tezi, İstanbul Üniversitesi, Fen Bilimleri Enstitüsü, Enformatik Ana Bilim Dalı, İstanbul

İnternet Kaynağı

Alabaş, M. 2019. "Python ile Veri Görselleştirme: Matplotlib Kütüphanesi" https://medium.com/datarunner/matplotlibkutuphanesi-1-99087692102b E. T: 30.11.2020

Anonim. 2020. "MSE,RMSE, MAE, MAPE ve Diğer Metrikler" https://veribilimcisi.com/2017/07/14/mse-rmse-mae-mape-metrikleri-nedir/ E.T: 30.11.2010

Amanet, H. 2020. "CRISP-DM Nedir?" https://textdecipher.com/crisp-dm-nedir/ E.T: 24.12.2020

Balık, A. 2018. "Seaborn ile Veri Görselleştirilmesi"https://www.veribilimiokulu.com/blog/seaborn-ile-veri-gorsellestirmesi/ E.T:30.11.2020

Breiman, L., 2001 "Random Forests", Machine Learning, 5-32, https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf E.T. 12.11.2020

Bulut, O., Yavuz, H. Ç. 2019 "Educational data mining: A tutorial for the rattle package in R ", *International Journal of Assessment Tools in Education*, 20-36 https://dergipark.org.tr/tr/download/article-file/859872 E.T. 20.12.2020

Cerebro. 2018. "Python Neden Bu Kadar Popüler" https://medium.com/kodcular/python-neden-bu-kadar-popüler-d7f0f6819de5 E.T:30.11.2020

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R., 2000. CRISP-DM 1.0 Step-By-Step Data Mining Guide. SPSS. https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf E.T: 20.12.2020

Chao, Wei-Lun. Machine Learning Tutori. DISP Lab, Graduate Institute of Communication Engineering, National Taiwan University. http://disp.ee.ntu.edu.tw/~pujols/Machine%20Learning%20Tutorial.pdf E.T 20.12.2020

Durna, M. 2019. "Veri Bilimi için Temel Python Kütüphaneleri-1: Numpy"https://medium.com/bilişim-hareketi/veri-bilimi-için-temel-python-kütüphaneleri-1-numpy-750429a0d8e5 E.T: 30.11.2020

Gedleç, Ş. , Yılmaz, H. B. 2020 "Karar Ağaçlarında Algoritma Seçimi"https://www.datasciencearth.com/karar-agaclarinda-algoritma-secimi/ E.T: 28.11.20

Gray, D. E. 2014. Doing Research İn The Real World. (3nd Edition b.) London: Sage. https://www.academia.edu/29567720/Doing_Research_in_the_Real_Worl_D Davi_E_Gray E.T: 30.11.2020

Gülcan M. 2018. "Python Pandas Kütüphanesi"https://medium.com/@wmuratgulcan/python-pandas-kütüphanesi-597209068238 E.T: 30.11.2020

Gürdal, Hakan, Çakıcı, Yılmaz. 2017. "Eğitsel veri madenciliği" Edirne. https://www.researchgate.net/profile/Hakan_Gueldal/publication/321098785_Egitsel_Veri_Madenciligi/links/5a0d79f24585153829b1bfb5/Egitsel-Veri-Madenciligi.pdf .E.T: 20.12 2020

Karaderili, Ş. 2018. "Hata Matrisini Anlamak" https://medium.com/@sengul_krdrl/hata-matrisini-anlamak-7035b7921c0f E.T: 30.11.2020

Narkhede, S. 2018. "Understanding AUC-ROC "https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5. E.T: 24.12.2020

Márquez-Vera, Carlos, Cano, Alberto, Romero, Cristóbal, Ventura, Sebastián, 2012 "Predicting Student Failure At School Using Genetic Programming And Different Data Mining Approaches With High Dimensional And İmbalanced Data" https://www.researchgate.net/publication/257518289_Predicting_student_failure_at_school_using_genetic_programming_and_different_data_mining_approaches_with_high_dimensional_and_imbalanced_data E.T:18.12.2020

Millî Eğitim Bakanlığı. 2018. "PISA 2018 Türkiye Ön Raporu" http://pisa.meb.gov.tr/wp-content/uploads/2020/01/PISA_2018_Turkiye_On_Raporu.pdf E.T: 09.12.2020

Millî Eğitim Bakanlığı. 2020." Örgün Eğitim İstatistikleri" https://istatistik.yok.gov.tr/ , https://sgb.meb.gov.tr/www/icerik_goruntule.php?KNO=396 E.T: 12.11.2020

Navlani, A. 2019." KNN Classification Using Scikit- Learn" https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn E.T: 28.11.2020

Navlani, A. 2019." Naive Bayes Classification Using Scikit-Learn"https://www.datacamp.com/community/tutorials/naive-bayes-scikit-learn E.T:28.11.2020

Navlani, A. 2019. "Understanding Logistic Regression in Python" https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python E.T: 28.11.20

Navlani, A. 2018. "Understanding Random Forest Classifiers in Python" https://www.datacamp.com/community/tutorials/random-forests-classifier-python E.T: 28.11.20

Navlani, A. 2019.”Support Vector Machines with Scikit-Learn”https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python E.T: 28.11.2020

Scikit-Learn Developers “Decision Trees”https://scikit-learn.org/stable/modules/tree.html?highlight=id3 , E.T: 26.11.21

Şeker, Şadi Evren, 2018. “Knime ile Uçtan Uca Veri Bilimi”, https://sadievrenseker.com/wp-content/uploads/veribilimi_knime.pdf , 8-64 , E.T: 16.11.2020

Taş, B.2019. “ROC Eğrisi ve Eğri Altında Kalan Alan(AUC)”https://medium.com/@bernatas/roc-e%C4%9Frisi-ve-e%C4%9Fri-alt%C4%B1nda-kalan-alan-auc-97b058e8e0cf E.T: 30.11.2020

T.C. Kalkınma Bakanlığı. 2013. “Onuncu Kalkınma Planı (2019-2023).” Ankara. https://www.sbb.gov.tr/wp-content/uploads/2019/07/OnbirinciKalkinmaPlani.pdf E.T: 07.12.2020

William, Villegas-Ch, Diego Buenaño-Fernández and Sergio Luján-Mora. 2018, Educatıonal Data Analysıs Applyıng A Kdd Methodology , 16th International Conference e-Society , Spain, 316 https://www.researchgate.net/profile/Piet_Kommers/publication/331174981_Edited_by_Piet_Kommers_and_Pedro_Isaias_Associate_Editor_Luis_Rodrigues_ISBN_978-989-8533-75-3/links/5c6ab38492851c1c9de7734d/Edited-by-Piet-Kommers-and-Pedro-Isaias-Associate-Editor-Luis-Rodrigues-ISBN-978-989-8533-75-3.pdf#page=316 E.T: 21.12.2020

Yüceoğlu, B. 2017. “Scikit-Learn ile Veri Analitiğine Giriş” http://www.veridefteri.com/2017/11/23/scikit-learn-ile-veri-analitigine-giris/ E.T:30.11.2020

EKLER

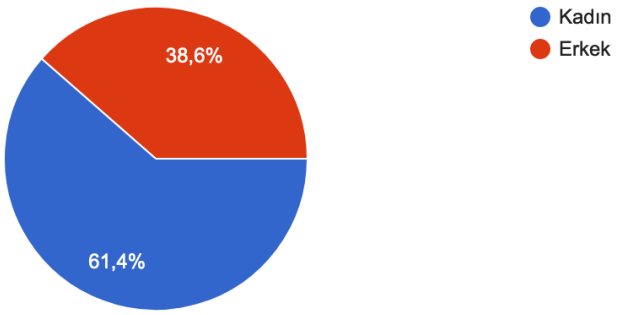
Ek 1. Etik Kurul Onayı

Ek 2. Anket Soruları

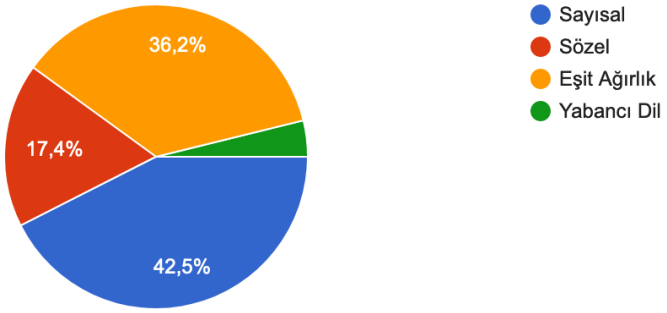
1. Cinsiyetiniz?
2. Üniversite sınavına hangi alandan girdiniz?
3. Lisede sorumlu olduğunuz alan derslerinizin karne notlarını beşlik sisteme göre işaretleyiniz.
4. Hangi tür liseden mezunsunuz?
5. Lisede alan dersleriniz için dershaneye gittiniz mi ya da özel ders desteği aldınız mı?
6. Üniversite sınavına hazırlanırken ders içeriklerine, soru bankalarına rahatlıkla ulaşabildiniz mi?
7. Günlük ortalama kaç saat ders çalıştınız?
8. Üniversite sınavına hazırlanırken dijital öğrenme platformlarından faydalandınız mı?
9. Lise öğreniminizde aldığınız eğitimi üniversite sınavı için yeterli buluyor musunuz?
10. Üniversite sınavına hazırlanma sürecinde bir iş yerinde çalışmak zorunda kaldınız mı?
11. Lisede akran zorbalığı yaşadınız mı?
12. Okuduğunuz lisenin rehberlik servisinden üniversite sınavı ile ilgi danışmanlık aldınız mı?
13. Lisede yaptığınız alan seçiminin doğru olduğunu düşünüyor musunuz?
14. Lisede sınıf mevcudunuz kaçtı?
15. Lise diploma notunuzu yazınız.
16. Üniversite sınavına giriş tarihiniz?
17. Lise Öğreniminizi hangi ilde gördünüz?
18. Lise öğrenimi sürecinde herhangi bir sosyal medya hesabınız var mıydı?
19. Üniversite sınavına hazırlık sürecinde herhangi bir sağlık problemi yaşadınız mı?
20. Lise Öğrenim Döneminde aile durumunu işaretleyiniz?
21. Lise döneminde ailenizin çalışma durumu işaretleyiniz?
22. Lise döneminde ailenizin gelir düzeyini işaretleyiniz?
23. Aile eğitim düzeyiniz?
24. Lise döneminde ailenizin medeni durumu nedir?

25. Kaç kardeşsiniz (Siz dahil)?
26. Lise öğreniminize devam ettiğiniz süreçte, üniversitede öğrenim gören kardeşiniz var mıydı?
27. Lise öğrenimi boyunca üniversite tanıtım günlerine ya da motivasyon amaçlı ziyaretlere katıldınız mı?
28. Üniversiteyi sınavsız geçiş hakkından faydalanarak ön lisans mı okudunuz?
29. Üniversiteye giriş sınavında barajı geçtiniz mi?
30. Üniversite sınavına ilk girişinizde tercih ettiğiniz üniversiteye yerleşmeye hak kazandınız mı?
31. Hangi üniversiteyi kazandınız?

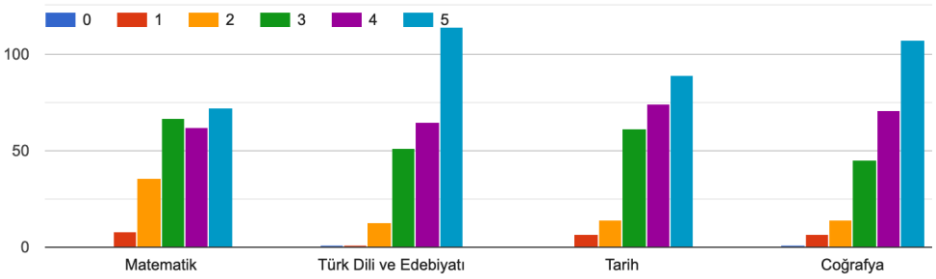
Ek 3. Veri Setinden Elde Edilen Bilgilerin Dağılımı



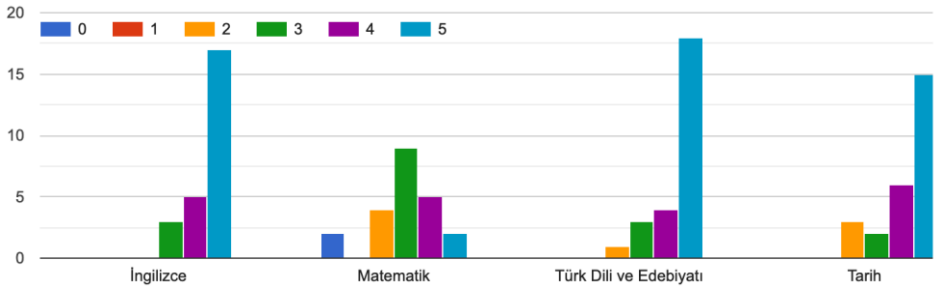
Cinsiyet Bilgisi



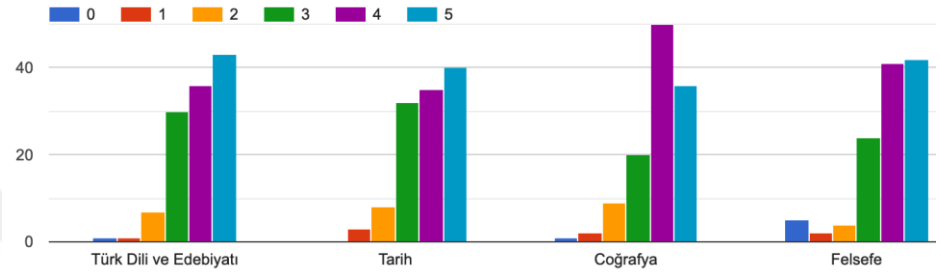
Alan Bilgisi



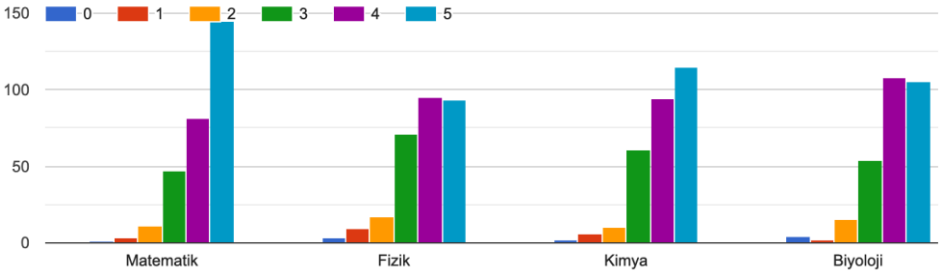
Eşit Ağırlık Not Dağılımı



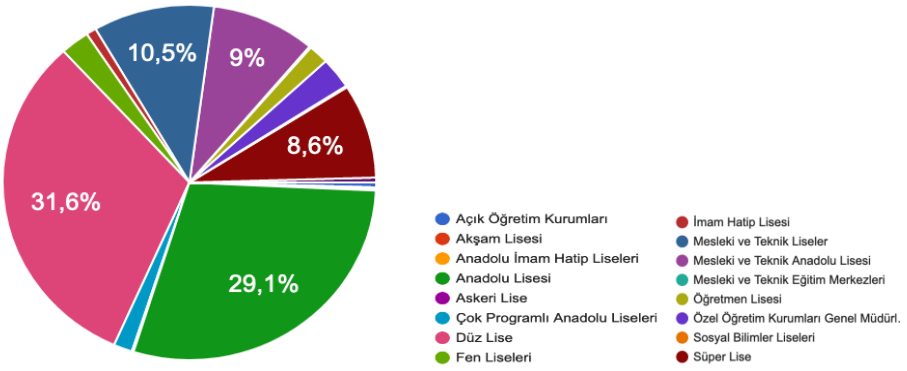
Yabancı Dil Bölümü Not Dağılımı



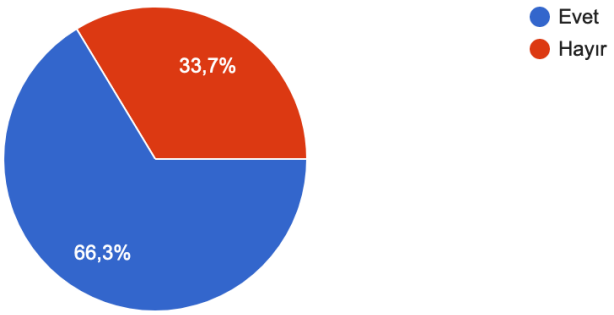
Sözel Not Dağılımı



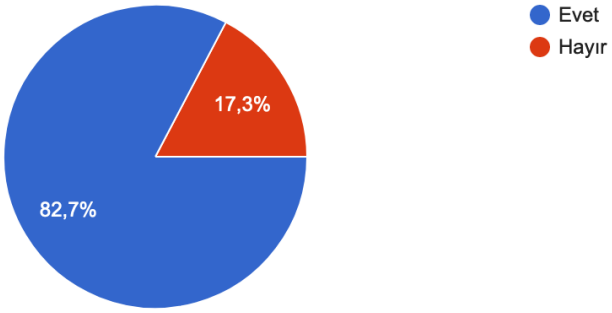
Sayısal Not Dağılımı



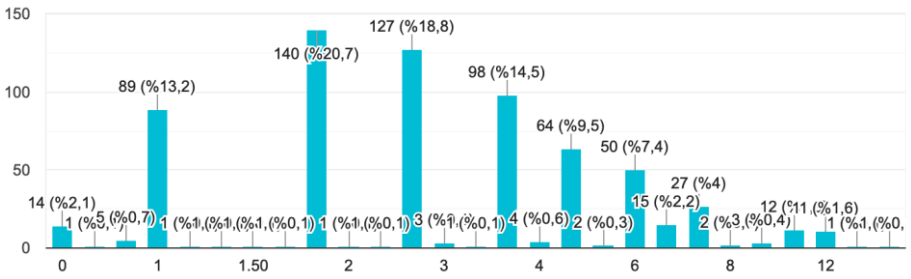
Lise Türleri



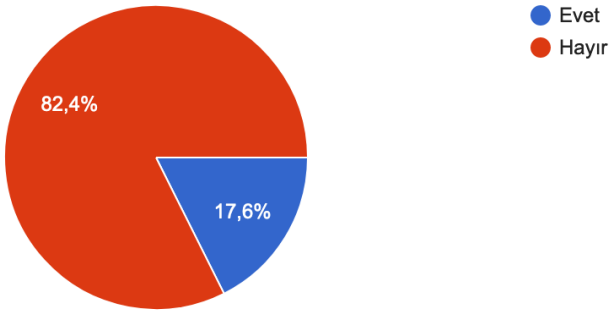
Özel Ders alma, Dershaneye Gitme Durumu



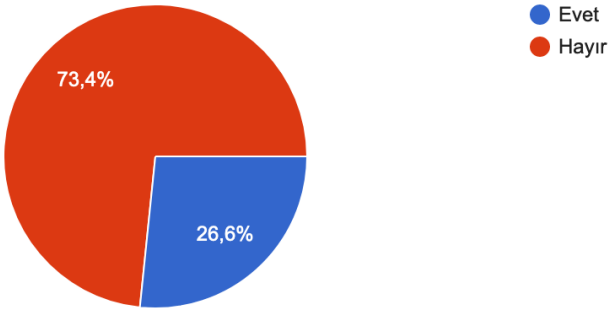
Ders İçeriklerine, Soru Bankalarına Kolay Erişim Durumu



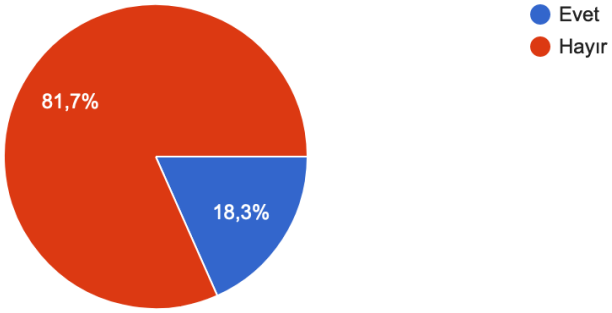
Günlük Ortalama Ders Çalışma Durumu



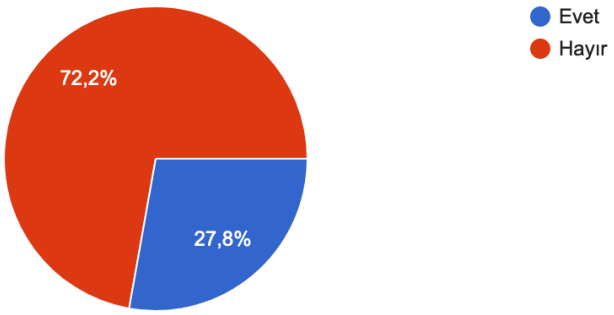
Dijital Öğrenme Platformlarından Faydalanma Durumu



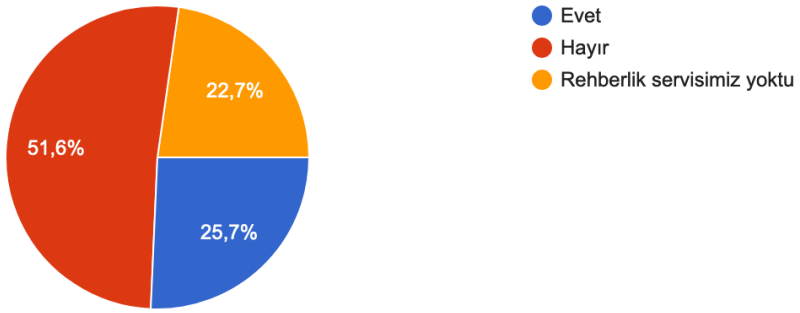
Lisede Alınan Eğitimi Yeterli Bulma Durumu



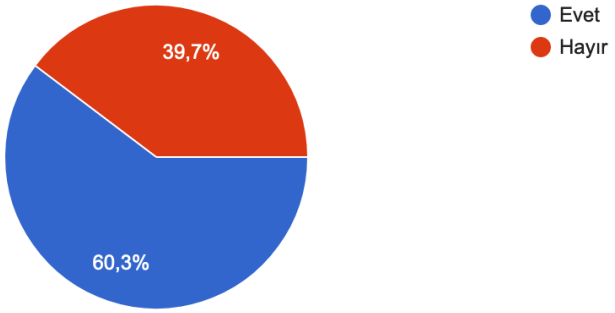
İşyerinde Çalışma Durumu



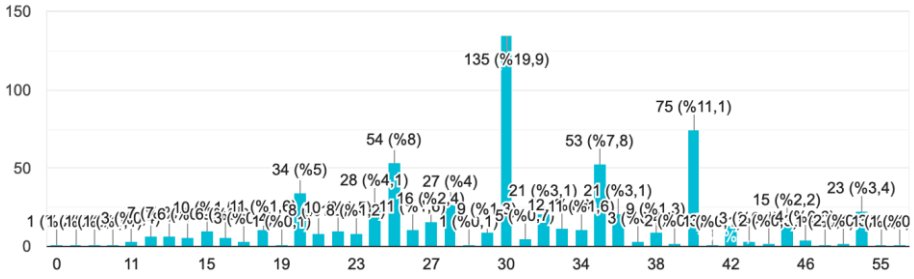
Akran Zorbalığı Yaşama Durumu



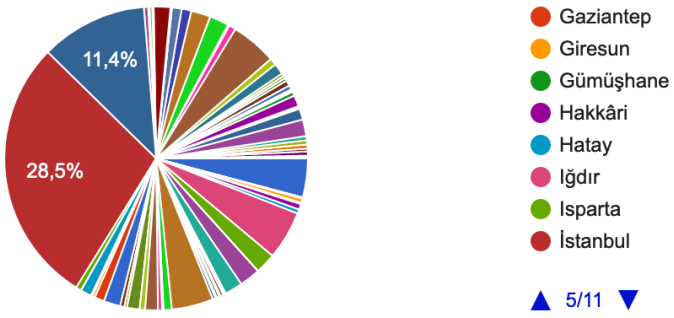
Rehberlik Servisi durumu



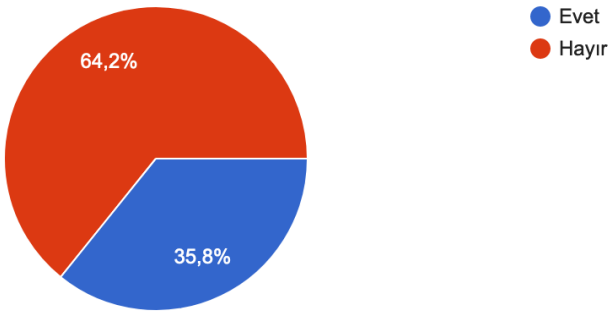
Lisede Alan Seçiminin Doğru Olma Durumu



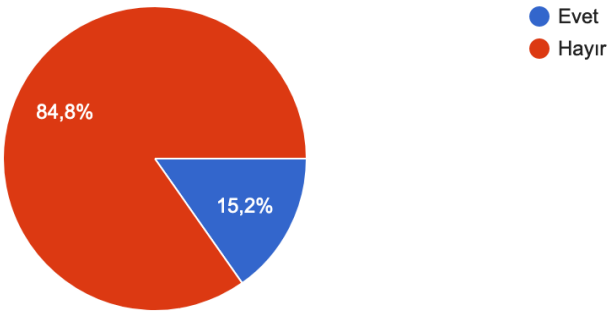
Sınıf Mevcudu Durumu



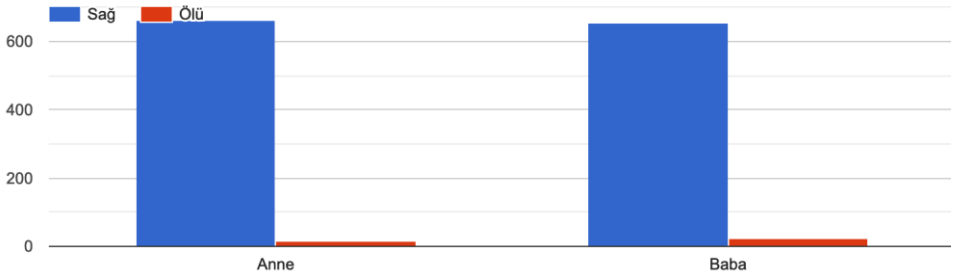
Lise Öğrenimi Görülen İl



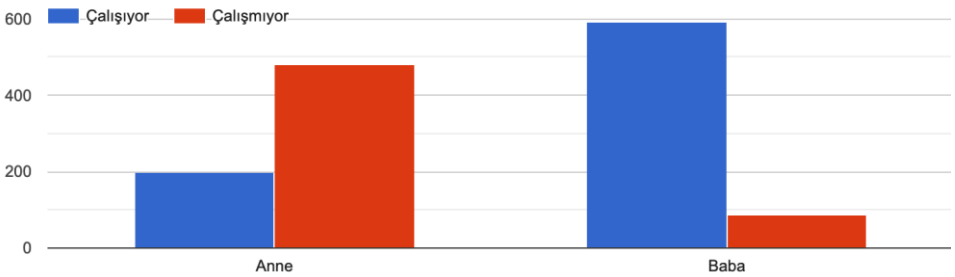
Sosyal Medya Kullanım Durumu



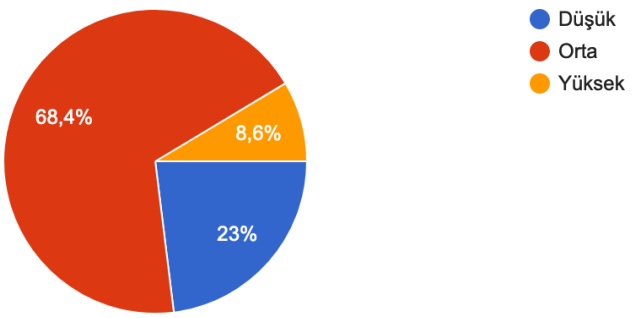
Sağlık Problemi Yaşama Durumu



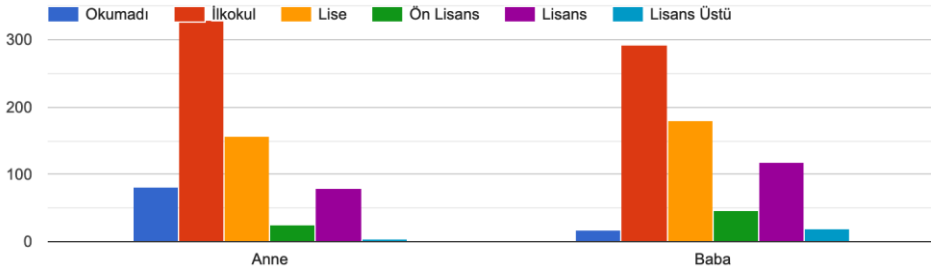
Ailenin Yaşama Durumu



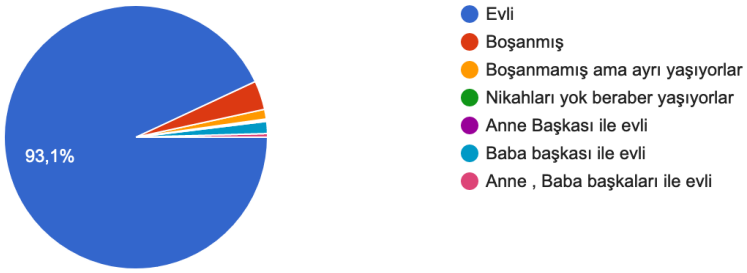
Ailenin Çalışma Durumu



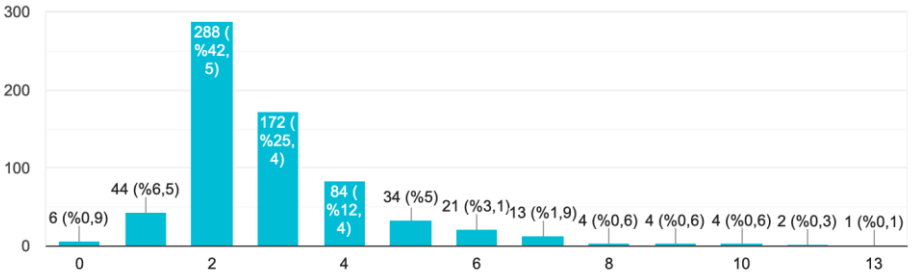
Ailenin Gelir Düzeyi



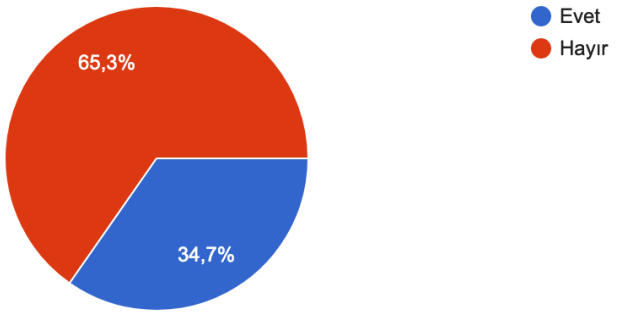
Ailenin Eğitim Düzeyi



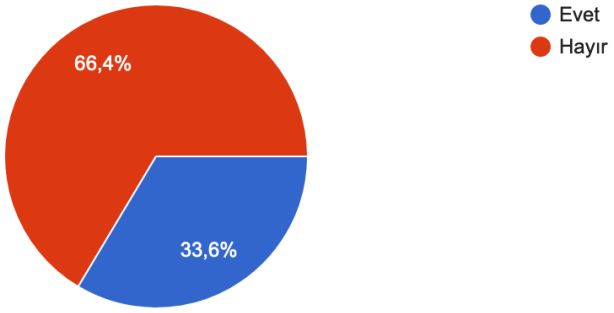
Ailenin Medeni Durumu



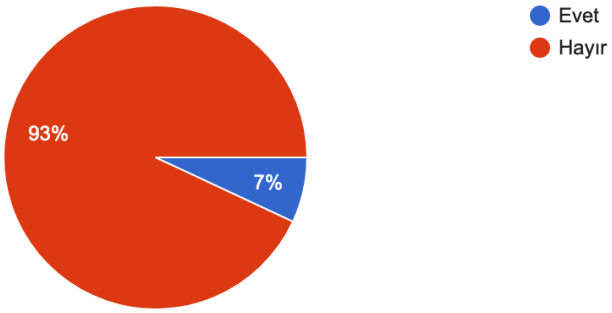
Kardeş Sayısı Durumu



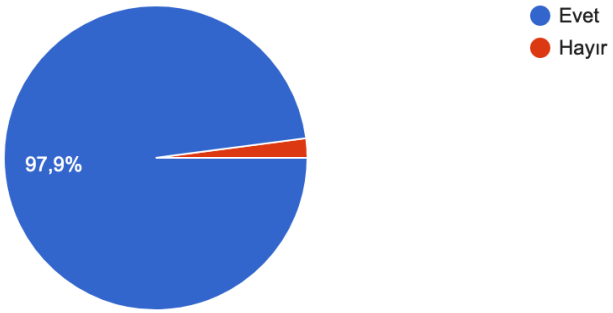
Üniversitede Kardeş olma Durumu



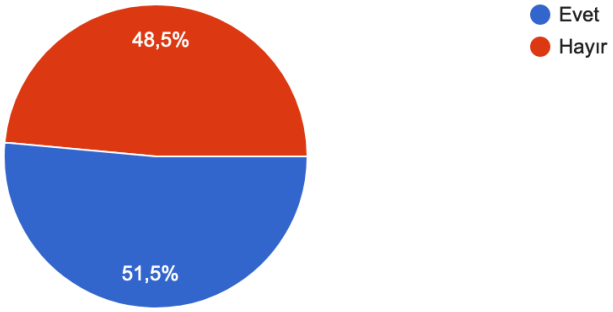
Lise Döneminde Üniversite Ziyareti



Sınavsız Ön Lisansa Geçiş Hakkı Olma Durumu

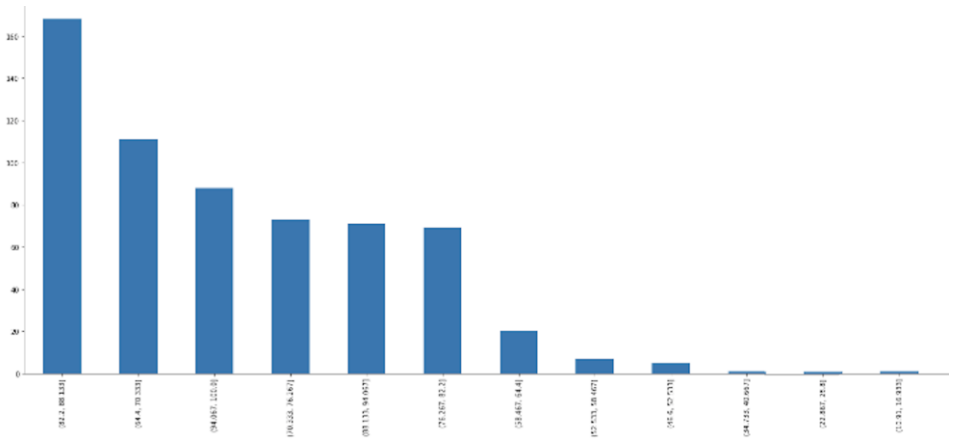


Sınavda Barajı Geçme Durumu

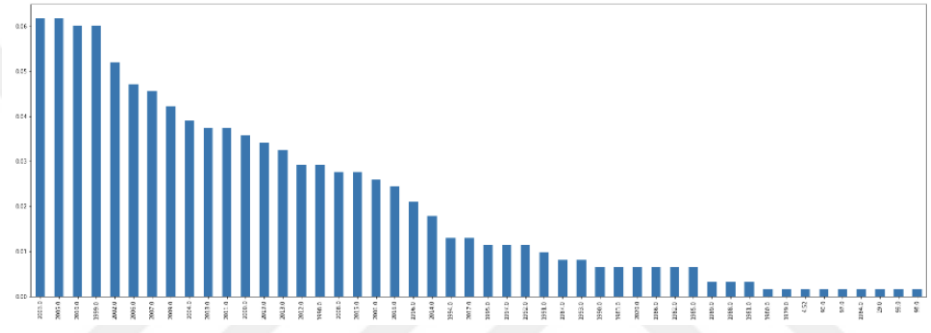


Tercih Edilen Üniversiteye Yerleşme Durumu

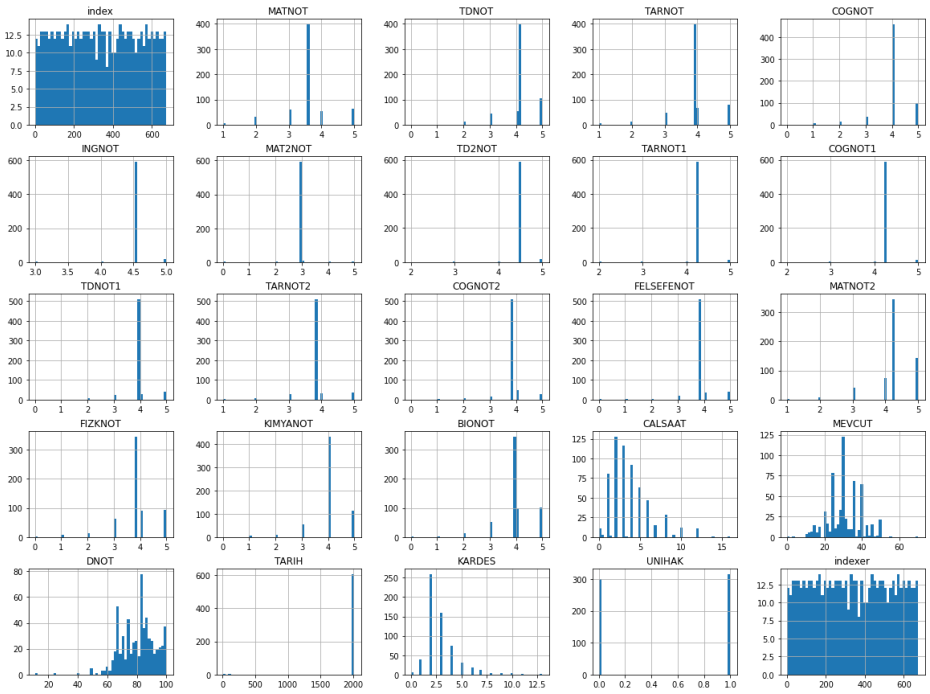
Kazanılan Üniversite Durumu



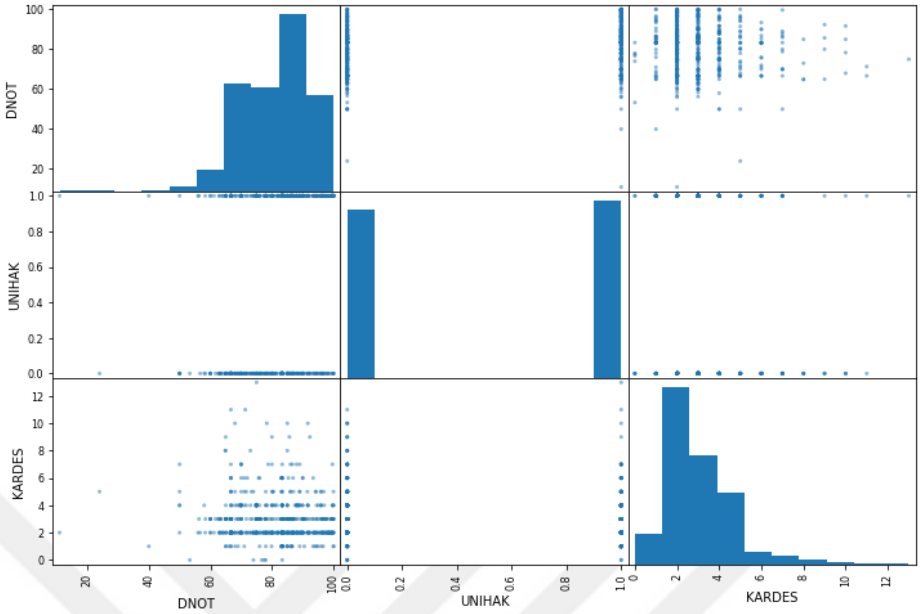
Diploma Notu Dağılımı



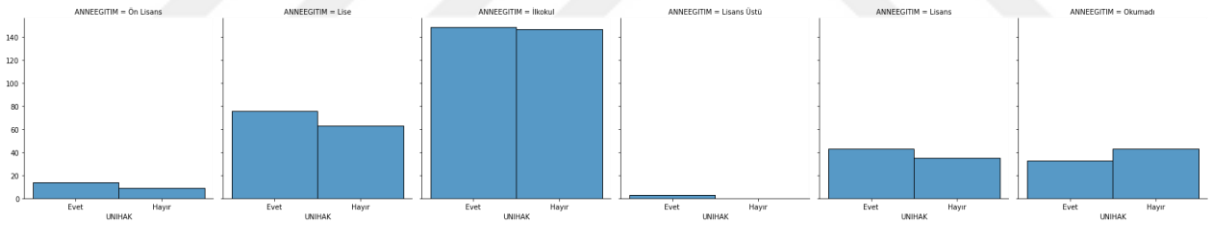
Yıllara Göre Sınava Giriş Dağılımı



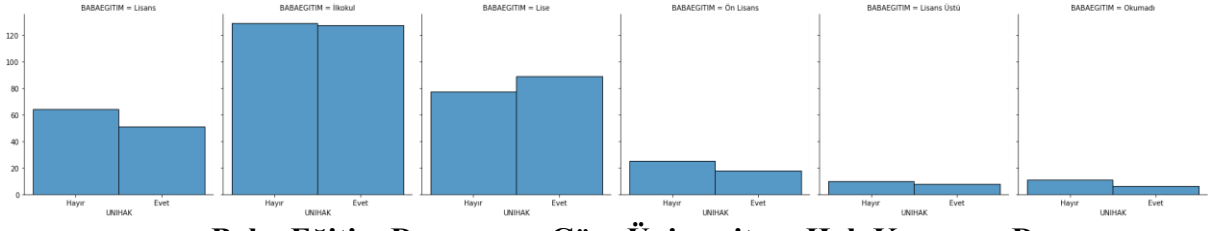
Veri Setinin Genel Dağılımı



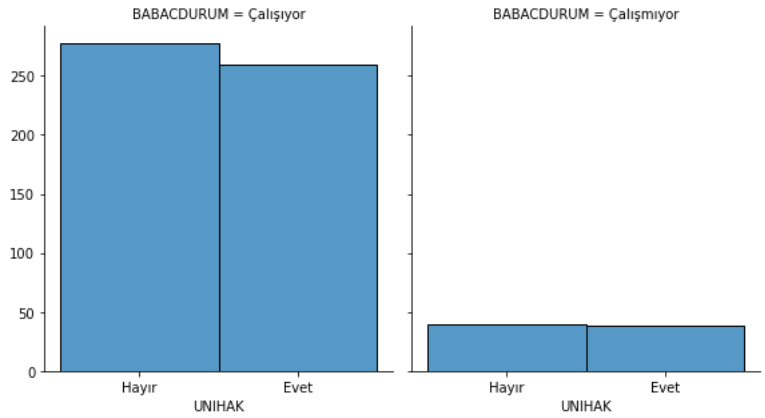
Veri Setinin Kardeş, Üniversite Kazanma Hakkı ve Diploma Notuna Göre Matrisi



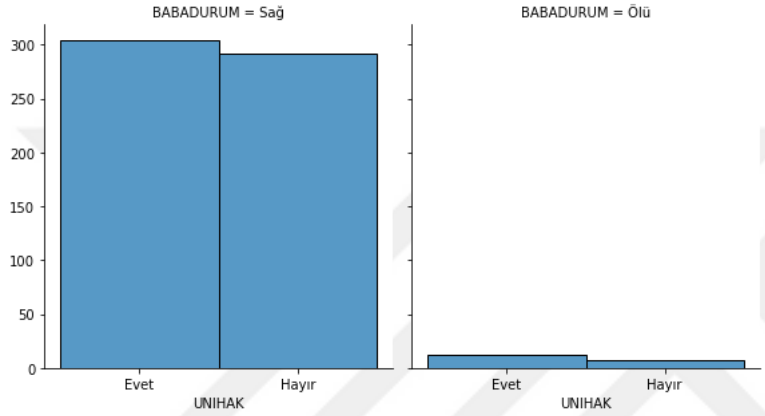
Anne Eğitim Durumuna Göre Üniversiteye Hak Kazanma Durumu



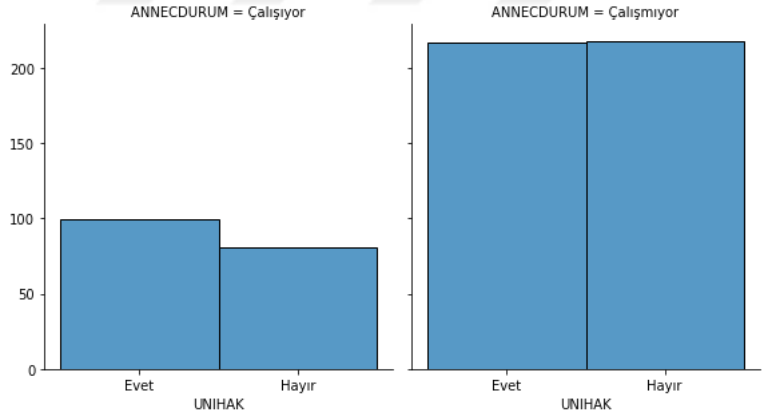
Baba Eğitim Durumuna Göre Üniversiteye Hak Kazanma Durumu



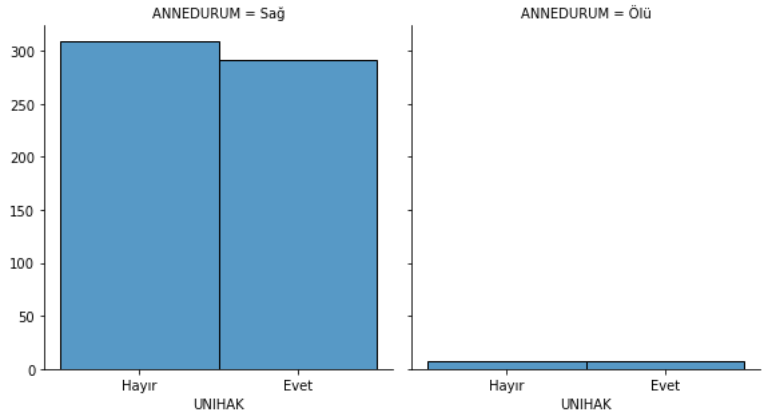
Babanın Çalışma Durumuna Göre Üniversiteyi Hak Kazanma Durumu



Babanın Yaşama Durumuna Göre Üniversiteyi Hak Kazanma Durumu



Annenin Çalışma Durumuna Göre Üniversiteyi Hak Kazanma Durumu



Annenin Yaşama Durumuna Göre Üniversiteyi Hak Kazanma Durumu

Cinsiyete Göre Tercih Edilen Üniversiteye Yerleşme Durumu

		Cinsiyetiniz		**Toplam**
Tercih Edilen Üniversiteye Yerleşme Durumu		Erkek	Kadın	
	Barajı Geçemeyen	31	30	61
	Evet	126	190	316
	Hayır	104	195	299
Toplam		261	415	676

Lise Türüne Göre Tercih Edilen Üniversiteye Yerleşmeye Hak Kazanma Durumu

	Tercih Edilen Üniversiteye Yerleşmeye Hak Kazanma Durumu			
		Evet	Hayır	Toplam
Açık Öğretim Kurumları	1	2	0	3
Akşam Lisesi	1	0	0	1
Anadolu İmam Hatip Liseleri	0	0	1	1
Anadolu Lisesi	4	127	66	197
Askeri Lise	0	0	1	1
Çok Programlı Anadolu Liseleri	3	3	5	11
Düz Lise	14	81	118	213
Fen Liseleri	0	10	7	17
İmam Hatip Lisesi	1	1	4	6
Mesleki ve Teknik Anadolu Liseleri	2	1	0	3
Mesleki ve Teknik Anadolu Lisesi	18	20	23	61
Mesleki ve Teknik Eğitim Merkezleri	1	0	0	1
Mesleki ve Teknik Liseler	15	20	36	71
Öğretmen Lisesi	0	9	3	12
Özel Öğretim Kurumları Genel Müdürlüğüne Bağlı Özel Okullar	0	12	7	19
Sosyal Bilimler Liseleri	0	1	0	1
Süper Lise	1	29	28	58
Toplam	61	316	299	676

Ek 4. Aşırı Öğrenmeye Bağlı Algoritma Sonuçları

Aşırı Uyum Ortalama Hata Değerleri

Algoritma	Ortalama Karesel Hata Değeri	Kök Ortalama Karesel Hata Değeri	Ortalama Mutlak Hata Değeri
Karar Ağacı	0.000000	0.000000	0.000000
Logistic Regression	0.038961	0.197386	0.038961
Rastgele Orman	0.064935	0.254824	0.064935
K-En Yakın Komşu	0.415584	0.644658	0.415584
Destek Vektör Makineleri	0.487013	0.697863	0.487013
Gaussian NB	0.058442	0.241747	0.058442

Aşırı Uyum Gösteren Modellerin Karışıklık Matrisi, Sınıflandırma Raporu

Karar Ağacı

```
[[79  0]
 [ 0 75]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        79
           1       1.00      1.00      1.00        75

    accuracy                           1.00       154
   macro avg       1.00      1.00      1.00       154
weighted avg       1.00      1.00      1.00       154
```

Lojistik Regresyon

```
[[73  6]
 [ 0 75]]
              precision    recall  f1-score   support

           0       1.00      0.92      0.96        79
           1       0.93      1.00      0.96        75

    accuracy                           0.96       154
   macro avg       0.96      0.96      0.96       154
weighted avg       0.96      0.96      0.96       154
```

Rastgele Orman

```
[[74  5]
 [ 5 70]]
              precision    recall  f1-score   support

           0       0.94      0.94      0.94        79
           1       0.93      0.93      0.93        75

    accuracy                           0.94       154
   macro avg       0.94      0.94      0.94       154
weighted avg       0.94      0.94      0.94       154
```

K-En Yakın Komşu

```
[[42 37]
 [27 48]]
              precision    recall  f1-score   support

           0       0.61      0.53      0.57        79
           1       0.56      0.64      0.60        75

    accuracy                           0.58       154
   macro avg       0.59      0.59      0.58       154
weighted avg       0.59      0.58      0.58       154
```

Destek Vektör Makineleri

```
[[79  0]
 [75  0]]
              precision    recall  f1-score   support

           0       0.51      1.00      0.68        79
           1       0.00      0.00      0.00        75

    accuracy                           0.51       154
   macro avg       0.26      0.50      0.34       154
weighted avg       0.26      0.51      0.35       154
```

Gaussian NB

```
[[79  0]
 [ 9 66]]
              precision    recall  f1-score   support

           0       0.90      1.00      0.95        79
           1       1.00      0.88      0.94        75

    accuracy                           0.94       154
   macro avg       0.95      0.94      0.94       154
weighted avg       0.95      0.94      0.94       154
```

Ek 5. Aşırı Uyum Sonrası Algoritma Sonuçları

Algoritma Ortalama Hata Değerleri

Algoritma	Ortalama Karesel Hata Değeri	Kök Ortalama Karesel Hata Değeri	Ortalama Mutlak Hata Değeri
Karar Ağacı	0.383117	0.618964	0.383117
Lojistik Regresyon	0.350649	0.592157	0.350649
Rastgele Orman	0.344156	0.586648	0.344156
K-En Yakın Komşu	0.415584	0.644658	0.415584
Destek Vektör Makineleri	0.487013	0.697863	0.487013
Gaussian NB	0.480519	0.693195	0.480519

Modellerin Karışıklık Matrisi, Sınıflandırma Raporu

Karar Ağacı

```
[[49 30]
 [29 46]]
              precision    recall  f1-score   support

           0       0.63      0.62      0.62        79
           1       0.61      0.61      0.61        75

    accuracy                           0.62       154
   macro avg       0.62      0.62      0.62       154
weighted avg       0.62      0.62      0.62       154
```

Lojistik Regresyon

```
[[53 26]
 [28 47]]
              precision    recall  f1-score   support

           0       0.65      0.67      0.66        79
           1       0.64      0.63      0.64        75

    accuracy                           0.65       154
   macro avg       0.65      0.65      0.65       154
weighted avg       0.65      0.65      0.65       154
```

Rastgele Orman

```
[[52 27]
 [26 49]]
              precision    recall  f1-score   support

           0       0.67      0.66      0.66        79
           1       0.64      0.65      0.65        75

    accuracy                           0.66       154
   macro avg       0.66      0.66      0.66       154
weighted avg       0.66      0.66      0.66       154
```

K-En Yakın Komşu

```
[[47 32]
 [32 43]]
              precision    recall  f1-score   support

           0       0.59      0.59      0.59        79
           1       0.57      0.57      0.57        75

    accuracy                           0.58       154
   macro avg       0.58      0.58      0.58       154
weighted avg       0.58      0.58      0.58       154
```

Destek Vektör Makineleri

```
[[79  0]
 [75  0]]
              precision    recall  f1-score   support

           0       0.51      1.00      0.68        79
           1       0.00      0.00      0.00        75

    accuracy                           0.51       154
   macro avg       0.26      0.50      0.34       154
weighted avg       0.26      0.51      0.35       154
```

Gaussian NB

```
[[70  9]
 [65 10]]
              precision    recall  f1-score   support

           0       0.52      0.89      0.65        79
           1       0.53      0.13      0.21        75

    accuracy                           0.52       154
   macro avg       0.52      0.51      0.43       154
weighted avg       0.52      0.52      0.44       154
```

Ek 6. Öznitelik Çıkarımı Yöntemlerine Ait Sonuçlar

Korelasyon Yöntemi Çıkartılan 30 Öznitelik

Öncelik Değeri	Öznitelikler
0	LISEIYETER
1	DNOT
2	LISETUR_Düz Lise
3	UNIZIYARET
4	TDNOT
5	DERSICERIK
6	DANISMANLIK_Hayır
7	SOSYALMEDYA
8	ALAN_Sözel
9	MATNOT
10	KARDES
11	MEVCUT
12	DERSHANEFLAG
13	FIZKNOT
14	MATNOT2
15	COGNOT
16	KIMYANOT
17	LISETUR_Mesleki ve Teknik Liseler
18	DOGRUYER
19	CALISMA
20	INGNOT
21	IL_Kocaeli
22	IL_Kırklareli
23	IL_Manisa
24	IL_Aksaray
25	AILEMEDENI_Boşanmış
26	TARIH
27	IL_İstanbul
28	COGNOT2
29	ANNEEGITIM_Lisans Üstü

Korelasyon Yöntemi Çıkartılan 30 Özniteliğin Yer Aldığı Hata

Algoritma	Ortalama Karesel Hata Değeri	Kök Ortalama Karesel Hata Değeri	Ortalama Mutlak Hata Değeri
Karar Ağacı	0.357143	0.597614	0.357143
Lojistik Regresyon	0.363636	0.603023	0.363636
Rastgele Orman	0.370130	0.608383	0.370130
K-En Yakın Komşu	0.480519	0.693195	0.480519
Destek Vektör Makineleri	0.487013	0.697863	0.487013
Gaussian NB	0.428571	0.654654	0.428571

Korelasyon Yöntemi Çıkartılan Modellerin Karışıklık Matrisi, Sınıflandırma Raporu

Karar Ağacı

```
[[43 32]
 [23 56]]
              precision    recall  f1-score   support

         0.0       0.65      0.57      0.61        75
         1.0       0.64      0.71      0.67        79

    accuracy                           0.64       154
   macro avg       0.64      0.64      0.64       154
weighted avg       0.64      0.64      0.64       154
```

Lojistik Regresyon

```
[[48 27]
 [29 50]]
              precision    recall  f1-score   support

         0.0       0.62      0.64      0.63        75
         1.0       0.65      0.63      0.64        79

    accuracy                           0.64       154
   macro avg       0.64      0.64      0.64       154
weighted avg       0.64      0.64      0.64       154
```

Rastgele Orman

```
[[45 30]
 [27 52]]
              precision    recall  f1-score   support

         0.0       0.62      0.60      0.61        75
         1.0       0.63      0.66      0.65        79

    accuracy                           0.63       154
   macro avg       0.63      0.63      0.63       154
weighted avg       0.63      0.63      0.63       154
```

K-En Yakın Komşu

```
[[38 37]
 [37 42]]
              precision    recall  f1-score   support

         0.0       0.51      0.51      0.51        75
         1.0       0.53      0.53      0.53        79

    accuracy                           0.52       154
   macro avg       0.52      0.52      0.52       154
weighted avg       0.52      0.52      0.52       154
```

Destek Vektör Makineleri

```
[[ 0 75]
 [ 0 79]]
              precision    recall  f1-score   support

         0.0       0.00      0.00      0.00        75
         1.0       0.51      1.00      0.68        79

    accuracy                           0.51       154
   macro avg       0.26      0.50      0.34       154
weighted avg       0.26      0.51      0.35       154
```

Gaussian NB

```
[[73  2]
 [64 15]]
              precision    recall  f1-score   support

         0.0       0.53      0.97      0.69        75
         1.0       0.88      0.19      0.31        79

    accuracy                           0.57       154
   macro avg       0.71      0.58      0.50       154
weighted avg       0.71      0.57      0.50       154
```

EK 7. Anova Yöntemi ile Çıkartılan 80 Öznitelik için Algoritmaların Hata Değerleri ve Sınıflandırma Raporları

Anova Yöntemi Çıkartılan 80 Öznitelik

Öznitelikler	
AILEMEDENI_Anne Başkası ile evli	IL_Karabük
AILEMEDENI_Boşanmış	IL_Kastamonu
AILEMEDENI_Nikahları yok beraber yaşıyorlar	IL_Kırklareli
ALAN_Sözel	IL_Kocaeli
ANNEEGITIM_İlkokul	IL_Kütahya
ANNEEGITIM_Lisans Üstü'	IL_Manisa
ANNEEGITIM_Lise	IL_Mardin
ANNEEGITIM_Okumadı	IL_Muş
BABAEGITIM_Okumadı	IL_Niğde
BIONOT	IL_Ordu
CALISMA	IL_Rize
CINSIYET_Kadın	IL_Sakarya
COGNOT	IL_Sivas
COGNOT1	IL_Tokat
COGNOT2	IL_Yalova
DANISMANLIK_Hayır	IL_Yozgat
DANISMANLIK_Hayır	INGNOT
DANISMANLIK_Rehberlik servisimiz yoktu	KARDES
DERSHANEFLAG	KARDES
DERSICERIK	KIMYANOT
DNOT	L_Trabzon
DOGRUYER	LISEIYETER
DOP	LISETUR_Askeri Lise
FIZKNOT	LISETUR_Düz Lise
IL_Adıyaman	LISETUR_İmam Hatip Lisesi
IL_Ağrı'	LISETUR_Mesleki ve Teknik Anadolu Lisesi
IL_Aksaray	LISETUR_Mesleki ve Teknik Liseler
IL_Amasya	LISETUR_Öğretmen Lisesi
IL_Bartın	LISETUR_Özel Öğretim Kurumları Genel Müdürlüğüne Bağlı Özel Okullar
IL_Bilecik	LISETUR_Sosyal Bilimler Liseleri
IL_Bitlis'	MATNOT
IL_Bolu	MATNOT2
IL_Çanakkale	MEVCUT
IL_Edirne	SOSYALMEDYA
IL_Elazığ	TARNOT1

IL_Eskişehir	TARNOT2
IL_Gaziantep	TD2NOT
IL_Hakkâri	TDNOT
IL_İstanbul	TDNOT1
IL_Kahramanmaraş	UNIZIYARET

Anova Yöntemi Çıkartılan 80 Özniteliğin Yer Aldığı Hata

Algoritma	Ortalama Karesel Hata Değeri	Kök Ortalama Karesel Hata Değeri	Ortalama Mutlak Hata Değeri
Karar Ağacı	0.454545	0.674200	0.454545
Lojistik Regresyon	0.435065	0.659595	0.435065
Rastgele Orman	0.402597	0.634506	0.402597
K-En Yakın Komşu	0.461039	0.678998	0.461039
Destek Vektör Makineleri	0.435065	0.659595	0.435065
Gaussian NB	0.415584	0.644658	0.415584

Anova Yöntemi Çıkartılan Modellerin Karışıklık Matrisi, Sınıflandırma Raporu

Karar Ağacı

```
[[40 31]
 [35 48]]
              precision    recall  f1-score   support

         0.0       0.53      0.56      0.55        71
         1.0       0.61      0.58      0.59        83

    accuracy                           0.57       154
   macro avg       0.57      0.57      0.57       154
weighted avg       0.57      0.57      0.57       154
```

Lojistik Regresyon

```
[[47 24]
 [42 41]]
              precision    recall  f1-score   support

         0.0       0.53      0.66      0.59        71
         1.0       0.63      0.49      0.55        83

    accuracy                           0.57       154
   macro avg       0.58      0.58      0.57       154
weighted avg       0.58      0.57      0.57       154
```

Rastgele Orman

```
[[49 22]
 [45 38]]
              precision    recall  f1-score   support

         0.0       0.52      0.69      0.59        71
         1.0       0.63      0.46      0.53        83

    accuracy                           0.56       154
   macro avg       0.58      0.57      0.56       154
weighted avg       0.58      0.56      0.56       154
```

K-En Yakın Komşu

```
[[34 37]
 [33 50]]
              precision    recall  f1-score   support

         0.0       0.51      0.48      0.49        71
         1.0       0.57      0.60      0.59        83

    accuracy                           0.55       154
   macro avg       0.54      0.54      0.54       154
weighted avg       0.54      0.55      0.54       154
```

Destek Vektör Makineleri

```
[[44 27]
 [37 46]]
              precision    recall  f1-score   support

         0.0       0.54      0.62      0.58        71
         1.0       0.63      0.55      0.59        83

    accuracy                           0.58       154
   macro avg       0.59      0.59      0.58       154
weighted avg       0.59      0.58      0.58       154
```

Gaussian NB

```
[[15 56]
 [16 67]]
              precision    recall  f1-score   support

         0.0       0.48      0.21      0.29        71
         1.0       0.54      0.81      0.65        83

    accuracy                           0.53       154
   macro avg       0.51      0.51      0.47       154
weighted avg       0.52      0.53      0.49       154
```

EK 8. Ki-Kare Yöntemi ile Çıkartılan 40 Öznitelik için Algoritmaların Hata Değerleri ve Sınıflandırma Raporları

Ki-Kare Yöntemi Çıkartılan 40 Öznitelik

Öznitelikler	
ALAN_Sözel	IL_Kocaeli
ANNEEGITIM_Lisans Üstü	IL_Manisa
ANNEEGITIM_Lise	IL_Mardin
ANNEEGITIM_Okumadı	IL_Sakarya
BABAEGITIM_Okumadı	IL_Tokat
DANISMANLIK_Hayır	IL_Trabzon
DERSHANEFLAG	IL_Yalova
DERSICERIK	KARDES
DNOT	KIMYANOT
DOGRUYER	LISEIYETER
FIZKNOT	LISETUR_Düz Lise
IL_Bilecik	LISETUR_İmam Hatip Lisesi
IL_Edirne	LISETUR_Mesleki ve Teknik Anadolu Lisesi
IL_Elazığ	LISETUR_Mesleki ve Teknik Liseler
IL_Eskişehir	LISETUR_Öğretmen Lisesi
IL_Gaziantep	LISETUR_Özel Öğretim Kurumları Genel Müdürlüğüne Bağlı Özel Okullar
IL_İstanbul	MEVCUT
IL_Kahramanmaraş	SOSYALMEDYA
IL_Karaman	TARNOT2
IL_Kırklareli	UNIZIYARET

Ki-Kare Yöntemi Çıkartılan 40 Özniteliğin Yer Aldığı Hata

Algoritma	Ortalama Karesel Hata Değeri	Kök Ortalama Karesel Hata Değeri	Ortalama Mutlak Hata Değeri
Karar Ağacı	0.422078	0.649675	0.422078
Lojistik Regresyon	0.428571	0.654654	0.428571
Rastgele Orman	0.363636	0.603023	0.363636
K-En Yakın Komşu	0.448052	0.669367	0.448052
Destek Vektör Makineleri	0.415584	0.644658	0.415584
Gaussian NB	0.532468	0.729704	0.532468

Ki-Kare Yöntemi Çıkartılan Modellerin Karışıklık Matrisi, Sınıflandırma Raporu

Karar Ağacı

```
[[39 32]
 [33 50]]
              precision    recall  f1-score   support

         0.0       0.54      0.55      0.55        71
         1.0       0.61      0.60      0.61        83

    accuracy                           0.58       154
   macro avg       0.58      0.58      0.58       154
weighted avg       0.58      0.58      0.58       154
```

Lojistik Regresyon

```
[[47 24]
 [42 41]]
              precision    recall  f1-score   support

         0.0       0.53      0.66      0.59        71
         1.0       0.63      0.49      0.55        83

    accuracy                           0.57       154
   macro avg       0.58      0.58      0.57       154
weighted avg       0.58      0.57      0.57       154
```

Rastgele Orman

```
[[54 17]
 [39 44]]
              precision    recall  f1-score   support

         0.0       0.58      0.76      0.66        71
         1.0       0.72      0.53      0.61        83

    accuracy                           0.64       154
   macro avg       0.65      0.65      0.63       154
weighted avg       0.66      0.64      0.63       154
```

K-En Yakın Komşu

```
[[34 37]
 [32 51]]
              precision    recall  f1-score   support

         0.0       0.52      0.48      0.50        71
         1.0       0.58      0.61      0.60        83

    accuracy                           0.55       154
   macro avg       0.55      0.55      0.55       154
weighted avg       0.55      0.55      0.55       154
```

Destek Vektör Makineleri

```
[[44 27]
 [37 46]]
              precision    recall  f1-score   support

         0.0       0.54      0.62      0.58        71
         1.0       0.63      0.55      0.59        83

    accuracy                           0.58       154
   macro avg       0.59      0.59      0.58       154
weighted avg       0.59      0.58      0.58       154
```

Gaussian NB

```
[[15 56]
 [16 67]]
              precision    recall  f1-score   support

         0.0       0.48      0.21      0.29        71
         1.0       0.54      0.81      0.65        83

    accuracy                           0.53       154
   macro avg       0.51      0.51      0.47       154
weighted avg       0.52      0.53      0.49       154
```

EK 9. Korelasyon, Anova, Ki-Kare Yöntemleri ile Çıkartılan Öznitelikler için Yapılan Testlerin Sonuçları

Yapılan Testlerdeki Sonuçların Bir Kısmı

Algoritma	Yüzde	K	ROC AUC Ort	ROC AUC STD	Doğruluk Ort	Doğruluk STD	Öznitelik Sayısı	Yöntem
Gaussian NB	30_70	2	73.77	7.08	60.94	6.80	40	Ki-Kare
Rastgele Orman	40_60	6	73.15	6.99	66.15	6.43	80	Ki-Kare
Rastgele Orman	25_75	10	72.96	6.95	63.52	6.50	80	Anova
Lojistik Regresyon	40_60	4	72.16	6.90	65.07	5.99	30	Ki-Kare
Gaussian NB	25_75	10	71.76	6.89	55.72	5.88	40	Anova
Rastgele Orman	25_75	10	71.09	6.88	66.16	6.67	30	Korelasyon
Gaussian NB	25_75	10	70.36	6.87	53.58	6.63	40	Korelasyon
Lojistik Regresyon	25_75	10	69.35	6.28	63.11	5.89	20	Anova
Lojistik Regresyon	25_75	10	67.90	6.24	62.27	5.47	40	Korelasyon
Destek Vektör Makineleri	25_75	10	65.91	6.15	61.81	5.50	80	Korelasyon
Destek Vektör Makineleri	20_80	6	64.54	6.14	60.72	5.43	60	Ki-Kare
Destek Vektör Makineleri	25_75	2	64.30	6.55	60.95	0.08	80	Anova
Karar Ağacı	25_75	10	62.68	6.46	62.69	6.27	20	Korelasyon
Karar Ağacı	25_75	2	62.05	5.60	62.05	5.55	80	Ki-Kare
Karar Ağacı	25_75	10	61.95	6.70	62.03	7.40	60	Anova
K-En Yakın Komşu	40_60	4	60.94	8.40	58.32	8.23	70	Ki-Kare
K-En Yakın Komşu	25_75	10	59.55	8.85	58.38	9.33	30	Korelasyon
K-En Yakın Komşu	25_75	2	59.42	0.36	59.43	0.35	30	Anova