

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

**MAKİNE ÖĞRENMESİ YÖNTEMİ İLE EĞİTİM
BAŞARISININ TAHMİNİ MODELİ**

Yüksek Lisans Tezi

Tezi Hazırlayan
Deniz ZİLYAS

İstanbul, 2023

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

**MAKİNE ÖĞRENMESİ YÖNTEMİ İLE EĞİTİM
BAŞARISININ TAHMİNİ MODELİ**

Yüksek Lisans Tezi

Tezi Hazırlayan

Deniz ZİLYAS

2020003034

0009-0004-9972-6158

Danışman

Dr. Öğr. Üyesi Atınç YILMAZ

İstanbul, 2023

YEMİN METNİ

Yüksek Lisans Tezi olarak sunduğum “Makine Öğrenmesi Yöntemi İle Eğitim Başarısının Tahmini Modeli” başlıklı çalışmanın, bilimsel ahlak ve geleneklere uygun şekilde tarafımdan yazıldığını, yararlandığım eserlerin tamamının kaynaklarda gösterildiğini, çalışmamın içinde kullanıldıkları her yerde bunlara atıf yapıldığını belirtir ve bunu onurumla doğrularım. 02/03/2023

Deniz ZİLYAS



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

02/03.2023

Enstitümüz *Bilgisayar Mühendisliği* Anabilim Dalı *Bilgisayar Mühendisliği* Programı yüksek lisans öğrencilerinden 2020003034 numaralı *Deniz ZİLYAS*'ın "*Beykent Üniversitesi Lisansüstü Eğitim – Öğretim Yönetmeliği*"nin ilgili maddesine göre hazırlayarak, Enstitümüze teslim ettiği "*Makine Öğrenmesi Yöntemi İle Eğitim Başarısının Tahmini Modeli*" konulu tezini, Yönetim Kurulumuzun 10/01/2023 tarih ve 2023/02 sayılı toplantısında seçilen ve Taksim yerleşkesinde toplanan biz jüri üyeleri huzurunda, Beykent Üniversitesi Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 29. maddesinin 3. fıkrası gereğince 45 dakika süre ile 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 At*** YIL***
(Beykent Üniversitesi)

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

ÜYE
Dr. Öğr. Üyesi So*** SER***
(Kütahya Dumlupınar Üniversitesi)

Adı ve Soyadı : Deniz ZİLYAS
Danışman : Dr. Öğr. Üyesi Atınç YILMAZ
Türü ve Tarihi : Yüksek Lisans Tezi, 2023
Alanı : Bilgisayar Mühendisliği
Anahtar Kelimeler : Makine Öğrenmesi, Tahmin, Performans, Regresyon,
Eğitim Başarısı

ÖZ

MAKİNE ÖĞRENMESİ YÖNTEMİ İLE EĞİTİM BAŞARISININ TAHMİNİ MODELİ

Günümüzde makine öğrenmesi yöntemleri etkin bir biçimde kullanılarak pek çok alanda yüksek performanslar göstermiştir. Çeşitli sektörlerde son yıllarda daha da yaygınlaşmaya başlamıştır. Makine öğrenmesini modellerinden elde edilebilecek başarılarla birçok sorun öngörülüp çözüme ulaştırılabilir.

Bu çalışmadaki amaçta, ortaokul öğrencileriyle yapılan anketten toplanan verilerle eğitim başarısının tahminini yapacak bir makine öğrenmesi modeli ortaya koymak ve öğrenciyi etkileyebilecek faktörlerinin önüne geçebilmektir. Anket soruları, öğrencinin başarısına tesir edebilecek etkenler araştırılarak oluşturulmuştur.

Çalışma kapsamında, çeşitli ortaokullarda eğitim gören 520 farklı öğrenciden kişisel verilerin korunması kanunu kapsamında 13 sorudan oluşan anket aracılığıyla veri toplanmıştır. Bu veriler hiçbir kurumla paylaşılmamış olup, gizliliği korunmuştur. Veri seti incelenerek bazı manipülasyon, ön işleme, görselleştirme işlemlerinden geçirilmiş , K-Nearest Neighbor (K-NN) , Random Forest (RF) , Lineer Regresyon, Bagged Trees, Gradient Boosting Regressor (GBM)ve Karar Ağaçları (DT) algoritmaları kullanılmıştır. Aralarından model başarısı en yüksek olan Random Forest (RF) algoritması seçilmiştir.

Çalışmada, veri manipülasyon işlemleri gerçekleştirildikten sonra model kurularak öğrencinin Türkçe notu üzerinden eğitim başarısının tahmini yapılmıştır. Türkçe dersi bağımlı değişken olarak seçilmiştir. Çalışmadaki ders

seiminin belirlenmesi, ana dilin Trke olması ve eēitim hayatından itibaren her dnem Trke dersi ile karřılařılmasından dolaydır.

alıřma neticesinde, algoritmanın performansı 0.88 ve R-squared deēeri 0.98 olarak belirlenmiřtir. Yeni girdilerle test edilen model Trke notu retmiřtir. ērencinin eēitim durumunu etkileyen en nemli faktrler Trke notuna baēlı olarak aralarındaki korelasyon ile aile geliri ve ders alıřma saati olarak belirlenmiřtir. Etkenler seilirken model eřitli senaryolarla defalarca test edilmiřtir ve korelasyon iliřkisi hesaplanmıřtır.



Name and Surname : Deniz ZİLYAS

Supervisor : Dr. Lecturer Atınc YILMAZ

Degree and Date : Master Dissertation, 2023

Major : Computer Engineering

Key Words : Machine Learning, Prediction, Performance, Regression,
Educational Success

ABSTRACT

A PREDICTIONAL MODEL OF EDUCATIONAL SUCCESS WITH THE MACHINE LEARNING METHOD

Today, machine learning methods have been used effectively and have shown high performance in many areas. It has become more widespread in various sectors in recent years. Many problems can be predicted and solved with the successes that can be obtained from machine learning models.

The aim of this study is to present a machine learning model that will predict educational success with the data collected from the survey conducted with secondary school students and to present the factors that may affect the student. The questions of the questionnaire were created by investigating the factors that may affect the success of the student.

Within the scope of the study, data were collected from 520 different students studying in various secondary schools through a questionnaire consisting of 13 questions within the scope of the law on the protection of personal data. This data is not shared with any institution and its confidentiality is protected. K-Nearest Neighbor(K-NN) , Random Forest (RF), Linear Regression, Bagged Trees, Gradient Boosting Regressor(GBM), Decision Trees (DT) algorithms were used. Among them, the Random Forest algorithm with the highest model success was selected.

In the study, after the data manipulation processes were carried out, the model was established and the estimation of the educational success of the student was made based on the Turkish grade. Turkish course was chosen as the dependent variable. The determination of the course selection in the study is due to the fact that the mother tongue is Turkish and that Turkish course has been encountered every semester throughout the education life of a student.

İÇİNDEKİLER

ÖZ

ABSTRACT

TABLolar LİSTESİ	iii
ŞEKİLLER LİSTESİ	iv
KISALTMALAR	v
SÖZLÜK	vi
GİRİŞ.....	1

BİRİNCİ BÖLÜM

1. TÜRKİYE’DE EĞİTİM BAŞARISINI ETKİLEYEN FAKTÖRLER.. 14

İKİNCİ BÖLÜM

2. MAKİNE ÖĞRENMESİ	16
2.1 Temel Kavramlar	16
2.2 Yapay Zeka	16
2.3 Derin Öğrenme.....	17
2.4 Makine Öğrenmesi	18
2.4.1 Öğrenme Türleri	19
2.4.2 Makine Öğrenmesi Algoritmaları.....	20
2.4.2.1 Basit Doğrusal Regresyon (SLR)	20
2.4.2.2 Çoklu Doğrusal Regresyon(MLR)	21
2.4.2.3 Ridge Regresyon.....	22
2.4.2.4 Lasso Regresyon.....	23
2.4.2.5 K Nearest Neighbors (K-NN).....	24
2.4.2.6 Karar Ağaçları (DT veya Decision Tree)	26
2.4.2.7 Random Forest (RF)	27
2.4.2.8 Naive Bayes (NB).....	28
2.4.2.9 Destek Vektör Makineleri (SVM)	29
2.4.2.10 Gradient Boosting Regressor (GBM)	29
2.5 Makine Öğrenmesinin Avantajları	29

2.6 Makine Öğrenmesinin Kullanım Alanları	30
2.7 Makine Öğrenmesinde Model Seçimi	30
2.8 Model Başarısı Değerlendirme Yöntemleri.....	31
2.9 Varyans ve Yanlılık Kavramları.....	33

ÜÇÜNCÜ BÖLÜM

3. MAKİNE ÖĞRENMESİ YÖNTEMİ İLE EĞİTİM BAŞARISININ	
TAHMİNİ MODELİ	35
3.1. Veri Setinin Hazırlanması	35
3.1.1 Öznitelik Bilgileri Ve Grafiksel Yorumları.....	35
3.1.2 Özniteliklerin Analizi	40
3.2. Modelin Oluşturulması.....	42
3.3 Çalışmada Kullanılan Algoritmaların Analizi	44
3.3.1 Random Forest Algoritması	44
3.3.2 K-NN Algoritması	45
3.3.3 Decision Tree Algoritması	45
3.3.4 Bagged Trees Algoritması.....	45
3.3.5 Çoklu Lineer Regresyon Algoritması.....	46
3.3.6 Gradient Boosting Machines Algoritması	46
3.4 Algoritmaların Değerlendirilmesi	47
3.5 Makine Öğrenmesi Yöntemlerinin Optimizasyonu Ve Model Seçimi	48
3.6 Random Forest İle Modelin Test Edilmesi	49
3.7 Feature Selection İle Öznitelik Seçimi.....	50
3.8 Web Arayüzünün Tasarlanması	52
SONUÇ	57
KAYNAKÇA.....	60
EKLER	64
Ek-1. Anket	64

TABLÖLAR LİSTESİ

Tablo 1. Veri Setinin Detayları	36
Tablo 2. Anket Sorularına Karşılık Gelen Değişken İsimleri	36
Tablo 3. Veri Seti İçerisindeki Değişkenlerin Tiplerinin Gösterimi	37
Tablo 4. Modellerin Başarı Yüzdeleri	47
Tablo 5. Modelin Performans Metrikleri	47
Tablo 6. Modellerin R-kare Değerleri	48
Tablo 7. Modelin Test Edildiği Örnek Veri	49
Tablo 8. Modellere Verilen Yeni Girdiler	51
Tablo 9. Modellerin Tahmin Ettiği Türkçe Notu Çıktıları	52

ŞEKİLLER LİSTESİ

Şekil 1. Cinsiyete Göre Türkçe Dersi Doğru Cevap Sayısı Ortalaması	14
Şekil 2. Lisansüstü Tezlerin Dağılımı	15
Şekil 3. Makine Öğrenmesi Ve Yapay Zeka	16
Şekil 4. Derin Öğrenme Katmanları	18
Şekil 5. Makine Öğrenmesi Yöntemleri	18
Şekil 6. Sınıflandırma Ve Regresyon Yöntemi	19
Şekil 7. Gözetimli Ve Gözetimsiz Öğrenme	20
Şekil 8. Basit Lineer Regresyonun Grafikselsel Gösterimi	21
Şekil 9. K-NN Adımlarının Gösterimi	24
Şekil 10. Euclidean Uzaklık Hesaplaması temsili görünümü; (Noktasal).....	25
Şekil 11. Manhattan Uzaklık Hesaplaması temsili görünümü; (Noktasal)	25
Şekil 12. Minkowski Uzaklık Hesaplaması temsili görünümü; (Noktasal)	26
Şekil 13. Karar Ağaçları Şematik Gösterimi	27
Şekil 14. Random Forest yapısının gösterimi.....	28
Şekil 15. Karışıklık matrisinin tablosal gösterimi	32
Şekil 16. Varyans Ve Yanlılığın Grafikselsel Gösterimi (Koptur, 2017).....	34
Şekil 17. Değişkenlerin value_counts().plot.barh() Parametresi İle Gösterimi..	38
Şekil 18. Catplot Parametresi İle Kategorik Değişkenlerin Gösterilmesi	39
Şekil 19. Kategorik Değişkenlerin Catplot Parametresiyle Görselleştirmesi	40
Şekil 20. Aykırı Gözleme Sahip Değişkenin Boxplot Grafiği	41
Şekil 21. Aykırı Gözleme Sahip Olmayan Değişkenin Boxlot Grafiği.....	41
Şekil 22. Kategorik Değişkenlerin Label Encoder İle Sayısallaştırılması.....	42
Şekil 23. Makine Öğrenmesi Yöntemlerinin Uygulanması Aşamaları	43
Şekil 24. Makine Öğrenmesinin Temel Süreci.....	44
Şekil 25. RF Algoritması İle Test Edilen Verinin Çıktısı.....	50
Şekil 26. Değişkenlerin Pvalue Değerleri.....	51
Şekil 27. Web Uygulamasında İzlenen Adımlar	53
Şekil 28. Value Değerlerinin Fonksiyonlara Uyarlanması	54
Şekil 29. Cmd Komutları.....	55
Şekil 30. Uygulamanın Arayüzü	56

KISALTMALAR

AI : Artificial Intelligence

ANN : Artificial Neural Network

BTR: Bagged Tree Regression

CSS : Cascading Style Sheets

DT : Decision Tree

FamSep: Family Separated

FN : False Negative

FP : False Positive

GBM : Gradient Boosting Regressor

HTML : Hypertext Markup Language

K-NN : K – Nearest Neighbor

ML: Machine Learning

MLR : Multiple Linear Regression

NB : Naïve Bayes

NLP: Neuro Linguistic Programming

RF : Random Forest

SLR : Simple Linear Regression

SMO : Sequential Minimal Optimization

StepMF: StepMotherFather

TN : True Negative

TP : True Positive

WStudying: WeekendStudying

YSA : Yapay Sinir Ağları

SÖZLÜK

Derin Öğrenme: Yapay sinir ağlarının büyük miktarda veriden öğrendiği Makine Öğrenmesi'nin bir alt dalıdır.

Makine Öğrenmesi: Sisteme verilen girdilere göre öğrenen ve veriler üzerinden tahmin yapabilen algoritmaların çalışmasını araştıran bir sistemdir.

Regresyon: Bir veya birden fazla tahmin değişkeninindeğerine dayalı olarak sürekli bir sonuç değişkenini tahmin edilmesine izin veren bir makine öğrenmesi yöntemidir.

Varyans: Verilerin aritmetik ortalamadan sapmalarının karelerinin toplamıdır.

Yanlılık: Verinin ne kadar yanlış olduğunu ölçer.

Yapay Zeka: İnsan zekasını taklit eden sistemler veya makineler anlamına gelen bir terimdir.

GİRİŞ

Türkiye’de eğitim sistemi 12 yıllık zorunlu eğitim olarak 3 kademeye ayrılmıştır. Şu anda 4+4+4 eğitim sistemini kullanılmaktadır. Son yıllarda eğitim sistemi birçok kez değiştirilmiştir. Eğitim sistemindeki değişiklikler bazı konularda öğrencileri olumlu ve/veya olumsuz yönde etkilemiştir. Bunlardan en önemlisi okul öncesi eğitimin gözardı edilerek, öğrencilerin küçük yaşta okula başlamalarıdır. Okuma-yazma, aritmetik konularda yetersiz olan bir çocuğun gelişiminde zorluklar yaşanmaktadır Buna karşın erken yaşlarda fark edilen yeteneklerin geliştirilmesinde geç kalınmaması gibi bazı avantajlardan da söz edilebilir.

Lise eğitimine geçmeden önce son basamak 8.sınıftır. Öğrenci bu aşamada hayatını etkileyecek bir döneme girmekte olup, lise eğitimi sayesinde üniversiteye girişe hazır hale gelecektir. Öğrenci üzerinde önceki senelere göre daha fazla baskı olabilmektedir. Sosyal çevre, aile ilişkisi, öğretmeniyle etkileşimi, özel ders desteği, ailelerin geliri, oyun oynama ve televizyon izleme süresi, uyku düzeni ve bunun gibi birçok etken öğrencinin başarı durumunu doğrudan etkileyebilmektedir. Çelenk’in sonuçlarına göre, aile üyelerinden eğitim yardımı alan ve okulla yakın iş birliği içerisine giren ailelerden gelen çocukların, okuduğunu anlama başarılarının daha yüksek olduğu sonucuna varılmıştır.

Türkiye’de yaklaşık olarak 13 bin özel ve 45 bin devlet ilkokul ile ortaokulu bulunmaktadır. Her yıl okul ücretlerinin artmasıyla beraber orta sınıf ve altındaki aileler çocuklarını özel okula gönderememektedir. Genel olarak özel okullarda eğitimin daha iyi verildiği düşüncesi mevcuttur. Bazı aileler özel ders ve dersane desteği sağlayabilirken, geliri düşük bir aileye sahip öğrenci kendi çabalarıyla öğrenmeye mecbur kalmaktadır. Hiçbir özel destek almadan başarılı olabilen öğrenciler de bulunmakta, ancak sayısı yüzdelik olarak bakıldığında oldukça düşüktür.

Ailenin öğrenciye karşı ilgisizliği, öğretmen tarafından verilen ödevlerde öğrencinin yetersiz kaldığı noktalarda gerekli yardımlaşmanın sağlanmaması da başarıyı etkilemektedir. Öğretmenler her ne kadar velilerin

daha özverili olmalarını isteseler de gerek ailenin eğitim konusundaki yetersizliği gerek iş yoğunluğu kaynaklı sebeplerden öğrenci bu konuda eksik kalmaktadır.Çelenk; eğitim açısından destekleyici bir tutum ve okul ile ortak program üzerinde görüş birliği içinde düzenli iletişim halinde olan ailelerden gelen çocukların daha başarılı olduğunu belirtmiştir Çalışma ortamı, maddi yetersizlik, ailenin kalabalık olması öğrencinin başarısızlığını ortaya çıkaran etkenlerden olabilmektedir. Maddi yetersizlik kısmen etkilese de gerekli kaynaklar aile tarafından öğrenciye sunulamamaktadır. Maddi durumdan yetersiz olan ailelerden de başarılı öğrencilerin varlığı nadiren de olsa gözlemlenmiştir..

Öğrencilikte, Türkçe; dersler arasında en çok öneme sahip ve öğrenilmesi güç olan derslerden biri olmuştur. Bunun en belirgin sebeplerinden biri öğrencilerin kitap okuma alışkanlığının olmamasıdır. Türkçe dilinin sürekli kullanılması iletişim açısından da bireyi güçlendirmektedir. İletişimi yüksek olan bir öğrenci soru sormaktan da çekinmeyecek ve kavrayamadığı herhangi bir dersle ilgili sorusunu öğretmeniyile paylaşabilecektir. Çekingen bireylerin yetişmesi genel olarak hayata karşı başarısızlıkları doğurabilmektedir.

Bu nedenle tezde amaç, Türkçe dersi notu üzerinden öğrencinin başarısını etkileyebilecek problemler arasında doğrudan etki edenlerin tespit edilerek ortaya çıkarılmasıdır. Çalışma sonunda program, sisteme verilen yeni girdilerle öğrencinin Türkçe notunu tahmin edebilecektir.

Çalışmayı kapsayan 13 farklı anket sorusu bulunmaktadır (soru sayısı artırılabilirdi, ancak belirli sayıdan sonra anketörlerin cevaplarının tutarlılığı soru işareti doğurabilirdi). Anket soruları bir öğrencinin başarı hayatını etkileyen etmenler araştırılarak oluşturulmuştur. Genel olarak ankette; aile durumu ve geliri, özel ders durumu, ders çalışma, oyun oynama, uyku düzeni, kardeş olma durumu, kendine ait oda durumu gibi etkenlere yer verilmiştir. Öğrencinin Türkçe notu konusunda doğru söylediğini anlamak adına ankete 3 gizli soru daha eklenmiş ancak uygulamada bu sorular değişken olarak kullanılmamıştır.

Anket sonuçlarından 519 gözlem elde edilmiştir. Öğrencilerden toplanan veriler Excel tablolarına aktarılmıştır. Veri seti toplamda 13 sütun ve 520 satırdan oluşmaktadır. Uygulama ise, Python programlama dilinde yazılmış olup, editör olarak Jupyter Notebook kullanılmıştır. Veriler .csv uzantısıyla Jupyter Notebook'a aktarılmıştır.

Veri seti içerisinde performansı etkileyecek durumlar göz önünde bulundurulmuştur. Eksik bir değerın olması modelin başarısını etkileyen en önemli etkenlerden biridir, ve bu sebeple eksik değer çok etkin bir problemdir. Böyle bir durumla karşılaşıldığında yapılması gereken 2 yöntem bulunmaktadır. Veri seti içerisinde eksik bir veri olup olmadığı **isnull** parametresi çağrılarak kontrol edilmektedir. Gözlemler içerisinde boş bir veriye rastlanılırsa o değere 0 verilebilir veya bulunduğu sütunun ortalaması atanabilir. Ancak bu çalışmada, **isnull** parametresi çağrılarak yapılan eksik veri analizinde herhangi bir eksik gözleme rastlanmamıştır.

Veri setinde toplamda 8 kategorik ve 5 sayısal değişken bulunmaktadır. Kategorik değişkenlerin özetlerini görselleştirmek için matplotlib kütüphanesindeki plot parametresinden yararlanılarak grafikler oluşturulmuş ve böylelikle değişkenlerin analizi kolaylaşmıştır. Her bir sayısal değişken için ortalama, maksimum değer, minimum değer, medyan ve standart sapma betimlemeleri yapılarak son adımda değişkenler arasındaki kolerasyon değerleri hesaplanmıştır.

Veri seti üzerinde kontrol edilmesi gereken diğer bir manipülasyon işlemi aykırı gözlem analizidir (outlier detection). Bu tez çalışmasında kutu grafiğı (boxplot) yöntemi ile aykırı gözlem analizi yapılmıştır. Boxplot yöntemi; veri çeyreklerini (veya yüzdelikleri) ve ortalamaları görüntüleyerek sayısal verilerin ve değişkenliğıın görsel olarak dağılımını göstermek için kullanılmıştır. Bu yöntem, veri analizinde sıklıkla kullanılan bir grafik türüdür. Kutu grafiğı yöntemi ile görselleştirilen değişkenler arasında Ders Çalışma

Süresi (studying) ve Gelir (income) değişkenlerinde aykırı gözleme rastlanmıştır. Baskılama yöntemi ile alt sınır değerleri belirlenmiş ve alt sınır değerinden aşağıda kalan değerler baskılanmıştır. Bununla birlikte modelin başarısında 0.01 değerinde artış gözlemlenmiştir.

Çalışmada veriyi modellemek için makine öğrenmesi yöntemlerinden Çoklu Lineer Regresyon, Random Forest , K-NN ve Bagged Tree Regression , Karar Ağaçları ve GBM yöntemleri kullanılmıştır. Modelin Türkçe notu tahmin etmesi beklendiği için 13 değişken içerisinde Türkçe notu bağımlı değişken olarak seçilmiştir. Algoritmalar, %80 oranında eğitim verisiyle eğitilmiştir. Eğitim sonrasında modelin tahmin başarısını değerlendirmek için model doğrulama (tuning) işlemi yapılmıştır. Bu bağlamda en yüksek performansı gösteren Random Forest algoritması seçilmiştir. Modelin başarı (score) değeri 1.88 8 ve 0.98 R^2 değeri ile tatminkar bir sonuç elde edilmiştir.

Çalışmanın görsel açıdan zenginleştirilmesi için bir web arayüzüne aktarılması planlanmıştır. Bu nedenle HTML-CSS kodlarıyla uygulamaya ait web arayüzü oluşturulmuştur. Jupyter Notebook ile geliştirilen uygulamanın farklı bir dizinde çalışması için Python'a ait Pickle modülü kullanılmıştır. Web arayüzüyle senkronizasyonun sağlanması Python'ın Flask kütüphanesiyle gerçekleştirilmiştir.

Çalışmada aşağıda belirtilen sıralamayla her bölümün içeriği kısa özetle tanıtılmıştır ;

Bölüm 1'de Türkiyede ki eğitim başarısıyla ilgili bilgi verilmiştir.

Bölüm 2'de Makine öğrenmesi konusuna girmeden önce temel kavramlar açıklanmış, makine öğrenmesi algoritmaları, tarihçesi ve kullanım alanlarından bahsedilmiştir.

Bölüm 3'de Makine Öğrenmesi İle Eğitim Başarısının Tahmini Modeli'nde kullanılan algoritmalar detaylandırılmış, seçilen yöntem esas alınarak oluşturulan uygulama ve uygulamanın çalışacağı arayüz yazılımı tanıtılmıştır.

Bu çalışmaya yönelik elde edilen sonuç ve öneriler sonuç bölümünde verilmiştir.

Literatür Taraması

Bu tez çalışmasında eğitim başarısıyla ilgili yapılan çalışmalar literatürde araştırılmış ancak makine öğrenmesi baz alınarak yapılmış fazla çalışmaya rastlanmamıştır. Bu sebeple öğrenci başarısızlığı ile ilgili çeşitli methodlar kullanılarak yapılan çalışmalara da yer verilmiştir.

(Onur Seveli, 2019), Bu araştırma makalesi çalışmasında göğüs kanseri teşhisinde 5 farklı makine öğrenmesi yöntemi kullanılarak algoritmaların performans karşılaştırmasını yapılmıştır. Çalışmasında Destek Vektör Makinesi (SVM), Naive Bayes (NB), Rastgele Forest (RF), K Nearest Neighbor(KNN) ve Lojistik Regresyon(LR) algoritmalarının test başarılarını karşılaştırmıştır. Test işlemi sonucunda en başarılı modelin %98,24 doğruluk oranıyla Lojistik Regresyon olduğu ortaya çıkartmıştır.

(Görmez, 2021), Adli bilişimde makine öğrenmesini kullanarak terör olaylarının tahmin edilmesi üzerine bir tez çalışması yapmış olup, çalışmasında Destek Vektör Makineleri(SVM) , Naive Bayes(NB) , K-Nearest Neighbor(K-NN) , Random Forest(RF), Karar Ağacı(DT – Decision Tree) ve Lojistik Regresyon(LR) algoritmalarını kullanılmıştır. Karar ağacı algoritması ile %91,16 doğruluk oranı elde etmiştir ve doğruluk oranının, boyut azaltma uygulanmadığı durumda %92,07'ye çıktığını ortaya koymuştur. Bu kapsamda, örnek bir veri seti üzerinden sorumlu grupları ve ilgili gerçekleri bulacak modellerin uygulanıp test edilmesi sonucu ham verilerin güvenlik uzmanlarının çalışmalarında kullanılabilecek yararlı bilgilere dönüşebileceği ve bu bilgiler doğrultusunda tahminlerde bulunulmasıyla incelemelerin yönlendirilebileceğini değerlendirmiştir.

(Ayдын, 2011), Kalp Ritim Bozukluğu Olan Hastaların Tedavi Süreçlerini Desteklemek Amaçlı Makine Öğrenmesine Dayalı Sistemin Geliştirilmesi adlı tez çalışmasında K-Nearest Neighbor (K-NN) algoritmasını kullanmıştır.Seçilme sebebi olarak düşük biansa (tarafılık) sahip bir fonksiyon olmasından kaynaklı olduğunu belirtmiştir. Çalışmanın sonunda elde edilen

sonuçlara göre sistemin daha fazla geliştirilmesi gerektiği sonucuna ulaşmıştır.. Gerekçe olarak sistem bileşenlerinin birbirlerinden ayrık olması yerine tümleşik bir sistemin geliştirilmesi gerektiğini ,bilginin ulaştığı ve depolandığı bilgisayar sunucusunun kaldırılıp yerine gömülü sistemlerin kullanılmasının uygun olacağını belirtmiştir.

(Nizam, 2014), Sosyal Medyada Makine Öğrenmesi İle Duygu Analizinde Dengeli Ve Dengesiz Veri Setlerinin Performanslarının karşılaştırılması adlı akademik makale çalışmasında, makine öğrenmesi yöntemlerinden gözetimli öğrenme yaklaşımlardan Naive Bayes (NB), Random Forest (RF), Sequential Minimal Optimization (SMO), Decision Tree ve K-Nearest Neighbors(K-NN) kullanarak sosyal medyada duygu analizi çalışması yapmıştır. Sosyal medya aracı olarak popüler olması açısından ve karakter kısıtlamasından kaynaklı veri analizinde kolaylık sağlayacağı düşünülerek Twitter kullanılmıştır. Çalışmanın sonucunda Türkçe film yorumlarının duygu analizinde ML algoritmaları daha performanslı çıktılar üretmiştir. Yöntemler arasında başarı performans karşılaştırılması yapılmış ve %72.33 ortalama doğruluk başarı oranıyla SMO algoritmasını seçmiştir.

(Çolakoğlu, 2020), Makine Öğrenmesi Algoritmaları ile Avrupa Havalimanları Analizi adlı tez çalışmasında 2 Makine Öğrenmesi algoritması kullanmıştır. Denetimsiz Öğrenme Algoritması olan Kümeleme ve Denetimli Öğrenme modellerinden yararlanmıştır. Değerlendirilen tüm parametreler Bayesyen Lineer Regresyon Algoritmasının daha iyi sonuçlar ürettiğini ortaya koymuştur. Kümeleme Algoritmasını benzer havalimanlarını tespit etmek amacı ile de kullanmıştır. Bu sayede artan havacılık faaliyetleri çerçevesinde gelecekte yapılacak her türlü havacılık organizasyonuna bir çıkarım sağlanabileceğini yorumlamıştır. Taşınan Yük Sayısı, Rota Sayısı, Şehir ve Ülke Nüfusu ve diğer teknik özellikler kullanılarak taşınan yolcu sayısına ilişkin trendi görebilmek açısından bir öngörü sağlanabileceği ortaya koymuştur.

(Akşehir, 2019), Bu makale çalışmasında Makine Öğrenmesi Teknikleri ile Banka Hisse Senetlerinin Fiyat Tahmini üzerine denetimli öğrenme yöntemlerinden Çoklu Lineer Regresyon(MLR) , Random Forest(RF) ve Karar

Ağaçları(DT) kullanmıştır. Sonuç olarak ,borsa endeks tahminlerinden yola çıkılarak , banka hisse senetlerinin bir gün sonraki kapanış değeri tahmininde oldukça başarılı olduğunu, yapılan analizlerden çıkan R – squared değeri göz önünde bulundurularak kullanılan tüm metodlardan %98 oranında başarı elde ettiğini ortaya çıkarmıştır.

(Koç Ustalı, 2021),Makine Öğrenmesi Teknikleri İle Hisse Senedi Fiyat Tahmini araştırma makalesinde, 3 farklı denetimli makine öğrenmesi algoritması olan RF, XGBoot, YSA kullanmış ve Python programlama dilinden yararlanmıştır. Çalışmasında daha önce yapılan hisse senedi tahmini literatür araştırmalarına da yer vermiştir. Firmalara ait 2010-2019 yılları arası verilerin 3 aylık ortalamasını alarak veri setini oluşturmuş ve 3 algoritma ile %70 oranında veriyle veri setini eğitmiştir. Her bir algoritma 10 kez çalıştırılmıştır. Finansal tablolarında 22 şirketin verilerine erişmiştir. Eğitilen veriler ile hisse senedinde gelecek fiyat tahmini yapmıştır. Bu algoritmaların başarı performanslarını karşılaştırmıştır. Sonuç olarak XGBoot ve RF algoritmaları ile daha doğru tahminler yapabileceği kanısına ulaşmıştır.

(Gök, 2017),Makine Öğrenmesi Yöntemleri İle Akademik Başarının Tahmini adlı makale çalışmasında, ortaokul öğrencileri üzerinde 24 soruluk anket düzenlenmiştir. Türkçe, Matematik dersi ve dönem sonu başarılarının ortalamasını regresyon ve sınıflandırma yöntemleri ile modellendirmiştir. Anket soruları öğrenci üzerinde etkisi olabilecek etkenler baz alınarak hazırlanmıştır. Çalışma Yalova İl MEB müdürlüğünce ortak yürütülmüştür. Çalışmada 24 öznitelik ve 1492 gözlem bulunmaktadır. Öznitelik seçimi yapılmış ve KÖAK yöntemi ile öznitelikler seçilmiştir. Türkçe , Matematik , Genel Başarı değişkenleri için ayrı öznitelikler seçilmiştir. Yöntem olarak K-NN, Destek Vektör Makineleri, Radyal Tabanlı DVM, Random Forest, Lojistik Regresyon, Naive Bayes algoritmalarından yararlanmıştır. Makale çalışmasının sonucunda, Genel Başarı Ortalaması için bütün kullanılan yöntemlerde başarılı sonuçlar elde etmiştir. Türkçe ve Matematik dersi puan tahmininde RF , not tahmininde Lojistik Regresyonun daha başarılı olduğunu vurgulamıştır.

(Ergin Tosunoğlu, 2021), Eğitimde Makine Öğrenmesi Araştırmalarındaki Güncel Eğilimler Üzerine İnceleme adlı araştırma makalesinde, NB, DT, SVM algoritmalarından yararlanmıştır. Araştırmadaki amaç, eğitim bilimleri alanında ML algoritmaları kullanılarak 2015-2020 yılları arasında yazılmış makale eğilimlerini incelemiştir. Toplamda 201 makale incelenmiştir. Sorgulama sonucu 184 makale kriterlere uymuştur. Sonuç olarak bu makalelerde RF algoritmasının makalelerde en çok tercih edilen algoritma olduğu sonucuna ulaşılmıştır.

(Mansur Alp Toçoğlu, 2019), Türkçe Metinlerde Duygu Analizi İçin Farklı Makine Öğrenmesi Yöntemlerinin Karşılaştırılması adlı araştırma makalesinde duygu analizi sınıflandırılması yapılmış ve RF, YSA, K-NN, SVM algoritmaları kullanılmıştır. Duygu analizinde mutluluk, korku, öfke,üzüntü, tikslenme ve şaşırma kategorilerine yer vermiştir. F5 yöntemi kullanılarak kelimelerin kökleri ilk 5 karakter bulunmuştur.Metrik olarak doğruluk yönteminden yararlanmıştır. Sosyal medya aracı olarak Facebook ve Twitter ele alınmıştır. Çalışmada YSA algoritmasının 5 katmanlı yapısı kullanılmıştır. Algoritmaların ürettiği sonuçlar 10 katmanlı çapraz doğrulama yöntemi kullanılarak değerlendirilmiştir.Çalışma sonucunda YSA ve SVM algoritmaları birbirine yakın sonuçlar üretmiştir.

(Ömer Çağrı Yavuz, 2021), Astronomik Gözlem Kalitesi Tahmininde Makine Öğrenmesi Algoritmalarının Kullanımı adlı araştırma makalesinde , K-NN, k-Means ve NB algoritmalarını kullanmıştır. Çalışma sonucunda makine öğrenmesi yöntemlerinin performansları ölçülmüştür. Veri setinde 12 nitelik ve 2557 gözlem bulunmaktadır.K-means algoritması ile 3,4,5 küme şeklinde kümelenmiştir. Sonuç olarak, NB yönteminin başarısı 3 küme için %97 , 4 küme için %92dir. K-NN yönteminin başarısı 3 küme için %95, 4 küme için %91 olmuştur.

(Gökhan Goy),Makine Öğrenmesi Yöntemleri İle Kredi Kartı Sahteciliği Tespiti adlı makale araştırmasında, kredi kartı kullanımının artmasına bağlı olarak oluşabilecek sahteciliğin tespiti için bu çalışmasını yürütmüştür. K-NN, Lojistik Regresyon, AdaBoostM1 algoritmalarından yararlanmıştır.

Çalışmasında Python programa dilini kullanmıştır. Gürültüye sebep olabilecek veriler üzerinde veri temizleme işlemi gerçekleştirmiştir. Bu işlemten sonra daha performanslı sonuçlar üretebileceği kanısına ulaşmıştır. Veri dengesizliğinin önüne geçmek için hibrit bir yöntem olan SMOTETomek'den yararlanmıştır. Çalışma sonucunda K-NN algoritması %99.9 doğruluk oranına sahip olduğu için bu çalışma kapsamında en performanslı sonuç veren yöntem kabul edilmiştir.

(Yavuz E. , 2020), Konutlarda Enerji Tüketimi Kestirimi İçin Derin Öğrenme Ve Makine Öğrenme Yöntemlerinin Karşılaştırılması adlı yüksek lisans tez çalışmasında, derin öğrenme için Konvolüsyonel Sinir Ağı ve yapay sinir ağı kullanmıştır. Makine öğrenmesinde RF, K-NN ve doğrusal regresyon modellerinden yararlanmıştır. Bu çalışma enerji tüketimi üzerinde tahmin kar sonuçlar üretmek ve enerji yönetiminde kolaylık sağlamak için yapılmıştır. Çalışmada 43320 veri üzerinde gözlem yapılmıştır ve veri seti %80 veri ile eğitilmiştir. Çalışma 10 katlı çaprazlama yöntemiyle doğrulanmıştır. Doğrusal Regresyon modeli en performanslı model olarak seçilmiştir.

(Şenel, 2020), Makine Öğrenmesi Algoritmaları Kullanılarak Kayısı İç Çekirdeklerinin Sınıflandırılması adlı araştırma makalesinde K-NN, SVM, ÇKA, Gaussian NB, Adaptive Boosting, RF, DT yöntemlerinden yararlanmıştır. Çalışmadaki amaç, insan etkenini ortadan kaldıracak bir sınıflandırma algoritmasının geliştirmektir. Çalışmada öz nitelik değerlendirilmesinden sonra RF yöntemi en performanslı yöntem kabul edilmiştir. Çalışmada kullanılan makine öğrenmesi yöntemi ile kayısı iç çekirdeği sınıflandırılması yapılabileceğini ortaya koymuştur.

(Ağca, 2021), Otel Oda Fiyatlarını Açıklamada Makine Öğrenmesi Algoritmalarının Kıyaslanması adlı araştırma makalesinde amaç, otel odalarının fiyatlarının en doğru şekilde tahmin edecek bir algoritma modeli ortaya koymaktır. Çalışma kapsamında RF, DT, Lojistik Regresyon, SVM, K-NN algoritmalarından yararlanılmıştır. Çalışmada Python programlama dilini kullanmıştır. Analizler yapıldıktan sonra lojistik regresyon modelinin uygun model olduğu kanısına varmıştır.

Bu tez çalışmasında kullanılan RF, DT, K-NN, MLR, BTR algoritmaları yapılan literatür araştırması sonucu belirlenmiştir. En çok tercih edilen ve genellikle yüksek tahminkar sonuçlar üreten algoritmalar olduğundan tez çalışmasında yer verilmiştir. Bu kapsamda, veriye yapılan manipölasyon, veri temizliği, veriyi modelleme aşamalarından sonra bütün kullanılan algoritmalar orta ve yüksek denilebilecek oranlarda performans göstermiştir.

Bazı çalışmalarda hazır veri setleri kullanılmıştır. Kaggle yada internet siteleri üzerinden ulaşılan veri setleri genellikle mükemmeliyete yakın bir tablo oluşturmaktadır. Gerekli veri manipölasyon işlemleri yapılmış olup bu veri setleri üzerinde oluşturulan modellerin sonuçları çok tahminkar ve model performansı genellikle %90'ın üstünde olmaktadır.

Ancak Makine Öğrenmesi İle Eğitim Başarısının Tahmini Modeli adlı tez çalışmasında veri seti özenle hazırlanan 13 anket sorusunun cevaplarına bağlı olarak oluşturulmuştur. Bu sebeple veri üzerinde uygulanan her aşama dikkatli bir şekilde ele alınmıştır.

Öğrencilerde Eğitim Başarısını Etkileyen Faktörler Üzerine Literatür Taraması

Bu bölümde tez çalışmasında kullanılan anket sorularının hazırlanma aşamasında yapılan literatür taramasına göre doğrudan veya dolaylı yoldan anket sorularının oluşmasını sağlayan referanslara yer verilmiştir.

(Araştırma Görevlisi Koçoğlu, 2010), İlköğretim 2. Kademe Öğrencilerinin Uyku Alışkanlıkları ve Uyku Sorunlarının Bazı Okul Fonksiyonlarına Etkisi adlı araştırmasında 6. 7. ve 8. Sınıf öğrencilerine yönlendirilen anket yöntemiyle veriler toplamıştır. Araştırmada ulaşılan sonuçlara göre okul çağındaki çocuklar için olumsuz uyku alışkanlıklarının olması ve yaşanan uyku sorunlarının düşük okul başarısıyla, devamsızlık ve gün içinde dikkati toplayamama problemleriyle yakından ilişkili olduğu kanısına varmıştır.

(Aslanargun, 2016), Sosyo Ekonomik Değişkenlerin Öğrencilerin Akademik Başarısı Üzerine Etkileri adlı çalışmasında, öğrencilerin başarısını etkileyen aile eğitimi, öğrenci cinsiyeti, ekonomik durum, kardeş sayısı gibi

faktörlerin başarıyla nasıl bir ilişki içerisinde olduğu yönünde bir araştırma yapmıştır. Öğrencilere ulaştırılan anket soruları ile veriler toplanmış ve analizi gerçekleştirilmiştir. Araştırmanın sonunda, tek çocuk veya çok kardeşe sahip olma durumunun, 2 kardeşe sahip olma durumuna oranla daha az başarılı olduğu, gelirin öğrenci başarısında pozitif etkisinin olduğu, cinsiyet faktörünün erkekler için olumlu yönde olduğu sonuçlarına ulaşmıştır.

(Dam, 2008), Öğrencinin Okul Başarısında Aile Faktörü adlı araştırmasında, otobiyografi tekniği ile öğrencilerden topladığı veriler doğrultusunda analiz yapmıştır. Öğrencilerden toplanan bilgilere göre; aileleriyle yaşadıkları iletişimsizlik, diğer kişilerle kıyaslanma durumu, ailelerinin fazla beklenti içerisine girdikleri, aile içi huzursuzluk, boşanma, ailede yaşanan ölüm, uygun çalışma ortamının bulunmaması, ekonomik durumlar başarıya etki eden faktörlerden olmuştur.

(Doç. Dr Şentürk, 2012), Parçalanmış Aile Çocuklarının Eğitimdeki Başarı/Başarısızlık Durumu adlı makalesinde, öğrencinin eğitim başarısında aile faktörünü incelemiştir. Anket, grup çalışması ve mülakat yöntemlerinden yararlanmıştır. Çalışma sonucunda, aileleri tarafından desteklenen çocukların başarısının pozitif yönde olduğu, aile içinde ölüm, boşanma gibi faktörlerde ise çocuğun eğitiminde negatif etkiler olabileceği kanısına ulaşmıştır.

(Gelbal, 2008), Sekizinci Sınıf Öğrencilerinin Sosyoekonomik Özelliklerinin Türkçe Başarısında Etkisi adlı çalışmasında, öğrencilerin Türkçe notundaki başarısını betimleyebilmek için Türkçe dersi testi uygulamıştır. Bu teste ek olarak öğrencilerden bazı bilgiler alınmıştır (Kardeş sayısı, anne eğitim düzeyi, ev içerisindeki olanaklar). Türkçe değişkeni bağımlı değişken kabul edilerek öğrencinin başarısını etkileyen etkenlerde evdeki olanakların başarı ile doğru orantılı olduğu, cinsiyetin başarıya etki ettiği, kardeş sayısındaki artışın başarıya negatif yönde etkilediği sonucuna ulaşmıştır.

(Cumhur, 2018), Öğretmenlerin Görüş Ve Önerileri Bağlamında Öğrencilerin Matematiksel Başarısını Etkileyen Faktörlerin İncelenmesi adlı araştırma makalesinde, öğrencinin Matematik dersi başarısını etkileyen etkenlerin ortaya çıkarılması ve çözüm önerileri ile ilgili öğretmen görüşlerine

ulařmıřtır. Yarı yapılandırılmıř g r řme yoluyla verileri elde etmiřtir.  alıřma sonunda  ğretmen g r řlerine g re aile,  ğrenme ortamı, okul dıřı yardımcı kurslar,  ğrenmeye ayrılan zaman ve teknolojinin bařarıyı ekledięi g r ř n  ortaya  ıkarmıřtır.

( avuř Erdem, 2014), Uyum Analizi Y ntemiyle Matematik Bařarısını Etkileyen Fakt rlerin  ncelenmesi adlı  alıřmasında,  ğrencilerin Matematik dersi  zerindeki bařarısını etkileyen etkenler  zerinde  alıřmıřtır.  ğrencilerden veriler anket yoluyla toplanmıřtır. Cinsiyet, Teog puanı, aile eęitim durumu, dersane ve  zel ders alma gibi sorulara yer vermiřtir.  alıřmasının sonunda,  ğrencilerin bařarısına babanın eęitim durumunun,  ğrencinin  zel eęitim almasının etki ettięini tespit etmiřtir. Ayrıca dershaneye giden  ğrencilerin bařarısında pozitif y nde bir etki olduęunu g zlemlemiřtir.

(Sarier, 2016), T rkiye'de  ğrencilerin Akademik Bařarısını Etkileyen Fakt rler: Bir Meta-Analiz  alıřması adlı arařtırmasında,  ğrencilerin bařarısını etkileyebilecek etkenler T rkiye'de yapılan 62  alıřma ve meta-analiz  alıřması yapılarak elde edilmiřtir.  alıřma sonunda, sosyo-ekonomik durumun  ğrencilerin bařarısında etkili olduęu kanısına ulařmıřtır.

(Yavuz A. , 2020), Orta ğretime Ge iř Sınavında  ğrenci Bařarısını Etkileyen Etmenler adlı tez  alıřmasında,  ğrencilerin bařarısıyla okul dıřındaki etmenlerin (Ailelerin ekonomik durumları, eęitim d zeyleri,  zel ders, dersane ve kursla) iliřkisini arařtırmıřtır. Anket y ntemiyle verileri elde etmiřtir.  alıřma sonunda vardıęı sonu lardan biri gelir durumunun  ğrenci bařarısının  zerinde etki bulunmaktadır řekilde olmuřtur.

(Karaaęa  Cing z, 2020), Ekonomik, Sosyal ve K lt rel Stat n n Akademik Bařarıya Etkisi adlı  alıřmasında,  ğrencilerin ekonomik, sosyal ve k lt rel stat lerini analiz edip dezavantajların oluřturduęu eřitsizlięin  l  tlerini ortaya  ıkarmaktır. Verilere PISA 2015 anketi aracılıęıyla elde etmiřtir.  alıřmada tarama modeli kullanılmıřtır.  alıřma sonunda,  ğrencinin ekonomik, sosyal ve k lt rel  zelliklerinin bařarı  zerinde  nemli bir etki d zeyine sahip olduęu kanısına ulařmıřtır.

(Okatan, 2021), PISA 2018 Türkiye Okuma Başarısının Bazı Değişkenler Açısından İncelenmesi adlı araştırma makalesinde, bazı değişkenler (cinsiyet, öğrencinin sosyo-ekonomik durumu) ESCS indeksi açısından incelemiştir. Veri toplama kaynağı olarak OECD'nin resmi web sitesinden yararlanmıştır. Araştırmasının sonunda, cinsiyet ve sosyo-ekonomik değişkenlerin okuma başarısında etkisi olduğu sonucuna ulaşmıştır.

(Sağlam, 2020), PISA 2018 Araştırmasına Etki Eden Duygusal Faktörlerin Veri Madenciliği Yöntemleri ile İncelenmesi adlı makalesinde, öğrencilerin matematik, fen dersi ve okuma puanlarını öğrencilerin mutlu ve üzgün olma durumu, hayatından memnun olup olmama durumu ile ilişkisini incelemeye yönelik olduğunu belirtmiştir. Çalışma sonundaki analiz ile öğrencilerin mutlu olduğunda derslerdeki başarısının arttığı, iyi okullarda okuyan ve ailesi tarafından ihtiyaçları giderilen bir çocuğun başarısında artış olabileceğini vurgulamıştır.

(Dr.Gürsakal, 2012), PISA 2009 Öğrenci Başarı Düzeylerini Etkileyen Faktörlerin Değerlendirilmesi adlı çalışmasındaki amaç, Türkiye örneğini istatistiksel yöntemler kullanarak analiz etmektir. Bu amaçla, ilk olarak toplam 170 okuldan 4996 15 yaşındaki öğrencide okumayı ve temel fen ve matematik eğitimini etkileyen faktörleri belirlemek için t ve F testi kullanılmaya çalışılmıştır. Daha sonra çok değişkenli bir istatistiksel analiz tekniği olan lojistik regresyon analizi kullanılarak öğrencilerin fen, matematik ve okuma puanlarını etkileyen faktörleri belirlemek için bir girişimde bulunulmuştur. Analiz sonuçları; öğrenci başarı oranlarının cinsiyet, okul yaşı ve ebeveyn eğitimi gibi değişkenlere göre farklılık gösterdiğini göstermiştir.

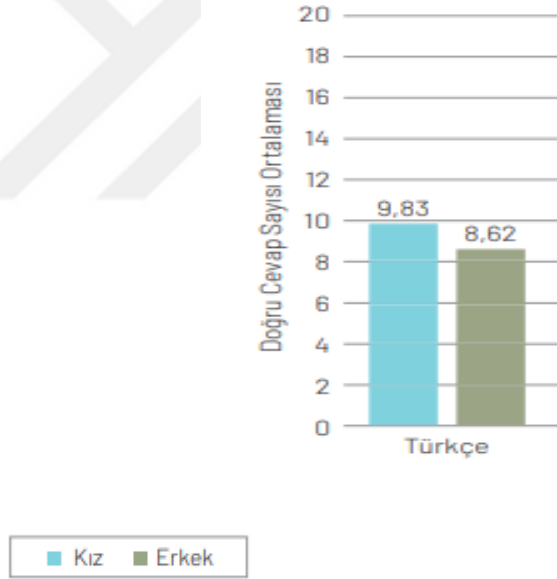
(Gök,Murat, 2017), Makine Öğrenmesi Yöntemleri İle Akademik Başarının Tahmini adlı araştırma makalesinde, öğrencilerin yaşam koşullarının ve sosyal çevrelerinin Türkçeyi nasıl etkilediğini belirlemek için ilk olarak 6, 7 ve 8. sınıflardaki öğrencilere 24 maddelik bir anket uygulanmış ve etkisini değerlendirilmiştir. Anket sorularına ebeveyn sağlığı, ebeveyn istihdamı, eğitim durumu, üvey baba durumu, öğrenci yaşı, kardeş sayısı, ev durumu, öğretmen durumu, oyun sıklığı, haftalık çalışma süresi, yaşam doyumu gibi sorulara yer vermiştir. Bu araştırma makalesinde kullanılan anket sorularından bazıları araştırma makalesi için referans olarak kullanılmıştır.

BİRİNCİ BÖLÜM

1. TÜRKİYE’DE EĞİTİM BAŞARISINI ETKİLEYEN FAKTÖRLER

Bu tez çalışmasında ortaya konulan problemin detaylandırılması için Türkiye’de eğitim başarısını etkileyen faktörlerden bahsedilmiştir.

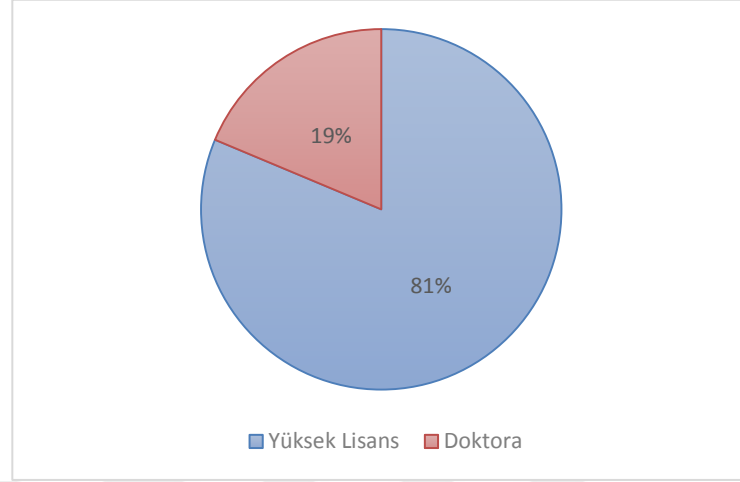
Milli Eğitim Bakanlığı 2022 verilerine göre 5 milyon 885 bin ortaokul öğrencisi bulunmaktadır. Bunlardan 1 milyon 84 bin 799’u 8.sınıf öğrencisidir. Ortaokul son sınıf öğrencilerinin sınavda yaptığı Türkçe dersi net ortalaması 20 soruda 9.22, standart sapması 4.22’dir. Bu oran 1.sınıftan itibaren 7 sene boyunca Türkçe eğitimi alan bir birey için oldukça düşüktür. Milli Eğitim Bakanlığı 2022 raporuna göre kız öğrencilerin Türkçe dersi net sayısı 9.83 iken erkek öğrencilerin 8.63’tür. Şekil 2’de bu dağılım görsel olarak gösterilmiştir.



Şekil 1. Cinsiyete Göre Türkçe Dersi Doğru Cevap Sayısı Ortalaması

Türkçe dersi bir öğrenci için en önemli derslerden biridir. Ana dilini iyi bilen bir birey hayatın tüm aşamalarında daha başarılı olması beklenir. Ayrıca okuma , anlama ve anlatım yeteneği gelişir. Anlama yeteneği gelişmiş öğrenci sayısal dersleri de daha iyi anlar. Okullardaki eğitim , öğrencinin okul dışındaki sosyal hayatı , ailesi , arkadaşlığı , öğretmenlerin yaklaşımları başarıyı etkileyen bir çok faktördendir.

Ulusal tez merkezinin 2022 yılına ait verilerine göre Makine Öğrenmesi anahtar kelimesi kullanılarak yapılan tarama sonucunda 664 adet teze ulaşılmıştır. Yazılan tezlerin çoğunluğunu Yüksek Lisans tezleri oluşturmaktadır.



Şekil 2. Lisansüstü Tezlerin Dağılımı

Literatürde Makine Öğrenmesi ile ilgili farklı algoritmalar kullanılarak çok sayıda çalışma yapılmıştır. Makine öğrenmesinin yaygınlaşması ve teknolojinin ilerlemesi ile beraber birçok sektörde kullanılmaya başlanmıştır.

İKİNCİ BÖLÜM

2. MAKİNE ÖĞRENMESİ

Makine öğrenimi, tükettikleri verilere göre öğrenen yada performansı iyileştiren sistemler oluşturmaya odaklanan bir yapay zeka (AI) alt bileşenidir. Makine öğrenmesinin öncüsü kabul görülen Arthur Samuel bunu “bilgisayara açıkça programlanmadan öğrenme kabiliyeti kazandıran bir çalışma alanı” olarak tanımlamıştır.

2.1 Temel Kavramlar

Makine öğrenimi disiplini, bilgisayarlara tam olarak tatmin edici bir yöntemin bulunmadığı görevleri gerçekleştirmeyi öğretmek için çeşitli örnekler kullanır. Makine öğrenimi ve yapay zeka çoğunlukla birlikte değerlendirilmektedir. Kimi durumlarda birbirinin yerine kullanılsada, aynı anlama gelmezler. Tüm makine öğrenimi çözümleri yapay zekanın altında değerlendirilirken, tüm yapay zeka çözümlerinin makine öğrenimi olmadığı önemli bir gerçektir. Şekil 1’de görsel olarak belirtilmiştir.



Şekil 3. Makine Öğrenmesi Ve Yapay Zeka

2.2 Yapay Zeka

Yapay Zeka 1950’li yıllarda yeni gelişmeye başlamış ve bilgisayar bilimcilerinde bugün cevap aradığı “bilgisayarlar düşünebilir mi?” sorusuna

cevap ararken ortaya çıkmıştır. Yapay zeka kısaca “normal insanlar tarafından yerine getirilen düşünsel faaliyetleri otonom hale getirme olarak tanımlanabilir. (Chollet, 2019, p. 4)

Yapay zekâ, belirlenen görevleri yerine getirmek amacıyla insan beynini taklit etmektedir. Görev sonucunda edindiği tecrübeler sayesinde kendisini tekrar ederek gelişebilen sistemler olarak isimlendirilmektedir.

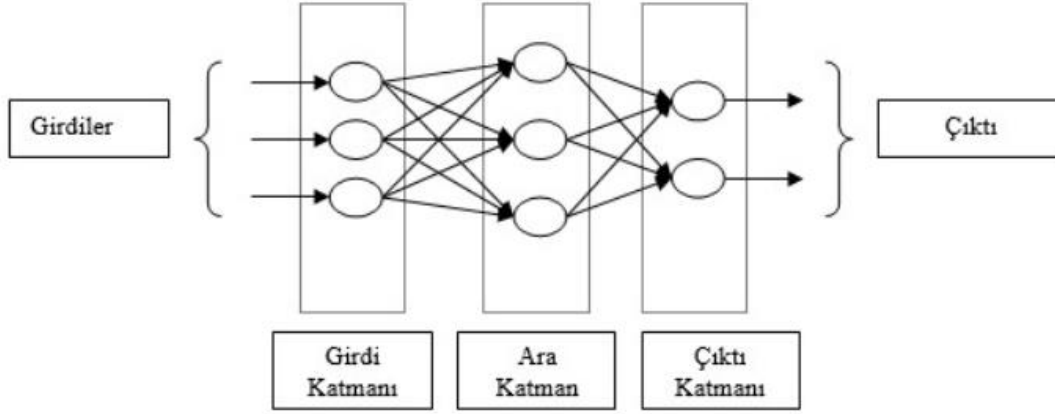
2.3 Derin Öğrenme

Derin öğrenme, yapay sinir ağlarının büyük miktarda veriden öğrendiği makine öğreniminin (Machine Learning veya kısaca ML) bir altkümesidir. En az bir adet yapay sinir ağının kullanıldığı ve birçok algoritma ile bilgisayarın eldeki verilerden yeni veriler elde etmesidir.

Derin derken gösterilmek istenen derin öğrenmenin birtakım derin bilgiler elde etmesi değil, birbirini takip eden gösterim katmanlarını ifade etmektir. Bu katmanlı gösterim, sinir ağı olarak adlandırılan ve birbirini takip eden katmanları olan model sayesinde öğrenilmektedir. Derin öğrenme, teknik olarak veri gösterimi öğrenmenin çok aşamalı bir yoludur. (Chollet, 2019, p. 8)

Derin öğrenmede yapay sinir ağları (YSA) kullanılır. Bir sinir ağında 3 farklı nöron katmanı bulunmaktadır: Giriş katmanı, gizli katmanlar, çıktı katmanı olarak değerlendirilmektedir. Nöronlar arasındaki iletişimler, giriş değerinin önemini belirleyen bir ağırlıkla ilişkilendirilmektedir. Şekil 4’te derin öğrenmenin katmanları gösterilmiştir.

Yapay sinir ağı insan beyninin öğrenme sürecinden etkilenecek ortaya atılmış bir yaklaşımdır. Bu yaklaşım ilk kez 1943 yılında insan beynindeki hücrelerin yapısının matematiksel modellenmesi oluşturularak gerçekleştirilmiştir. (McCullonch, 1943)

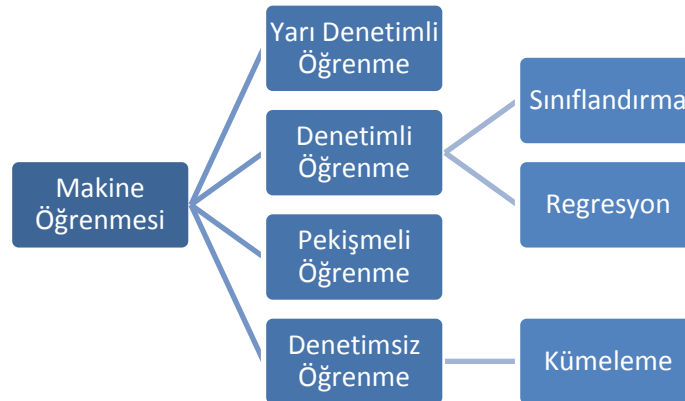


Şekil 4. Derin Öğrenme Katmanları

2.4 Makine Öğrenmesi

Makine öğrenmesi matematiksel ve istatistiksel yöntemler ile mevcut girdilerden çıkarımlar yapan bu çıkarımlarla bilinmeyene dair öngörülerde bulunan bir yöntemdir. Makine öğrenmesi olmazsa bir veriyi modellemek oldukça zordur. Kullanılan metodlar geçmişe dönük girdileri baz alır. İleriye doğru yeni girdilere göre öngöründe bulunabilecek en uygun yöntemin geliştirilebilmesini sağlar.

Makine öğrenmesinin yöntemleri; denetimli, denetimsiz, pekişmeli ve yarı gözetimli algoritmalarıdır. Şekil 5’te Makine öğrenmesinde kullanılan öğrenme türlerinin hiyerarşik yapısı gösterilmiştir.



Şekil 5. Makine Öğrenmesi Yöntemleri

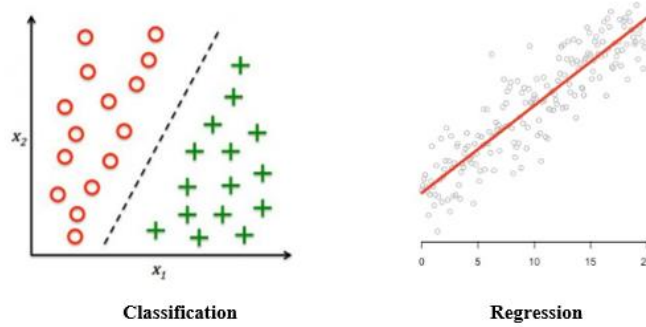
2.4.1 Öğrenme Türleri

Gözetimli Öğrenme : Denetimli öğrenme olarak da bilinir. Veri setinde giriş ve çıkış bilgilerinin bulunduğu öğrenme türüdür. Verileri sınıflandırmak ve sonuç bilgilerimi doğru bir şekilde tahmin eden modelleri eğitmek için etiketli veri sınıflarının kullanılması olarak tanımlanır. Denetimli öğrenmede bir bağımlı y değişkeni ve bağımsız X değişkenlerinin birliktaki etkileşimleri sonucunda ortaya çıkardıkları durumun ne olduğu bilgisidir. Denetimli öğrenme yöntemlerinde sınıflandırma(classification) ve regresyon kullanılır. Naive bayes, yapay sinir ağları, basit ve çoklu doğrusal regresyon, lojistik regresyon(LR), rastgele orman veya random forest(RF) ve destek vektör makinesi (SVM) gibi yaklaşımlar bu öğrenme türünün örneklerindedir.

Regresyon: Sayısal girdiler sisteme verilir ve bu sayısal değerler arasındaki ilişki matematiksel fonksiyonla gösterilir. Sistem çıktı olarak sayısal bir değer üretir.

Sınıflandırma: Sistem verilen girdiler doğrultusunda çıktı değişkeni kategorik bir değişken ise bu yöntem sınıflandırmadır.

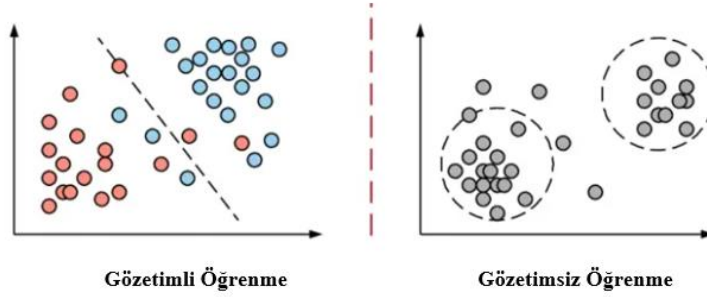
Sınıflandırma ve regresyona ait grafiksel gösterimler Şekil 6'da verilmiştir.



Şekil 6. Sınıflandırma Ve Regresyon Yöntemi

Gözetimsiz Öğrenme: Denetimsiz öğrenme türü olarakta bilinmektedir. Makine öğrenmesi yöntemlerini çözümlmek ve etiketsiz veri sınıflarını kümelemek için kullanılır. Bu öğrenme türünde y bağımlı değişkeni ve etiket bilgisi bulunmamaktadır. Etiket bilgisi olmadan girdi bilgileri ile tespit yapılır. Girdiler kendi içinde organize olarak matematiksel bir ilişki kurar. Dolayısıyla

sınıf bilgisi olmayan veri içerisindeki grupların ortak özellikleri ortaya çıkar. Yapay sinir ağları, K Nearest Neighbor (K-NN) , K means (kümeleme) en yaygın kullanılan yöntemlerdendir.



Şekil 7. Gözetimli Ve Gözetimsiz Öğrenme

Yarı Gözetimli Öğrenme: Yarı gözetimli öğrenme, denetimli ve denetimsiz tür arasında bir orta yol sunar. Eğitim esnasında, etiketsiz bir veri kümesinden sınıflandırmaya ve nitelik ayrıştırmaya yön vermek için daha küçük, etiketli bir veri kümesi kullanır. Yarı denetimli öğrenme, denetimli bir öğrenme modelini eğitmek için yeterli etiketlenmiş veriye sahip olmama problemini çözebilir.

Pekiştirmeli Makine Öğrenmesi: Denetimli öğrenmeyle benzerlik gösteren ancak modelin basit veri kullanılarak eğitilmediği bir davranışsal makine öğrenmesi yöntemidir. Bu makine öğrenmesi türü, deneme yanılma metoduyla öğrenir. Belirli bir problem için en iyi teklifi ya da prensibi geliştirmek üzere başarılı sonuçlar dizisi pekiştirilir.

2.4.2 Makine Öğrenmesi Algoritmaları

Makine öğrenmesinde kullanılan birçok algoritma mevcuttur. Bu alt başlıkta en çok tercih edilen yöntemler anlatılacaktır.

Regresyon tahmin modelleri aldığı değerle sürekli bir sonuç üretir. Bu sürekli değerler genel olarak miktarlar ve boyutlardır.

2.4.2.1 Basit Doğrusal Regresyon (SLR)

Basit doğrusal regresyon, iki sürekli değişken arasındaki ilişkileri özetlemesini ve analizini sağlayan istatistiksel bir metottur. Doğrusal regresyon,

giriş değişkenleri (x) ile tek çıktı değişkeni (y) arasında doğrusal bir bağlantı olduğunu varsayan bir yöntemin bulunduğu doğrusal bir algoritmadır.

Basit doğrusal regresyon, uygulanabilecek en basit makine öğrenmesi metotudur. Sadece bir x değişkeni olduğundan çok kompleks bir yapıya sahip değildir. Daha fazla değişken probleme dahil olduğunda basit doğrusal regresyon modeli yetersiz kalır. Bu durumda Çoklu Doğrusal Regresyon , Polinomal Regresyon , Lasso Regresyon, Elastik Regresyon , Ridge Regresyon modelleri kullanılır.

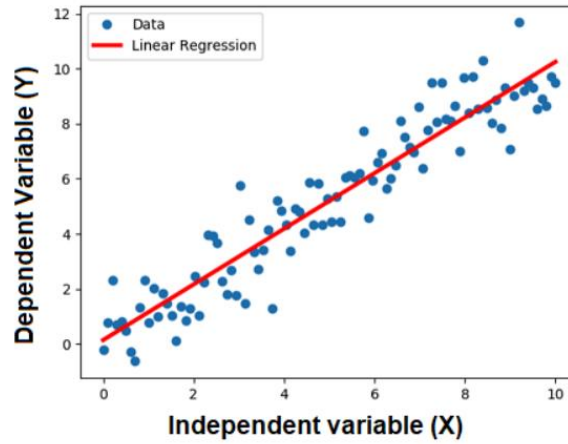
$$y = \beta_0 + \beta_1 X + \epsilon$$

Formülde Y değeri bağımlı değişkeni ifade eder.

β_0 , X ,0 olduğunda y nin öngörülen değeridir.

β_1 , regresyon katsayısıdır.

X, y bağımlı değişkeni üzerinde etki eden bağımsız değişkendir.



Şekil 8. Basit Lineer Regresyonun Grafiks gösterimi

2.4.2.2 Çoklu Doğrusal Regresyon(MLR)

İki veya daha fazla bağımsız değişken (X) ile bir bağımlı değişken (y) arasındaki ilişkiyi tahmin etmek için kullanılan yöntemdir.

Çoklu doğrusal regresyon(MLR veya Multiple Linear Regression), basit doğrusal regresyonla aynı öngörülerde bulunur:

Varyansın saflığı: Öngörümüzdeki hatanın boyutu, y değişkeninin değerleri arasında önemli ölçüde değişmez.

Gözlemlerin bağımsızlığı: Veri seti içerisindeki gözlemler sayımlama olarak geçerli metotlar kullanılarak elde edilmiştir ve değişkenlerin arasında saklı ilişkiler bulunmamaktadır.

Çoklu doğrusal regresyon modelinde, bazı bağımsız değişkenlerin birbiriyle ilişkili içinde olması mümkündür. Bu nedenle regresyon yönteminde geliştirmeden önce bunları kontrol etmek gerekmektedir. Eğer iki bağımsız değişken çok yüksek oranda ilişkiliyse o zaman regresyon yönteminde sadece biri kullanılmalıdır.

$$y = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n + \epsilon$$

y = bağımlı değişkenin tahmin edilen değeridir.

β_0 = y kesişimi sabitidir. (tüm parametreler 0'a olduğunda y 'nin aldığı değerdir.)

$\beta_1 X_1$ = birinci bağımsız değişkenin (X_1) regresyon katsayısıdır (β_1).

... = test ettiğiniz birçok bağımsız değişken için aynısını yapılır.

$\beta_n X_n$ = n (son) bağımsız değişkenin regresyon katsayısıdır.

ϵ = model hatası olarak bilinmektedir (y tahminimizde ne kadar varyans (tutarlılık) var olduğunu ortaya koymaktadır).

2.4.2.3 Ridge Regresyon

Ridge regresyon, verilerdeki ortak değişkenlerin çoklu bağlantı sorununu çözmek için klasik bir veri modelleme metotudur. (Kennard, 1970, pp. 55-67) Burada çoklu iletişim, çoklu regresyon yönteminde birden fazla tahminci değişkenin yüksek oranda korelasyonu olduğu bir durumu ortaya çıkarmıştır. Çoklu iletişim mükemmel ise, regresyon katsayıları belirsizdir ve standart sapmaları ebebidir. Mükemmelden düşükse, regresyon katsayıları aşıkarak olmakla birlikte büyük standart sapmalara sahiptir ve bu katsayıların büyük bir doğrulukla öngörülemediği anlamına gelir. (Gujarati, 1998, pp. 209-212)

Çok doğrusal regresyon verilerini incelemekte kullanılır. Buradaki amaç hata kareler toplamını minimize eden regresyon katsayılarını, bu katsayılara bir alpha değeri uygulayarak bulmaktır. Over-fitting'e(aşırı düzeltme) karşı direnç gösterir.

Tüm değişkenler ile iletişim kurar, ilgisi olmayan değişkenleri ortadan kaldırmaz, yalnızca katsayılarını sıfıra yaklaştırır. Modeli yaratırken ceza değeri alpha için iyi bir değer bulmak gerekmektedir

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Yukarıdaki formüldeki amaç hata kareler toplamını minimize eden küçük varyanslı regresyon katsayılarını bulmaktır.

y_i = Veri serindeki gerçek değerlerdir.

$\hat{y}_i$ = Bağımsız X değişkenleriyle tahmin edilen değerlerdir.

$$SSE_{L_2} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^p \beta_j^2$$

$\alpha \sum_{j=1}^p \beta_j^2$ = Ceza terimidir.

α = Parametresi ayar parametresidir. Çapraz doğrulama yöntemi ile belirlenir.

Çapraz doğrulama yöntemi cross validation olarak da bilinir ve veri setini k adet alt parçaya ayırır. Belirlenen alt kümelerden birisi dışarda bırakılır . Kalan diğer parçalar ile oluşturulan model dışarıda bırakılan set ile test edilir.

2.4.2.4 Lasso Regresyon

Ürettiği modelin öngörü doğruluğunu ve açıklanabilirliğini arttırmak için hem değişken seçimi hem de düzenleme işlemi yapmaktadır. Ridge regresyon modelinde olduğu gibi hata kareler toplamını (MSE) minimize eden regresyon katsayılarını, katsayılara alpha ceza değerini uygulayarak bulmaktadır.

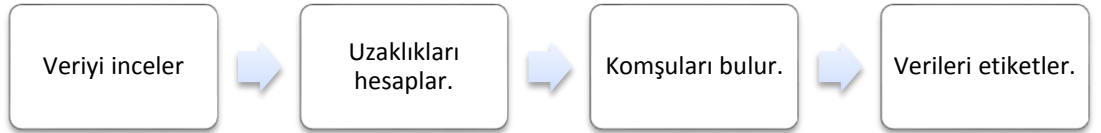
$$SSE_{L_1} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^p |\beta_j|$$

$\alpha \sum_{j=1}^p |\beta_j|$ = Ridge regresyondan farkı ceza uygularken katsayıların karelerini almak yerine mutlak değerlerini kullanır.

2.4.2.5 K Nearest Neighbors (K-NN)

K Nearest Neighbor (K-NN) içerisinde öngörülecek değerin bağımsız değişkenlerinin oluşturduğu vektörün en yakın komşularının hangi sınıfta yoğun olduğu bilgisi üzerinden sınıfını öngörmeye dayanır. K-NN hem sınıflandırma hem regresyon problemlerinde kullanılan bir algoritmadır.

K-NN adımları Şekil 6’da belirtilmiştir.



Şekil 9. K-NN Adımlarının Gösterimi

- i. Veriyi inceleme aşamasında k parametresi belirlenir.
- ii. Uzaklık hesabı yapılırken 3 farklı formül kullanılır; Öklid, Minkowski , Manhattan.
- iii. İncelenecek veriye en yakın komşuyu seçer.
- iv. Veriyi etiketler yani beklenen gözlem değerinin sınıfı olarak kabul edilir.

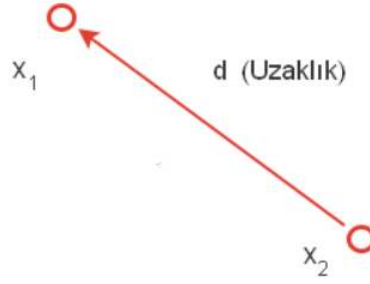
Euclidean (Öklidyen) Uzaklık hesaplaması;

$$\text{Öglid} = \sqrt{\sum_{i=1}^k (x_i - y_i)^2}$$

k = Boyut sayısıdır (komşu sayısı).

x_i = Veri seti içerisindeki değerlerdir.

y_i = Tahmin edilecek değerlerdir.



Şekil 10. Euclidean Uzaklık Hesaplaması temsili görünümü; (Noktasal)

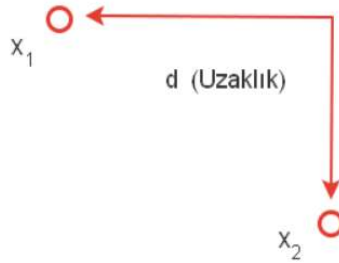
Manhattan Uzaklık hesaplaması;

$$Manhattan = \sum_{i=1}^k |x_i - y_i|$$

k = Boyut sayısıdır (komşu sayısı).

x_i = Veri seti içersindeki değerlerdir.

y_i = Tahmin edilecek değerlerdir.

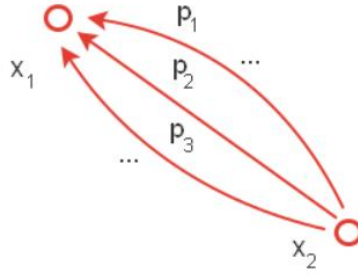


Şekil 11. Manhattan Uzaklık Hesaplaması temsili görünümü; (Noktasal)

Minkowski Uzaklık hesaplaması ;

$$Minkowski = \left(\sum_{i=1}^k |x_i - y_i|^p \right)^{1/p}$$

p integer bir değerdir. p değeri 1'e eşit olduğunda Manhattan uzaklığına , p değeri 2 olduğunda Öglid uzaklığına eşit olur.



Şekil 12. Minkowski Uzaklık Hesaplaması temsili görünümü; (Noktasal)

2.4.2.6 Karar Ağaçları (DT veya Decision Tree)

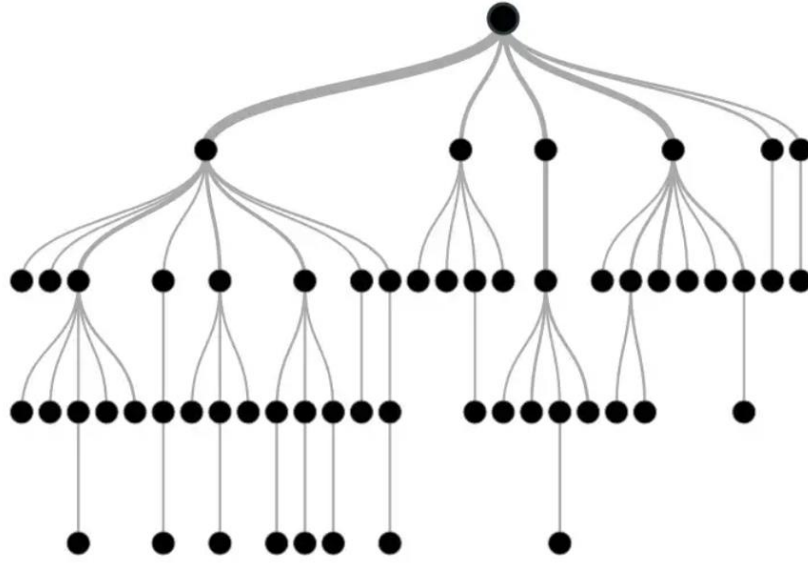
Sınıflama ve regresyon metodlarının bir algoritması olan karar ağaçları, yorumlanmalarının kolay olması, veri tabanı sistemleri ile kolay bir şekilde entegre edilebilmeleri ve güvenilirliklerinin yüksek olması nedenleri ile sınıflama metodlarının içerisinde en yaygın kullanıma sahip modeldir. (Ayık Y.Z, 2007, pp. 441-454).

Karar ağaçları kök , düğüm ve yapraklardan oluşur. İlk hücresi köktür. Kök hücreyi seçerken çeşitli yöntemler bulunur; Gini , Entropi, Cart algoritmalarıdır.

Gini Metodu : Kök hücreyi bulabilmek için her bir düğüm için Gini değeri hesaplanır.

Entropi Metodu: Entropi metodu , daha dengeli bir ağaç ortaya çıkarmak için kullanılır.

Cart: İki gözlem birimi kalana kadar ağacı bölmeye devam eder.

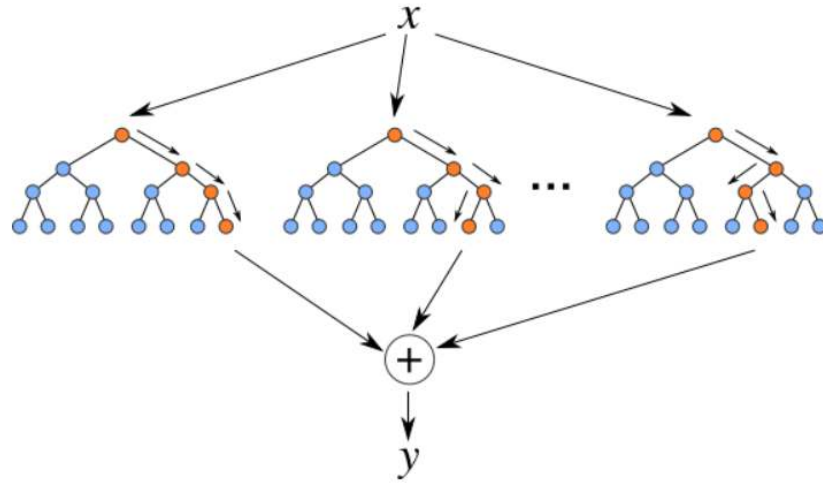


Şekil 13. Karar Ağaçları Şematik Gösterimi

2.4.2.7 Random Forest (RF)

Random forest yöntemi ,gözetimli sınıflandırma metodlarından biridir. Hem regresyon hem de sınıflandırma problemlerini ortaya çıkarmakta kullanılır. Model, birçok karar ağacı üreterek sınıflandırma işlemi sırasında sınıflandırma değerini yükseltmek hedeflenmektedir. Random forest modeli birbirinden bağımsız olarak çalışan birden fazla karar ağacının bir araya gelmesi sonucu, aralarından en yüksek puan alan değerin seçilmesidir.

Ağaç sayısının artması kesin bir sonuç elde etme oranını arttırmaktadır. Karar ağaçları modeli ile arasındaki temel fark ise, random forest modelinde kök düğümü bulma ve düğümleri bölme işlemlerinin rastgele gerçekleşmesidir.



Şekil 14. Random Forest yapısının gösterimi

2.4.2.8 Naïve Bayes (NB)

Naïve Bayes belirli bir sınıf için girişlerin birbirinden bağımsız olduğunu varsayarak sınıf koşulu olasılıkları tarafsız öngörmektedir. Naïve Bayes, sınıfları koşullu marjinal yoğunlukları bulmak amacıyla ayırt edici sınıflar problemini azaltmaktadır. (Molina R., 1994). Naïve Bayes, birbirleri ile ilişkili giriş verisi içermedikçe diğer modellere karşı iyi bir performans göstermektedir (Hastie, 2009, p. 533).

Bayes Teoremi

Bayes teoremi, bir rassal değişken için koşullu ihtimaller ile önsel ihtimaller arasındaki ilişkiyi göstermektedir.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$P(A|B)$: B olayı gerçekleştiğinde , A olayının meydana gelme ihtimalidir.

$P(B|A)$; A olayı gerçekleştiğinde , B olayının meydana gelme ihtimalidir.

$P(A)$ ve $P(B)$: A ve B olaylarının önsel ihtimalidir.

Naive Bayes Sınıflandırma Modeli

Naive Bayes sınıflandırma problemi, birden fazla özellikten ve bir hedef değişkeninden oluşmaktadır.

$$p(C|F_1, \dots, F_n) = \frac{P(C)pP(F_1, \dots, F_n|C)}{p(F_1, \dots, F_n)}$$

C verilen sonuç, F ise özellikleri temsil eder. Naive bayes ,bütün koşullu ihtimallerin çarpımıdır.

2.4.2.9 Destek Vektör Makineleri (SVM)

Bu model ikinci dereceden optimizasyon problemini çözmektedir. Yüksek boyutlu değişken uzayda doğrusal olmayan karar sınırlarını tanımlama yeteneğine sahiptir (Hsu, 2010).

Destek Vektör Makinesi modeli, iki ya da daha fazla sınıfı birbirinden ayırmaktadır. Sınıflandırma ve regresyon problemlerinden kullanılır. Sınıfları birbirinden ayırmak için sonsuz adet düzlem belirlenebilir. Destek Vektör Makinesi modelindeki amaç, sınıfları birbirinden ayıran ve sınıf örneklerinin hiper-düzleme olan uzaklıklarının maksimum olduğu bir hiper-düzlemin belirlenmesidir.

2.4.2.10 Gradient Boosting Regressor (GBM)

Makine öğrenmesine ait bu yöntem hem sınıflandırma hem de regresyon problemlerinin çözümünde kullanılan bir algoritmadır. Bu yöntemin yaklaşımı, tek bir karar ağacının yeteri kadar kuvvetli olmadığı yönündedir. Bir karar ağacı yöntemi uygulanır. Ortaya çıkan hata ile yeni bir karar ağacı oluşturulur. İşlem hata minimuma ininceye kadar devam ettirilir.

2.5 Makine Öğrenmesinin Avantajları

Makine öğrenmesinin birden fazla uygulama alanı vardır. Makine öğrenmesi projeleriyle elde edilen en önemli avantajlar şunlardır:

- i. İçgörülerini ortaya çıkarma; Makine öğrenmesi, hem yapılandırılmış hem de yapılandırılmamış verilerde desenin tanımlanmasında yardımcı olmaktadır.
- ii. Veri bütünlüğünü; Makine öğrenmesi ,veri madenciliği konusunda oldukça iyi bir performans göstermektedir.

- iii. Risk azaltma; Dolandırıcılık sürekli olarak değişmektedir. Makine öğrenmesi, dolandırıcılık denemeleri başarılı olmadan önce bunları yakalamaktadır.
- iv. Müşteri davranışlarını öngörme; Makine öğrenmesi, desen ve davranışların belirlenmesinde yardımcı olmaktadır. Müşteriyle ilgili verileri araştırıp ürün önerilerinin iyileştirilmesine olanak sağlar. (Azure, n.d.)

2.6 Makine Öğrenmesinin Kullanım Alanları

Makine öğrenmesi teknoloji ilerledikçe her geçen gün kullanım alanı daha da artmaktadır.

- i. Görüntü tarama; Makine öğreniminin en yaygın kullanım alanlarından biridir. Makine öğrenme algoritmalarının yardımı ile yüz ve fotoğraf algılamasının yanında el yazısını da algılayabilmektedir.
- ii. Siber güvenlik; Makine öğrenimi algoritmaları, siber işlemleri takip etmektedir. Yasal olmayan herhangi bir işlemi ayırt eder, engeller veya bildirmektedir.
- iii. Duygu Analizi; Doğal Dil İşleme (NLP) adı verilen bu süreçte, makine öğrenimi modeli yorum vb gibi müşteri verilerini olumlu ve olumsuz olarak analiz etmektedir. (Kabacan, 2020)
- iv. Satış ve Pazarlama; Satış fırsatlarını öngörmeye ve görevleri otomatikleştirmekte iyi bir performans göstermektedir.
- v. Finans; Yapay öğrenme ile çalışan yazılımlarla yatırım yönetimi kararlarını güçlendirmektedir.

2.7 Makine Öğrenmesinde Model Seçimi

Regresyon algoritmaları ile modelleme yapılan çalışmalarda nümerik bir değer çıktısı beklenmektedir. Sınıflandırma metotları kullanılarak kurulan modelde ise çıktı kategorik değişken olmalıdır.

- Regresyon için açıklanabilirlik oranı(R^2) ve RMSE değeri önemlidir.
- Sınıflandırma için doğru sınıflandırma oranı önemlidir.

- Yani tahmin ile gerçek değer arasında farkları en küçük olan model seçilmelidir.

2.8 Model Başarısı Değerlendirme Yöntemleri

MSE değeri: Regresyon metotlarından biri ile modelleme yapılacaksa bu doğrulama yöntemi kullanılır. Gerçek değer y_i ve tahmin edilen $\hat{y}_i$ değer farkları karelerinin toplamının ortalamasıdır. Ortalama Kare Hatası, öngörülen sonuçların gerçek sayıdan ne kadar farklı olduğuna dair bir sayı vermektedir.

$$MSE = \frac{\sum(\hat{y}_i - y_i)^2}{n}$$

n =Gözlem sayısıdır.

y_i =gerçek değerlerdir.

$\hat{y}_i$ =tahmin edilen değerlerdir.

RMSE değeri: Bir veri kümesindeki öngörülen $\hat{y}_i$ değerleri ile gerçek y_i değerleri arasındaki ortalama kare farkının karekökünü hesaplamaktadır. RMSE düşükse, model veri kümesine o kadar iyi uymaktadır. RMSE değeri , MSE değerinin kareköküdür. (Zach, 2021)

$$RMSE = \sqrt{\frac{\sum(\hat{y}_i - y_i)^2}{n}}$$

R^2 değeri: Verilerin regresyon hattına ne kadar yakın olduğunun istatistiksel bir metriğidir. Belirleme katsayısı olarak da bilinmektedir. Özetlemek gerekirse R-kare, doğrusal regresyon metodları için uygunluk ölçüsüdür.

$$R^2 = 1 - \frac{\text{Explained Variation}}{\text{Total Variation}}$$

Explained Variation: Açıklanan Varyasyon

Totaş Variation: Toplam Varyasyon

Adjusted R^2 : Ayarlanmış R-kare, her bir bağımsız değişkeni telafi etmektedir. Verilen her değişken, algoritmayı mümkün olanın üstünde geliştirdiğinde artar.

$$Adjusted R^2 = 1 - (1 - R^2) \frac{n - 1}{n - p - 1}$$

p= Bağımsız değişken sayısıdır.

n= Gözlem sayısıdır.

Karışıklık matrisi: Makine öğrenimi, sınıflandırma problemi için bir performans metriğidir. Tahmin edilen değerleri ve gerçek değerleri içeren bir tablodur. Verideki var olan durum ile sınıflama algoritmasının, doğru ve yanlış tahminlerinin sayısını göstermektedir.

		Var olan Durum	
		Pozitif Durumlar	Negatif Durumlar
Tahmin	Pozitif	TP	FP
	Negatif	FN	TN

Şekil 15. Karışıklık matrisinin tablosal gösterimi

TP:True Pozitif

FN:False Negatif

FP:False Pozitif

TN:True Negatif

Accuracy(Doğruluk): Modelde doğru olarak yapılan öngörülerim tüm tahminlere oranıdır.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FN + FP)}$$

Recall(Anma): Pozitif durumların ne kadar başarılı öngörüleceğini göstermektedir.

$$Recall = \frac{TP}{(TP + FN)}$$

Precision(Kesinlik): Pozitif olarak öngörülen bir durumdaki başarıyı göstermektedir.

$$Precision = \frac{TP}{(TP + FP)}$$

F-Measure: Recall ve Precision'ın harmonik ortalamasıdır. (Uzun, 2022)

$$F - Measure = \frac{2 * Precision * Recall}{(Precision + Recall)}$$

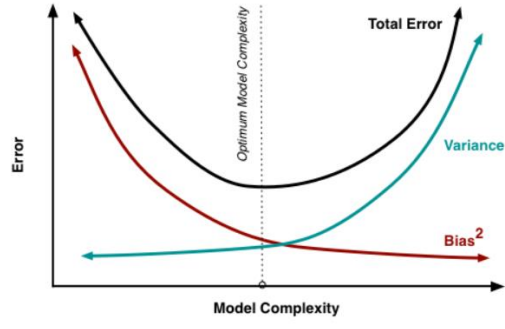
2.9 Varyans ve Yanlılık Kavramları

Yanlılık (bias), algoritmanın ortalama öngörülen değerler ile gerçek değer arasındaki farka denmektedir. Algoritmada eğitilen model yüksek biasa sahipse eğitim verisini daha basite indirgemiş olur. Bu da eğitim ve test verisinin yüksek hatasına sebep olmaktadır.

Yanlılık, algoritmanın problemine çözümünü içermediğini göstermektedir. Algoritmanın zayıf kaldığı bu durum eksik öğrenmedir. Yüksek yanlılığa sahip bir algoritmanın, eğitim verisini eksik öğrenme ihtimali daha fazladır.

Varyans (variance), belirli bir veri noktası için algoritma öngörüsünün değişkenliğini söyleyen bir değerdir. Algoritma yüksek varyansa sahipse öğrenme setinde genelleme yapabilecek bir örüntü bulmaz. Bunun yerine veri setindeki gözlemleri ezberlemeye başlar. Dolayısıyla daha önce görmediği verilerde iyi bir genelleme yapamaz. Bu da eğitim verisinde düşük hata ve test verisinde de yüksek hata elde edilmesine sebep olur.

Bir modelin karmaşıklığını artırmak, varyansını artırır ve yanlılığı azaltmaktadır. Aksine, bir modelin karmaşıklığını azaltmak ise, yanlılığı artırır ve varyansı azaltmaktadır. Ancak unutulmamalıdır ki, her iki model de aynı hataya sahip olabilir. (Koptur, 2017)



Şekil 16. Varyans Ve Yanlılığın Grafikselsel Gösterimi (Koptur, 2017)

ÜÇÜNCÜ BÖLÜM

3. MAKİNE ÖĞRENMESİ YÖNTEMİ İLE EĞİTİM BAŞARISININ TAHMİNİ MODELİ

Bu bölümde, eğitim başarısının tahminine ait toplanan veriler makine öğrenmesi yöntemleriyle modellenerek en yüksek model başarısını sağlayan yöntem seçilmiştir. Bu uygulama ile elde edilen veri seti eğitim ve test verisi olarak ayrılmıştır. Çoklu Lineer Regresyon , Random Forest , K-NN , Bagged Trees Regresyon, GBM, Karar Ağaçları algoritmaları ile model eğitilip performansları karşılaştırılmıştır.

Ayrıca uygulamanın görsellik açısından zengin görünmesi amacıyla HTML ve CSS kullanılmıştır. Python'a ait Flask kütüphanesiyle modelin eş zamanlı web üzerinde çalışması için bir uygulama arayüzü tasarlanmıştır.

Model ile ilgili analizlerin yapılacağı ortamların kurulumu yapılmıştır. Bu aşamada veri manipülasyonu, veri görselleştirme , veriye ait hesaplamalar, modelin algoritmasının oluşturulması ve modelin performans değerlendirmesi işlemlerini gerçekleştirmek için Jupyter Notebook editörü kullanılmıştır.

3.1. Veri Setinin Hazırlanması

Bu çalışmadaki veriler ilkokul 8.sınıf öğrencilerine hazırlanan 13 soruluk anketle toplanmıştır. Toplamda 519 veri elde edilmiştir. Veriler ilkokul öğrencilerine öğretmenleri aracılığıyla ve sosyal medya kaynaklarından ulaştırılmıştır. Anket soruları Ekler bölümünde verilmiştir.

3.1.1 Öznitelik Bilgileri Ve Grafikselleştirme

Anket cevapları özenle kontrol edilerek excele aktarılmıştır. Anket sorularının içerisinde eksik veri analizi yapılmıştır. Daha sonrasında oluşturulan excel dosyası Jupyter Notebook ortamına “read.csv” komutuyla ve .csv uzantısıyla alınmıştır. Veri setindeki değişkenler ve ilk 10 verinin bulunduğu örnek tablo , Tablo 1’ de gösterilmiştir.

Tablo 1. Veri Setinin Detayları

Sex	K	K	K	Er	Er	K	Er	K	K	Er
Sibling	E	E	H	E	E	H	E	H	E	E
Coaching	E	H	H	E	H	E	H	H	H	H
HaveRoom	E	H	E	H	E	E	E	E	H	H
FamSep	H	H	H	E	H	E	H	E	H	H
FamilyAlive	E	E	E	E	E	E	E	E	E	E
StepFM	H	H	H	E	H	H	H	E	H	H
WStudying	E	H	E	H	H	E	E	E	E	H
Studying	2	1	3	4	3	3	5	2	4	2
Sleeping	9	10	8	9	6	9	7	8	8	7
Playing	1	1	0	3	5	0	1	1	3	4
Income	9000	8900	11000	9200	9800	8500	11000	7000	10000	6000
TurkNote	85	67	94	50	72	87	80	76	85	56

Anket soruları öğrencilere Türkçe sorulmuştur. Ancak öznitelik belirlerken tercih edilen ingilizce kısaltmalar olmuştur. Yapılan literatür taramasında tablolarda yer alan değişken isimlerinin çoğunlukla ingilizce belirtildiği gözlemlenmiştir. Bu da daha akılda kalıcı ve etkileyici görünmesine sebep olmaktadır.

Tablo 2’de anket sorularına karşılık gelen İngilizce kısaltmalar gösterilmiştir.

Tablo 2. Anket Sorularına Karşılık Gelen Değişken İsimleri

Cinsiyetiniz	Sex
Kardeşiniz var mı ?	Sibling
Özel eğitim alıyor musunuz?	Coaching
Kendine ait bir odanız var mı?	HaveRoom
Anneniz ve babanız ayrı mı?	FamSep
Anneniz veya babanız hayatta mı?	FamilyAlive
Üvey anneniz veya babanız var mı?	StepFM
Hafta sonları ders çalışıyor musunuz?	WStudying
Günde kaç saat ders çalışıyorsunuz?	Studying
Günde kaç saat ders uyuyorsunuz?	Sleeping
Ortalama kaç saat oyun oynuyorsunuz?	Playing
Aile geliriniz nedir?	Income
Türkçe Notunuz nedir ?	TurkNote

Veri seti içerisinde 8 adet kategorik değişken ve 5 adet sürekli (sayısal) değişken bulunmaktadır. Tablo 3 'te değişkenlerin tipleri belirtilmiştir.

Tablo 3. Veri Seti İçerisindeki Değişkenlerin Tiplerinin Gösterimi

Sex	object
Sibling	object
Coaching	object
HaveRoom	object
FamilySepareted	object
FamilyAlive	object
StepFatherMother	object
WeekendStudying	object
Studying	int64
Sleeping	int64
Playing	int64
Income	int64
TurkishNote	int64

Veri setindeki değişkenler yukarıda bahsedildiği üzere İngilizce diliyle kısaltılmıştır.

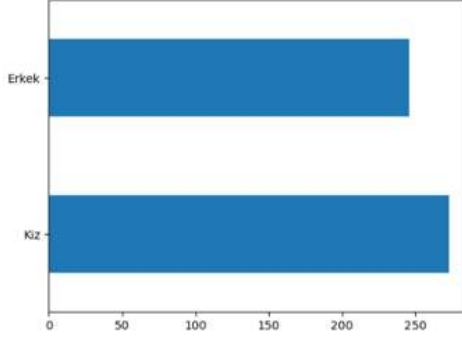
Anket sorularını cevaplayan öğrenciler, “value_counts()” metoduyla veri seti içerisinde belirlenmiştir.

Kategorik değişkenler Matplotlib kütüphanesi kullanılarak farklı parametrelerle görselleştirilmiştir.

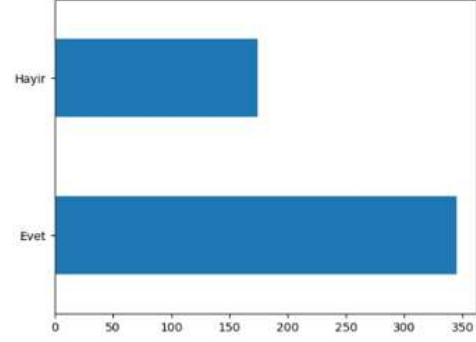
Şekil 17’de “value_counts().plot.barh()” parametresi kullanılmıştır. Veri setinden örnek olarak cinsiyet (Sex) ve kardeşe sahip olma (Sibling) kategorik değişkenleri grafiksel olarak betimlenmiştir. Bu grafiklere göre anket sorularını cevaplayan 519 öğrenciden 273 tanesi kız öğrenci, 246 tanesi erkek öğrencidir. Ve 343 öğrenci kardeşe sahip, 174 öğrencinin ise kardeşi olmadığı sonucuna varılmıştır. Böylelikle anketi cevaplayan daha fazla kız öğrenci olduğu ve kardeşe sahip olma öğrencilerin çoğunlukta olduğu gözlemlenmiştir.

```
data["Sibling"].value_counts().plot.barh();
kat_df["Sibling"].value_counts()
```

```
Out[14]: Kız    273
         Erkek   246
         Name: Sex, dtype: int64
```



```
Evet    345
Hayır   174
         Name: Sibling, dtype: int64
```



Şekil 17. Değişkenlerin value_counts().plot.barh() Parametresi İle Gösterimi

Şekil 18’de Seaborn kütüphanesine ait catplot parametresi kullanılarak özel eğitim alma durumu (Coaching), ailenin ayrılık durumu (FamilySeperated), ailenin hayatta olma durumu (FamilyAlive) ve üvey anne babaya sahip durumu (StepMotherFather) değişkenleri “sns.catplot()” parametresiyle görsel olarak Türkçe notuyla ilişkilendirilmiştir. Görseller göz önünde bulundurulduğunda;

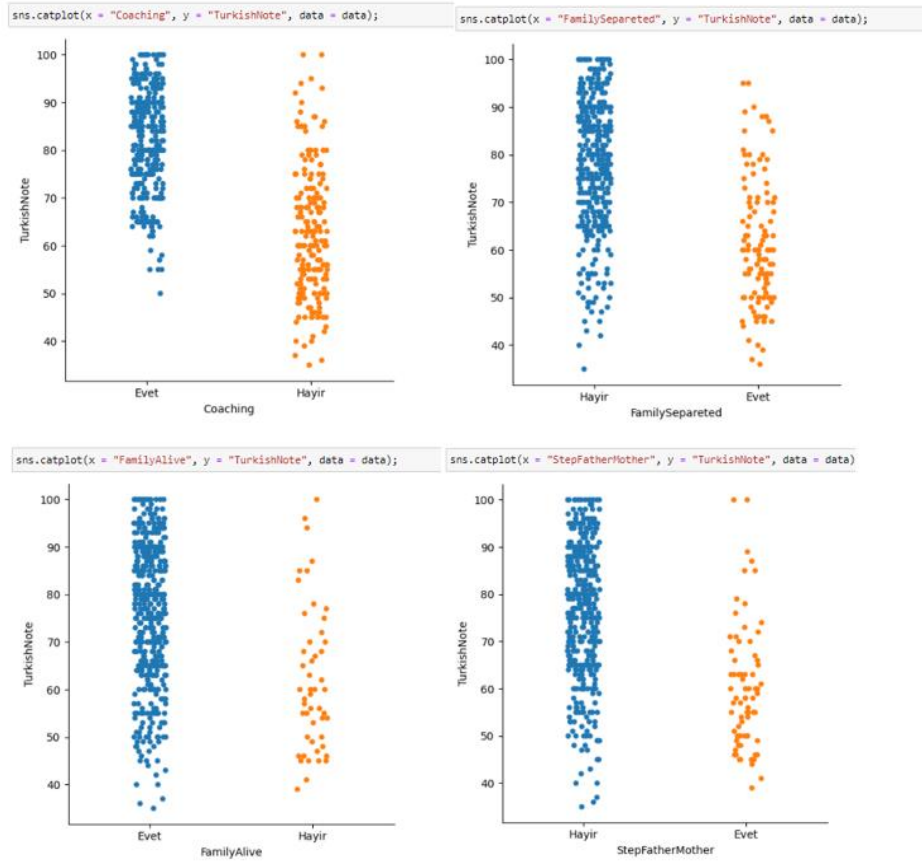
i)Veri seti içerisinde ağırlıklı olarak özel eğitim desteği alan öğrencilerin notlarının 70 ile 100 aralığında olduğu ,

ii)Ailesi ayrı olmayan öğrencilerin Türkçe notunun 60 ile 98 arasında olduğu,

iii)Ailesi hayatta olan öğrencilerin notlarının 50 ile 100 arasında yoğunlukta olduğu,

iv)Üvey anne babaya sahip olmayan öğrencilerin 60 ile 100 arasında not aldığı,

Sonuçlarına ulaşılmıştır.



Şekil 18. Catplot Parametresi İle Kategorik Değişkenlerin Gösterilmesi

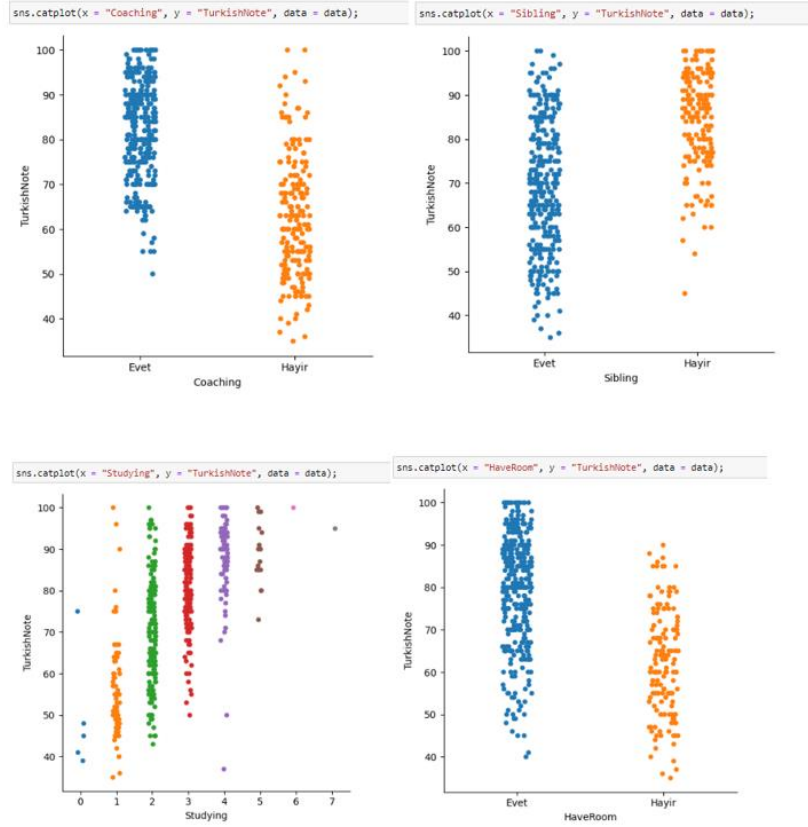
Şekil 19’da Odaya sahip olma (HaveRoom), özel eğitim alma (Coaching), kardeşe sahip olma (Sibling), ders çalışma (Studying) değişkenlerinin catplot parametresi ile görsel olarak betimlenmesi gösterilmiştir. Görsel yorumlandığında;

i)Özel eğitim alan öğrencilerin not aralığının 65 ile 100 , almayan öğrencilerin 45 ile 75 arasında olduğu,

ii)Kardeşe sahip öğrencilerin ağırlıklı olarak 45 ile 90, olmayan öğrencilerin 75 ile 100 arasında ağırlıklı olduğu,

iii)Genellikle 2,3,4 saat ders çalışan öğrencilerin notlarının 55 ile 100 arasında olduğu,

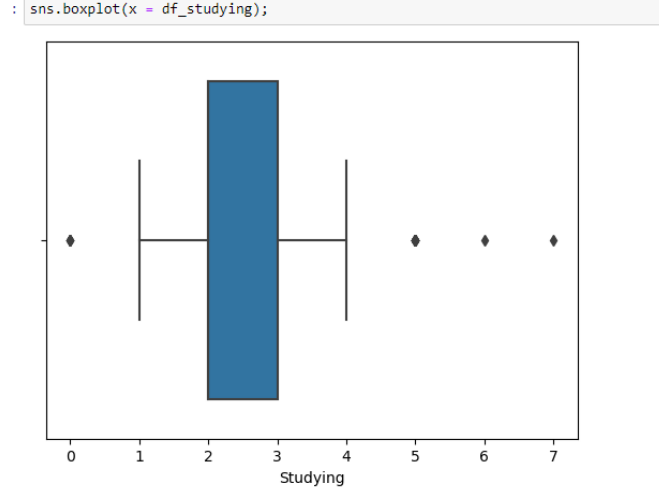
iv)kendi odasına sahip öğrencilerin notlarının 65 ile 100 , olmayan öğrencilerin 50 ile 75 arasında olduğu sonucuna varılmıştır.



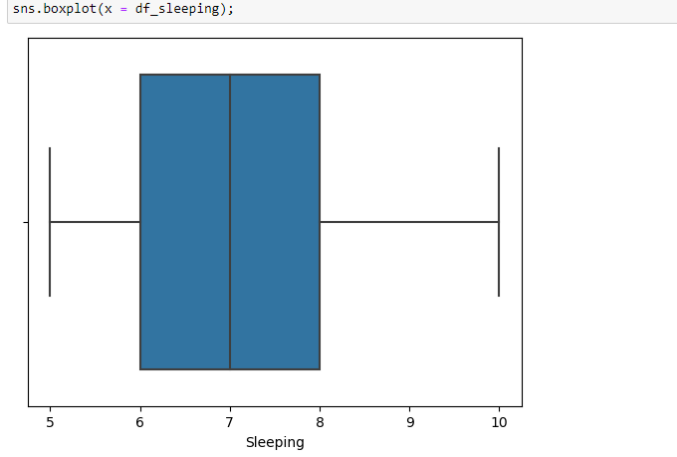
Şekil 19. Kategorik Değişkenlerin Catplot Parametresiyle Görselleştirilmesi

3.1.2 Özniteliklerin Analizi

Değişkenler Boxplot grafiği ile ayrı ayrı gösterilmiştir. Boxplot(Kutu Grafiği) veri çeyreklerini (veya yüzdeleri) ve ortalamaları görüntüleyerek sayısal verilerin ve değişkenliğin görsel olarak dağılımını göstermek için kullanılır. Veri analizinde sıklıkla kullanılan bir grafik türüdür. Boxplot ile değişkenlerin gösterilmesi veri setini etkileyecek aykırı gözlemlerin ortaya çıkarılması amacıyla yapılmıştır. Bir boxplot grafiğinde simetri yok ise o değişkende aykırı gözleme rastlanır. Şekil 20' de veri setinde aykırı gözleme rastlanmış bir değişkenin boxplot grafiği ile gösterimi yapılmıştır. Şekil 21'de ise simetri bir yapı gösteren, aykırı gözleme sahip olmayan bir değişkenin boxplot grafiği gösterilmiştir. Boxplot grafiğinin gösterimi için Python'ın Seaborn kütüphanesinden yararlanılmıştır.



Şekil 20. Aykırı Gözleme Sahip Değişkenin Boxplot Grafiği



Şekil 21. Aykırı Gözleme Sahip Olmayan Değişkenin Boxlot Grafiği

Yapılan boxplot grafiği analizi sonucunda Income (gelir) ve Studying (Ders Çalışma) değişkenlerinde aykırı gözleme rastlandığı için baskılama yöntemi ile aykırı gözlemler alt sınıra baskılanmıştır. Aykırı gözlemlerin belirlenerek bu yöntemin uygulanmasıyla modelin başarı performansı 0,01 oranında artmıştır.

Veri seti içersindeki kategorik değişkenleri sayısal değişkenlere dönüştürebilmek için Label Encoder yöntemi kullanılmıştır. Evet- Hayır ve Erkek-Kız kategorik değişkenleri bu yöntemle 0-1 olarak dönüşüme uğramıştır. Şekil 22’de kategorik değişkenlerin sayısallaştırılmasına ilişkin görsel verilmiştir.

```

: for column in data.columns[1:8]:
    data[column] = labelEncoder.fit_transform(data[column])
data
#kız=1 , erkek=0 , evet=0 , hayır =1

```

	Sex	Sibling	Coaching	HaveRoom	FamilySeparated	FamilyAlive	StepFatherMother	WeekendStudying	Studying	Sleeping	Playing	Income	TurkishNote
0	1	0	0	0	1	0	1	0	2.0	9	1	9000	85
1	1	0	1	1	1	0	1	1	1.0	10	1	8900	67
2	1	1	1	0	1	0	1	0	3.0	8	0	11000	94
3	0	0	0	1	0	0	0	1	4.0	9	3	9200	50
4	0	0	1	0	1	0	1	1	3.0	6	5	9800	72
...	...	...	...	...	...	...	...	...	...	...	...	...	...
514	0	0	0	0	1	0	1	1	2.0	9	3	12000	63
515	0	0	1	1	0	0	0	1	1.0	7	3	9000	50
516	1	1	0	0	1	0	1	0	3.0	7	2	10900	74
517	0	0	1	1	0	0	0	1	1.0	8	3	9000	44
518	0	1	1	0	0	1	0	1	1.0	8	4	9000	57

519 rows x 13 columns

Şekil 22. Kategorik Değişkenlerin Label Encoder İle Sayısallaştırılması

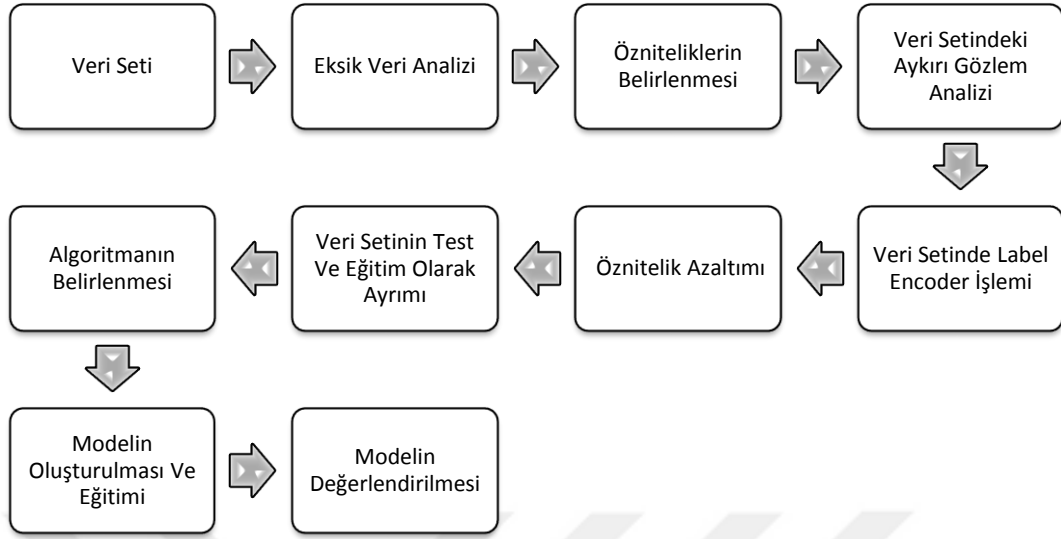
Türkçe Dersi notu bağımlı değişkeniyle sayısal değişkenlerin arasındaki korelasyon hesaplanmıştır ve negative yönde etki eden 2 X değişkeni (Playing, WeekendStudying) veri setinden çıkartılmıştır. Bu sonuca sadece korelasyon hesaplama yöntemiyle ulaşılmamıştır. Genel olarak değişken sayısının artması modelin performansını kötü etkileyebileceğinden model kullanılan algoritmalar ile %80 oranında veriyle eğitildikten sonra defalarca test edilmiştir. Ve korelasyon değerine ek olarak yapılan test gözlemlerince bu 2 değişken veri setinden çıkarıldıktan sonra çok daha iyi sonuçlar ürettiği ortaya konulmuştur.

3.2. Modelin Oluşturulması

Değişken sayısının çok olması modelin başarı performansını etkileyen en önemli faktörlerden biridir. Modelde toplamda 13 değişken bulunmaktadır. Türkçe notu değişkeni y bağımlı değişkeni, 12 değişken ise X değişkeni olarak belirlenmiştir. Model oluşturulduktan sonra veri seti içerisinde X değişkenleri y değişkenini tahmin edecektir.

Çalışma kapsamında sürekli değişken tahmin edileceğinden regresyon yöntemleri ve modeli oluşturmak için önceden söz edilen 6 ayrı algoritma kullanılmıştır.

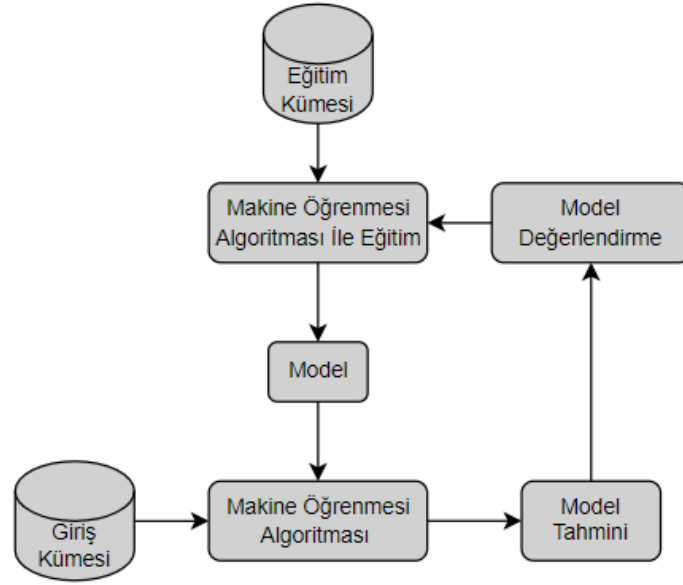
Makine öğrenmesi modeli oluşturulurken belirli aşamalardan geçmelidir. Bu aşamalar Şekil 23’ de gösterilmiştir.



Şekil 23. Makine Öğrenmesi Yöntemlerinin Uygulanması Aşamaları

Algoritmaların modellenmesi aşamasında y değişkeni (Türkçe notu) veri setinden çıkartılmıştır. “ X_{train} , X_{test} , y_{train} , y_{test} ” parametreleri kullanılarak değişkenler test ve eğitim (train) olarak ayrılmıştır. Veri setinin % 20’si test için %80’i ise eğitim için belirlenmiştir. “Random_state” parametresi 42 olarak belirtilmiştir. Random_state parametresi kod her çalıştığında veri kümesini aynı şekilde bölmektedir

Modeller; LinearRegression, KNeighborsRegressor, RandomForestRegressor, BaggingRegressor, DecisionTreeRegressor, SVR fonksiyonlarıyla oluşturulmuştur. Makine öğrenmesindeki temel süreç Şekil 24’te açıklanmıştır.



Şekil 24. Makine Öğrenmesinin Temel Süreci

3.3 Çalışmada Kullanılan Algoritmaların Analizi

Çalışmada daha önceki bölümlerde belirtildiği üzere 6 adet denetimli öğrenme türü olan Random Forest , Multiple Linear Regression, Decision Tree, Bagged Trees, Gradient Boosting Machines, K-NN algoritmaları kullanılmıştır.

Bu bölümde, kullanılan algoritmaların analizi ve değerlendirilmesi yapılacaktır.

3.3.1 Random Forest Algoritması

Random Forest modeli oluşturulurken RandomForestRegressor kütüphanesi import edilmiştir. Bu kütüphane içerisinde RandomForestRegressor() fonksiyonu kullanılarak “rf_model” isimli model oluşturulmuştur. Random State parametresi 42 değeri verilerek veri seti eşit parçaya bölünmüştür. X_train ve y_train parametreleri fit fonksiyonuyla sisteme verilmiştir.

Model eğitildikten sonra “model.Score(X,y)” fonksiyonu ile başarı performans yüzdesi %88 olarak belirlenmiştir. Bu oran, tahmin kar sonuçların doğruluğunun daha yüksek olduğu gözlemlenebilir. Model, %98 R-kare oranına sahiptir. R-kare değeri modelin regresyon hattına yakınlığıdır. Modelin

uyumluluğu anlamına da gelir. Yani Random Forest algoritması kullanılan model %98 oranında yeni girdilere uyum sağlayacak şekilde eğitilmiştir.

3.3.2 K-NN Algoritması

K-NN algoritmasında, KNeighborsRegressor kütüphanesinden faydalanılmıştır. Model “knn_model” ismi ile oluşturulmuştur. Algoritmadaki en iyi komşu sayısı, “n_neighbors” parametresi kullanılarak 5 olarak belirlenmiştir. Modelde algoritma metriği olarak euclidean(öglid) kullanılmıştır. Bu metriğe de , “effective_metric_” parametresiyle erişilmiştir.

Modelin başarı performans yüzdesi %44 olarak model.Score parametresiyle değerlendirilmiştir. Bu başarı değeri göz önünde bulundurulduğunda düşük tahminlik değere sahiptir. Algoritmanın regresyon hattına yakınlık oranı (modelin uyumluluğu, R-kare) %29 olarak belirlenmiştir. Başarı oranı gibi , yeni girdilerin modele uyumu oldukça düşüktür.

3.3.3 Decision Tree Algoritması

Decision Tree yada Karar Ağaçları olarak billinen bu algoritmada CART metodu seçilmiştir. DecisionTreeRegressor fonksiyonu kullanılarak model oluşturulmuştur. Model X_train ve y_train parametreleri ile fit edilmiştir.

Model performansı değerlendirildiğinde %83’lük bir başarıya sahip olsada model uyumluluk oranı (R-kare) %34 olarak hesaplanmıştır. Model yüksek bir performans başarısına sahip olabilir ancak verilen girdilerin modele uyumluluk oranı yüzdesel olarak düşüktür. Yani regresyon hattından uzaklaşmıştır. Bu da modelin şaşırtıcı tahminler üretebilmesine neden olabilmektedir.

3.3.4 Bagged Trees Algoritması

Model BaggingRegressor fonksiyonu kullanılarak, “bag_model” ismi ile oluşturulmuştur. Algoritma diğer tüm yöntemlerdeki gibi %20’lik oranda veri setiyle eğitilmiştir.

Model başarısı yüzdesi %80 olarak üretilmiştir. Random Forest ve Decision Tree algoritmasından sonra en yüksek başarıya sahip yöntemdir. Modelin R-kare değeri %61 olarak hesaplanmıştır. Modelin uyumluluğu yüksek

olmasada diğer algoritma başarılarına göre regresyon hattına Random Forest algoritmasından sonra en yakın modeldir.

3.3.5 Çoklu Lineer Regresyon Algoritması

Çoklu Lineer Regresyon modelinde, LinearRegression fonksiyonu kullanılmıştır ve model “model” ismiyle adlandırılmıştır.

Algoritmanın formülüne göre ;

$$y = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n + \epsilon$$

β_0 sabiti model.intercept_ parametresinin çağrılmasıyla sistem tarafından 47.38 olarak hesaplanmıştır.

Diğer X bağımsız değişkenlerine ait sabitler değerler ise ($\beta_1, \dots, \beta_{10}$) model.coef_ parametresiyle;

```
array([ 2.36984478e+00,  5.12852936e+00, -7.74608201e+00, -6.44312847e+00,
        5.48366296e+00, -8.87890590e-01, -1.82986986e+00,  4.93936336e+00,
        6.75501982e-01,  7.67826758e-04])
```

olarak sistem tarafından hesaplanmıştır.

Bu sabitler her bir X bağımsız değişkeni ile çarpılıp toplandıktan sonra y bağımlı değişkenini tahmin etmeye çalışmaktadır. Modelin başarı performansı ölçülmüştür ve %67 olarak belirlenmiştir. R-kare , modelin uyumluluğu değeri %66 olarak değerlendirilmiştir. Bu sonuç göz önünde bulundurulduğunda, çok yüksek olmayan ancak orta dereceli tahminkar sonuçlar elde edilebilecek bir model ortaya çıkarmıştır.

3.3.6 Gradient Boosting Machines Algoritması

Bu denetimli öğrenme algoritmasına, GradientBoostingRegressor kütüphanesi import edilerek “gbm_model” ismiyle model oluşturulmuştur.

Modelin başarı performansı %77 olarak değerlendirilmiştir. Ancak R-kare değeri , Decision Tree algoritması gibi düşük bir değere sahiptir. Sistem tarafından %33 olarak hesaplanmıştır. Model iyi bir başarı yüzdesine sahip olsa da bazen yanıltıcı çıktılar verebileceği öngörülmektedir.

3.4 Algoritmaların Değerlendirilmesi

Tez çalışmasında kullanılan 6 adet algoritmanın; model başarıları, R-kare değerleri (model uyumluluğu) , RMSE ve MSE metrikleri karşılaştırılarak en tahminkar sonuç verecek algoritma seçilmiştir.

Algoritmaların model başarıları Tablo 4 'te gösterilmiştir.

Tablo 4. Modellerin Başarı Yüzdeleri

	K-NN	MLR	RF	BTR	DT	GBM
Model Score(X,y)	0,44	0,67	0,88	0,80	0,83	0,77

Tablo 4 göz önünde bulundurulduğunda en yüksek performansı gösteren model Random Forest olmuştur.

Modelin başarısına göre sıralandırılması ;

RF>DT>GBM>BTR>MLR>K-NN

Bu oranlara göre en yüksek performansı gösteren Random Forest algoritması, en düşük performansa ait algoritma K-NN olmuştur.

Algoritmaların RMSE, MSE değerleri Tablo 5 'te gösterilmiştir.

Tablo 5. Modelin Performans Metrikleri

	K-NN	RF	MLR	BTR	DT	GBM
RMSE	12,77	8,56	9,06	9,53	11,34	10,29
MSE	61,48	61,46	82,16	61,46	61,55	75,46

RMSE değeri düşük olan algoritmanın performansı yüksektir. RMSE değerine göre performans sıralaması ;

RF>MLR >BTR>GBM>DT>K-NN

RMSE metrik değeri göz önünde bulundurulduğunda Random Forest(RF) algoritması en iyi performansı gösteren model, K-NN algoritması en düşük performansı gösteren model olarak belirlenmiştir.

Tablo 6’da modellere ait R-kare deęerleri verilmiřtir.

Tablo 6. Modellerin R-kare Deęerleri

	K-NN	RF	MLR	BTR	DT	GBM
R-kare	0.29	0.98	0.66	0.61	0.34	0.33

R-kare deęerlerine gre algoritmaların sıralaması;

RF>MLR>BTR>DT>GBM>K-NN

Model uyumluluklarına bakıldığında RF algoritması regresyon hattına en yakın yntemdir. K-NN algoritması ise 0.29 deęeri ile regresyon hattına en uzak yntem olmuřtur.

3.5 Makine ęrenmesi Yntemlerinin Optimizasyonu Ve Model Seęimi

Decision Tree ve GBM algoritmalarının model bařarısı yksektir. Ancak R-kare deęerinin dřk olmasından kaynaklı yeni girdiler sisteme verildiğinde yanıtıcı çıktı retebileceğinden bu algoritmalar tercih edilmemiřtir.

K-NN algoritması dřk bařarı ve regresyon hattına sahiptir. İyi sonu ıktısı alınması olduka gtr. Algoritma bu sebeple tercih edilmemiřtir.

BTR ve MLR algoritmalarının bařarı yzdesine bakıldığında yksek olsada, model uyumluluęu en yksek algoritma deęildir. Daha iyi tahmin retebilecek yeni girdilerle uyum saęlayabilecek bir algoritma modeli sonucuna ulařıldığı iin BTR ve MLR yntemleri kullanılmamıřtır.

Yapılan analizler ve deęerlendirmeler sonucunda %88 bařarı oranı ,%98 model uyumluluęu ve en dřk RMSE deęerine sahip olduęu iin en dzgn tahminkar sonular retebilecek algoritma Random Forest’tır. Yeni girdilerle modelin test ařamasında yanıtıcı sonular olmaması iin bu tez alıřmasında RF yntemi seilmiřtir.

Kullanılan dięer 5 algoritmanın RF modeli kadar bařarılı olmamasının sebebi veri setinde kullanılan gzlem sayısının az olmasından kaynaklı olabileceęi dřnlmektedir. Deęiřken sayısının artması modelin bařarısını

negatif yönde etkileyebilir. Değişken sayısına bağlı olarakta başarısızlık söz konusu olabilir.

Bu algoritmalarının seçilmesinin nedeni, yapılan literatür taraması sonucunda en çok tercih edilen ve model başarı performansları yüksek olacağı öngörülen yöntemler olmasından kaynaklıdır. GBM algoritması fazla tercih edilen bir makine öğrenmesi algoritması değildir ancak literatüre yeni bir algoritma tanımı ve kullanımı örneklendirmek için tez çalışmasında bu algoritmayada yer verilmiştir. GBM yöntemi, literatürde en çok kullanılan algoritmalar kadar başarı (%77) göstermiştir. Ancak R-kare (model uyumluluk) değerinin düşük olmasından kaynaklı model seçiminde elenmiştir. Tez çalışmasında, beklenen çıktı değeri sürekli değişken olduğundan uygulamada modele doğrusal ve doğrusal olmayan regresyon modelleri uygulanmıştır.

3.6 Random Forest İle Modelin Test Edilmesi

Model, Random Forest algoritması kullanılarak daha önce hiç görmediği verilerle test edilmiştir. Şekil 25'te Random Forest ile test edilen test verisinin çıktısı verilmiştir. Label Encoder işleminden sonra kategorik değişkenlerin sayısallaştırıldığı için 0 ve 1'e karşılık gelen kategorik değişkenler belirtilmiştir.

Kız=1 , Erkek=0 , Evet=0 , Hayır=1

Tablo 7. Modelin Test Edildiği Örnek Veri

Öznitelikler	Test Değeri
Sex	Erkek
Sibling	Hayır
HaveRoom	Evet
Coaching	Evet
FamilySeperated	Hayır
FamilyAlive	Evet
StepMotherOrFather	Hayır
Studying	5 Saat
Sleeping	7 Saat
Income	10000TL

```
yeni_veri = [[0],[1],[0],[0],[1],[0],[1],[5],[7],[10000]]
yeni_veri = pd.DataFrame(yeni_veri).T
```

```
yeni_veri
```

	0	1	2	3	4	5	6	7	8	9	
	0	0	1	0	0	1	0	1	5	7	10000

```
rf_model.predict(yeni_veri)
```

```
array([86.95])
```

Şekil 25. RF Algoritması İle Test Edilen Verinin Çıktısı

Farklı farklı senaryolarda model defalarca test edilmiştir. En performanslı sonucu üretmek için bütün değişkenler gözlemlenmiştir. Sonuç olarak Playing(Oyun) ve WeekendStudying(Haftasonu Ders Çalışma) değişkenleri modelin tahmin performansını negatif yönde etkilediğinden bu değişkenler mevcut veri seti içersinden çıkarılmıştır. Model bu değişkenler çıkarıldıktan sonra tekrar test edilmiştir. Test sonucunda algoritma daha yüksek performanslı sonuç üretmiştir.

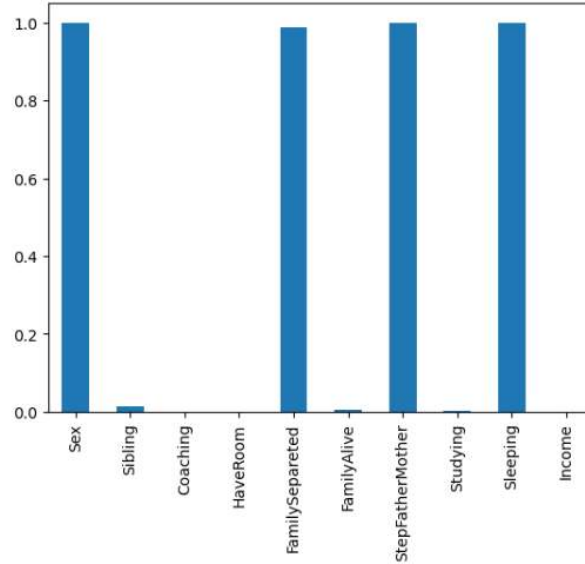
Sonuç olarak cinsiyeti erkek , kardeşi olmayan , özel eğitim alan , kendi odasına sahip, anne ve babası ayrı olmayan ,ailesi hayatta olan , üvey anne , babaya sahip olmayan, ortalama 5 saat ders çalışan , 7 saat uyuyan , gelir durumu orta-düşük olan öğrencinin notu 86.95 olabileceğine yönelik sistem bir tahminde bulunmuştur.

3.7 Feature Selection İle Öznitelik Seçimi

Değişken sayısının fazla olması modelin performansını etkileyebilecek etkenlerdendir. Bu tez çalışmasında, veri setindeki 13 değişken arasından en etkili öznitelikleri tespit edebilmek için feature selection işlemi yapılmıştır. Girdiler arasında kategorik değişken bulunduğu için ve çıktı değişkeni sürekli değişken olduğundan Chi-Kare yöntemi kullanılmıştır.

Chi-Kare: Değişkenler arasında bağımlılık olup olmadığını kontrol eder. Çıktı değişkeni sürekli dir. Değişkenlerin p-değeri 0.5'ten az ise yüksek önemli değişken olarak kabul edilir.

Değişkenlere Chi-Kare yöntemi uygulandıktan sonra $pvalue < 0.5$ olan değişkenler seçilmiştir. Seçilen değişkenler ; Sibling, Coaching, HaveRoom, FamilyAlive, Studying, Income'dır. Özniteliklerin pvalue değerleri Şekil 26'da gösterilmiştir.



Şekil 26. Değişkenlerin Pvalue Değerleri

Pvalue değeri 0,5 ten büyük olan değişkenler veri setinden çıkarılmıştır. Model güncel değişkenlerle tekrar eğitilmiştir. Tüm modellere aynı girdiler verilmiş ve sonuçlar karşılaştırılmıştır.

Evet=0 , Hayır=1

Tablo 8. Modellere Verilen Yeni Girdiler

Öznitelikler	Test Değeri
Sibling	1
Coaching	0
HaveRoom	0
FamilyAlive	0
Studying	4
Income	12000

Tablo 8’e göre kardeş sahip olmayan, özel eğitim alan, kendine ait odası olan, ailesi hayatta, 4 saat ders çalışan ve aile geliri 12000 olan bir öğrencinin alabileceği Türkçe Notu, 6 adet makine öğrenmesi algoritması tarafından tahmin edilmiş ve Tablo 9’da verilmiştir.

Tablo 9. Modellerin Tahmin Ettiği Türkçe Notu Çıktıları

Modeller	Türkçe Notu Tahmini
RF	94.74
DT	94.5
MLR	91.33
BTR	88.14
GBM	92.1
K-NN	84.1

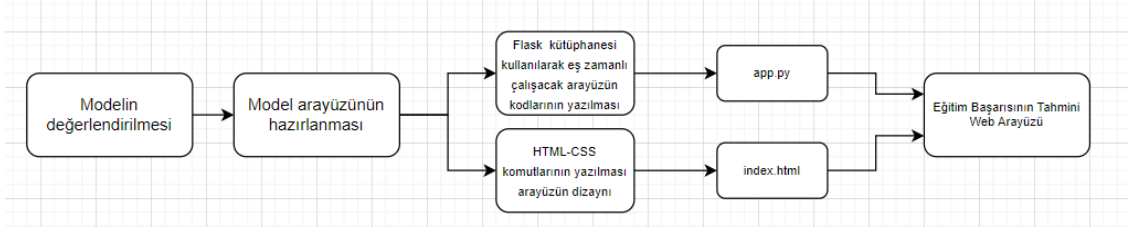
Girdiler göz önünde bulundurulduğunda, en tahminkar sonuç üreten algoritmalar RF, DT, GBM algoritmaları olmuştur. Tablo 4’te modellerin başarı performansı incelendiğinde algoritmalar beklenen sonuçları üretmiştir.

3.8 Web Arayüzünün Tasarlanması

Jupyter Notebook editörü kullanılarak modelin tüm manipülasyon ve istatistiksel işlemleri aşamaları gerçekleştirilmiştir. Çalışmada Python’a ait Flask kütüphanesi kullanılarak model ile eş zamanlı çalışmasını sağlayacak bir web arayüzü uygulaması oluşturulmuştur. Uygulamanın farklı bir dizinde çalışması için Python’a ait Pickle modülü kullanılmıştır. Jupyter Notebook editörü üzerinde modele, yazma yetkisi tanımlanarak “model.pkl” uzantılı dosya oluşturulmuştur.

HTML komutlarının yazılması için bir index.html dosyası oluşturulmuştur. Uygulamanın eş zamanlı çalışmasını sağlayacak kodlar için app.py uzantılı bir dosya oluşturulmuştur.

Arayüz tasarlanırken ve uygulamaya alınırken izlenen adımlar Şekil 27’de verilmiştir.



Şekil 27. Web Uygulamasında İzlenen Adımlar

Uygulamanın farklı bir ortamda çalışmasını sağlamak için projenin içinde olduğu dosya dizinine cmd ekranından ulaşılarak “python –m venv venv” komutuyla virtual environment yaratılmıştır. Bu environmentın aktif olarak uygulamada kullanılabilmesi için “activate.bat” komutuyla aktifleştirilmiştir.

Uygulama, projenin içerisine flask kütüphanesi “pip3 install flask” komutuyla eklenerek bu kütüphanenin tüm objelerine erişilecek hale getirilmiştir.

Arayüz tasarlanırken HTML komutlarını desteklemek için dizayn aşamasında Bootstrap kütüphanesinden faydalanılmıştır. Bootstrap içinde CSS kodlarını belli kalıplarda bulunduran bir kütüphanedir. Tasarım için Bootstrap kütüphanesinden form, label, footer komutları alınmıştır. Form yapısı, anket sorularından alınacak cevapları oluşturulan URL’e başarılı bir şekilde gönderilmesi için, label yapısı, kullanıcıdan alınacak inputları oluşturmak için, footer yapısı, sayfanın en altında Tahmini Gör butonunu ayrı bir yapıda tutmak için kullanılmıştır.

Her bir anket sorusu ayrı inputlarla kullanıcıdan alınmıştır. Oluşturulan html dosyası, app.py dosyasına render_template komutuyla bağlanmıştır.

App.py dosyasının, html ile eş zamanlı çalışabilmesi için index.html içerisinde her bir input için belirlenen value değerleri fonksiyonlarda belirtilmiştir. Bu işlem sonrasında doldurulan inputlar arka planda fonksiyonları aktif etmiştir.

Şekil 28’de html dosyasındaki kullanıcıdan alınacak inputların value değerlerinin fonksiyonlara uyarlanmasıyla ilgili örnek verilmiştir.

```
def setRadioValues(value):  
    if value == 'kiz':  
        return 1  
    else:  
        return 0  
  
def setContact(value):  
    if value == 'evet':  
        return 1  
    else:  
        return 0
```

Şekil 28. Value Değerlerinin Fonksiyonlara Uyarlanması

Verilen örnekte, cinsiyet inputuna girilen kız veya erkek olma durumuna göre fonksiyonda if koşulu kullanılmıştır. Label Encoder işleminde kız cinsiyeti için 1 değeri, erkek cinsiyeti için 0 değeri şeklinde bir encode dönüşümü yapılmıştır. Buna bağlı olarak çalışacak kod bloğunda cinsiyeti kız olan değerler için 1, diğer durumda yani erkek cinsiyete sahip olma durumunda 0 değerini dönecektir. Diğer kategorik değişken inputları evet veya hayır değerine sahiptir. Bu sebeple encode işleminde belirlenen hayır için 0, evet için 1 değeri if bloğunda value değerine bağlı olarak çalışacaktır.

Html dosyasındaki değişkenler app.py dosyası içerisindeki değişkenlerle fonksiyonlar aracılığıyla bağdaştırıldıktan sonra Numpy kütüphanesine ait “np.array” komutuyla array haline getirilmiştir. Buna bağlı olarak Jupyter Notebook editörü üzerinde içeri alınan excel dosyasındaki değişkenlerle eşleşmiş olacaktır.

Böylelikle dizayn edilen html dosyası app.py dosyası çalıştırıldığında aktif olarak cmd ekranı üzerinden bir link üretilerek eş zamanlı modelle çalışmış olacaktır.

Uygulamanın arayüzüne erişmek için cmd ekranında kullanılan komutlar Şekil 29 'de verilmiştir.

```
Command Prompt - flask run
Microsoft Windows [Version 10.0.19045.2251]
(c) Microsoft Corporation. All rights reserved.

C:\Users\DenizZilyas>cd C:\Users\DenizZilyas\Desktop\MLTEZ

C:\Users\DenizZilyas\Desktop\MLTEZ>venv\Scripts\activate

(venv) C:\Users\DenizZilyas\Desktop\MLTEZ>set FLASK_APP=app.py

(venv) C:\Users\DenizZilyas\Desktop\MLTEZ>set FLASK_ENV=development

(venv) C:\Users\DenizZilyas\Desktop\MLTEZ>flask run
'FLASK_ENV' is deprecated and will not be used in Flask 2.3. Use 'FLASK_DEBUG' instead.
'FLASK_ENV' is deprecated and will not be used in Flask 2.3. Use 'FLASK_DEBUG' instead.
'FLASK_ENV' is deprecated and will not be used in Flask 2.3. Use 'FLASK_DEBUG' instead.
* Serving Flask app 'app.py'
* Debug mode: on
WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
* Running on http://127.0.0.1:5000
Press CTRL+C to quit
* Restarting with watchdog (windowsapi)
'FLASK_ENV' is deprecated and will not be used in Flask 2.3. Use 'FLASK_DEBUG' instead.
'FLASK_ENV' is deprecated and will not be used in Flask 2.3. Use 'FLASK_DEBUG' instead.
'FLASK_ENV' is deprecated and will not be used in Flask 2.3. Use 'FLASK_DEBUG' instead.
* Debugger is active!
* Debugger PIN: 320-142-307
```

Şekil 29. Cmd Komutları

FLASK_APP=app.py komutuyla hangi python uygulamasının çalışacağı belirtilir.

FLASK_ENV=development komutu kullanılarak, oluşturulan environmentın development modunda çalışacağı söylenir.

Uygulamanın içinde olduğu dizin “flask run” ile çalıştırılarak bir link üretilir. Bu link ile arayüze erişim sağlanır.

Uygulamanın arayüzü Şekil 30’da gösterilmiştir. Öğrencilere sunulan anket soruları böylelikle bir uygulamaya aktarılarak model ile eş zamanlı çalışacak hale getirilmiştir. 10 anket sorusu cevaplandıktan sonra Tahmini Gör butonuna basılarak girdilerle bağlantılı bir sonuç yani Türkçe Notu üretilcektir.

1-Cinsiyetiniz
☒ Kız ☐ Erkek

2-Kardeşiniz var mı?
☒ Evet ☐ Hayır

3-Özel ders alıyor musunuz?(Dershane, özel hoca ,kurs vb..
☒ Evet ☐ Hayır

4.Kendinize ait bir odanız var mı?
☒ Evet ☐ Hayır

5.Anneniz ve babanız ayrı mı?
☐ Evet ☒ Hayır

6.Anneniz veya babanız hayatta mı?
☒ Evet ☐ Hayır

7.Üvey anneniz veya babanız var mı?
☐ Evet ☒ Hayır

8.Günde kaç saat ders çalışıyorsunuz?

9.Günde kaç saat uyuyorsunuz?

10.Ailenizin aylık geliri nedir?

[Tahmini Gör](#)

Tahmin: [69.0829278]

Şekil 30. Uygulamanın Arayüzü

SONUÇ

Bu tez çalışmasında, öğrencilerden toplanan verilerle bir veri seti yaratılmış, 6 farklı algoritma kullanılarak modeller oluşturulmuş ve en tahminkar sonuç üreten Random Forest algoritması seçilmiştir. Model %80 oranında veriyle eğitilerek Türkçe Dersi notu tahmini yapacak algoritma ile eş zamanlı çalışacak arayüze aktarılmıştır.

Model %88 başarı oranına ve %98 doğruluk oranına(R-squared) sahiptir. Yapılan korelasyon incelemesinde Income (gelir) (%44) ve Studying (ders çalışma) (%64) değişkenleri model üzerinde pozitif yönde etkiye sahiptir. Sleeping (uyku) değişkeni model üzerinde negatif yönde (-%32) etkiye sahip olduğu hesaplanmıştır. Uyku değişkeninin negatif etkiye sahip olmasına rağmen modelin tahmin sonucunu negatif yönde etkilemediği daha çok nötr etkiye sahip olmasından kaynaklı veri setinden diğer 2 değişken gibi çıkarılmamıştır. Bu değişken modelin performansını arttıracak bir etkiye sahip olmadığı kanısına varılmıştır.

HTML kullanılmasının tez çalışmasına katkısı uygulama görsel olarak zenginleştirilmiş ve kullanıcılara karmaşık gelmeyecek basit bir arayüzde sunulmuştur.

HTML de hazırlanan arayüz arka planda flask kütüphanesi kullanılarak python kodlarıyla desteklenmiştir ve tezin en önemli aşamalarından biri olmuştur.

Bu çalışma Türkiye’de mevcut eğitim sistemindeki öğrencilerin hangi etkenlere bağlı olarak başarılı yada başarısız olduğu sonucuna ulaşmak için yapılmıştır.

Gelir durumu yüksek olan yada daha çok ders çalışan bir öğrencinin notunun daha yüksek olabileceği sonucuna ulaşılmıştır. Veri setinde toplamda 519 gözlem vardır. Ancak veri seti büyütülürse modelin performansı pozitif yönde olacaktır. Bunun sebebi eğitim ve test için ayrılan veri setinde kullanılacak verinin çok daha fazla olmasından kaynaklıdır. Böylelikle model

daha iyi eğitilecek ve çok daha tahminkar sonuçlar verebilecektir. Model yeni verilerle 50 kere test edilmiştir.

Bu veriler 2021 aralık ayında toplanmıştır bu sebeple test aşamasında gelir değişkeni göz önünde bulundurulurken 2021 yılı baz alınarak test edilmiştir.

Bu tez çalışmasının literatüre katkısı yeni bir çalışma konusu olarak tanımlanabilir. Yapılan literatür taramasında eğitim alanında fazla çalışma yapılmamış olması üzerine tez çalışması tasarlanmıştır. Ayrıca çalışmada kullanılan GBM algoritması yaygın olarak kullanılmadığından literatüre yeni bir algoritma tanımı kazandırmak amaçlanmıştır.

Bilgisayar mühendisliği kapsamında olan tez, makine öğrenmesi ile yaratılan bir modelin eğitildikten sonra sonuçları eş zamanlı çalışacak bir ortam üzerinden görüntüleme imkânı sağlamıştır. Bu kapsamda, bir Web arayüzü tasarlanmış olup arka planda yeni girdilerden bir tahmin üretmesi klasik çalışmalardan farkını göstermiştir.

Genel olarak yayınlanan tez, makale, rapor çalışmaları sağlık ve teknoloji üzerine olmuştur. Ve bu çalışmalardan çıkan sonuçlar sağlık alanında kullanılarak bazı hastalıkların önlenmesine olanak sağlamıştır.

Türkiye’deki eğitim başarısı genel bir problemdir. Bu alanda daha fazla çalışma yapılarak eğitim sorununa bir çözüm üretmek gerekmektedir. Öğrencileri etkileyen faktörler üzerine yoğunlaşılmalıdır. Ayrıca eğitim kalitesine müdahalede bulunulmadır. Öğrencilere ders çalışması için gerekli teşvikin sunulmasında yegane sorumlu ailesi sonrasında öğretmenleridir. Ders çalışma isteği öğrencilerin öğretmenlerine, ailesiyle ilişkisine bağlı olarak değişebilmektedir. Ezber sisteme dayanan bir eğitim söz konusu olduğunda öğrencilerin çalışma isteğini üzerinde negatif etkileri olmaktadır.

Gelir durumu anne ve babanın yaptığı işle ilişkilidir. Aile geliri iyi olmayan başarılı öğrenciler de bulunmaktadır. Bu konuda genelleme yapılmamalıdır. Ancak bu tez çalışmasındaki sonuca göre gelir durumu iyi bir aileye sahip öğrencilerin notlarının daha yüksek olduğu ortaya çıkarılmıştır.

Bu modele bağılı olarak, okullarda her öğrenci için yapılacağı varsayılırsa ve alınan çıktıları göre başarısız olabilecek öğrenciler üzerinde yoğunlaşırsa Türkiye’deki eğitim başarısında artış sağlanabilir. Veri setindeki gözlem sayısının artışı daha sağlıklı ve güçlü sonuçlar alınmasını sağlayacaktır.

Literatür araştırmasında Murak Gök’e ait Makine Öğrenmesi Yöntemleri İle Akademik Başarının Tahmini adlı makale çalışması giriş bölümünde referans olarak gösterilmiştir. Kullanılan algoritmalar, anket sorularının benzerliği yönünden bu tez çalışmasına benzemektedir. Çalışmada Türkçe ve Matematik dersindeki puan, not, genel başarı tahminleri üretecek algoritmalar kullanılmıştır. Puan tahmininde en performanslı algoritma Rastegale Orman veya Random Forest algoritması olmuştur. Makine Öğrenmesi İle Eğitim Başarısının Tahmini tez çalışmasında kullanılan 6 algoritma içerisinde de en başarılı yöntem Random Forest (%88) kabul edilmiştir. Bu tez çalışmasından farkı çevrimdışı bir araştırmadır ve ek olarak matematik dersi de bir bağımlı değişken olarak kabul edilmiştir. Ayrıca makale çalışmasında sadece model performans değerlendirilmesi yapılmıştır. Akademik başarıya etkisi olabilecek öznitelikler için ayrı bir sosyolojik çalışma gerektiğini ileriye sürmüştür.

KAYNAKÇA

- Ağca, Yılmaz. (2021). «Otel Oda Fiyatlarını Açıklamada Makine Öğrenmesi Algoritmalarının Kıyaslanması», *Journal Of Business Research-Turk* 2021, 13(1), 45 0-46
- Akşehir, Zinnet Duygu ve Kılıç, Erdal . 2019. «Makine Öğrenmesi Teknikleri ile Banka Hisse Senetlerinin Fiyat Tahmini », *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi (2019 Cilt:12- Sayı:2)* -30
- Aydın, Fatih. 2011. «Kalp Ritim Bozukluğu Olan Hastaların Tedavi Süreçlerini Desteklemek Amaçlı Makine Öğrenmesine Dayalı Sistemin Geliştirilmesi. », *Yüksek Lisans Tezi*
- Ayık Ziya, Özdemir Abdulkadir, Yavuz Uğur. 2007. « Lise Türü Ve Lise Mezuniyet Başarısının, Kazanılan Fakülte İle İlişkisinin Veri Madenciliği Tekniği İle Analizi», *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi. 10(2): 441-454*
- Azure, M. Retrieved from «What is machine learning: » <https://azure.microsoft.com/tr-tr/resources/cloud-computing-dictionary/what-is-machine-learning-platform> , *Microsoft Azure*.
- Bengio, Yoshua, LeCun, Yann, & Hinton, Geoffrey. 2015. « *Deep Learning*. » https://www.researchgate.net/publication/277411157_Deep_Learning
- Berker Kılıç, P. Y. (2021). Yargıtay Kararlarının Suç Türlerine Göre Makine Öğrenmesi Yöntemleri İle Sınıflandırılması . *Yargıtay Kararlarının Suç Türlerine Göre Makine Öğrenmesi Yöntemleri İle Sınıflandırılması* .
- Çelenk , Süleyman. 2003. «Okul Başarısının Ön Koşulu: Okul Aile Dayanışması », *Çelenk, S. İlköğretim-Online 2 (2), 2003 sf. 28-34*
- <https://dergipark.org.tr/tr/download/article-file/429518>
- Chollet, François. 2019. « Deep Learning With Python. », *Buzdağı Yayınevi, p.4-8*
- Çolakoğlu, Ahmet Arif. 2020. « Makine Öğrenmesi Algoritmaları ile Avrupa Havalimanları Analizi. », *Yüksek Lisans Tezi*

- David G. Kleinbaum, v. Klein Melanie. 2010. «*Logistic Regression: A Self-Learning Text (Statistics for Biology and Health)*» », Third Edition. New York: Springer 2010
- Tosunoğlu, Ergin. Yılmaz, Ramazan. Özeren, Eren. Sağlam, Zeynep (2021). «Eğitimde Makine Öğrenmesi: Araştırmalardaki Güncel Eğilimler Üzerine İnceleme. », *Cilt: 3 Sayı: 2 Yıl: 2021 E-ISSN: 2687-1750*
- Gök, Murat. (2017). «Makine Öğrenmesi Yöntemleri İle Akademik Başarının Tahmin Edilmesi. », *Yalova Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 77100, YALOVA*
- Goy, Gökhan. Gezer, Cengiz. Güngör, Vehbi Çağrı. «Makine Öğrenmesi Yöntemleri ile Kredi Kartı Sahteciliği Tespiti. »
<https://ieeexplore.ieee.org/document/8906995>
- Görmez, Burak. 2021. « Makine Öğrenmesi Algoritmaları İle Terör Olaylarının Tahmin Edilmesi Çalışması » , *Yüksek Lisans Tezi*
- Grubbs, 1969. «Procedures for Detecting Outlying Observations in Samples. *Technometrics*»
<http://dx.doi.org/10.1080/00401706.1969.10490657>
- Gujarati, Damodar N. ve Madsen, Jakob B. 1998. «Basic econometrics. *Journal of Applied Econometrics*(13).», *4th edition* : 209-212
<http://zalamsyah.staff.unja.ac.id/wp-content/uploads/sites/286/2019/11/7-Basic-Econometrics-4th-Ed.-Gujarati.pdf>
- Hastie, Trevor , Tibshirani, Robert, Friedman, Jerome H. 2009. «*The elements of statistical learning: data mining, Inference and Prediction* », *2nd edn.* , Springer, New York, USA p. 533.,
- Haykin , Simon. 1999. «*Neural Networks: A Comprehensive Foundation.* . » New Jersey: Prentice Hall Inc
- Haykin, Simon. 2009. «*Neural Networks and Learning Machines.* » . NJ: Pearson Education Inc

- Hosmer, David W. , Lemeshow, Stanley, Sturdivant, Rodey X. 2013. «*Applied Logistic Regression.* » 3rd Edition, John Wiley & Sons,.
- Hsu, C.-W. C.-C.-J. 2010. « *A Practical Guide to Support Vector Classification* » Department of Computer Science. » National Taiwan University, Taipei, Taiwan 16
- James, G. Witten, Hastie, Trevor, & Tibshirani, Robert .2017. «*An Introduction to Statistical Learning.* » , New York Heidelberg Dordrecht London : Springer
- Kabacan, Duygu. 2020. Retrieved from Makine Öğreniminin Gerçek Hayattaki Kullanımına En İyi Örnekler: <https://www.yapayzekatr.com/2020/09/23/makine-ogreniminin-gercek-hayattaki-kullanimina-en-iyi-ornekler/>
- Kennard, Hoerl. 1970. «Biased estimation for nonorthogonal problems. Technometrics(12). » , p 55-67
- Koptur, Murat. 2017. Retrieved from Makine Öğrenimi : <https://makineogrenimi.wordpress.com/2017/05/30/yanlilikvaryans-ikilemi-biasvariance-tradeoff/>
- Toçoğlu, Mansur Alp. Çelikten, Azer. Aygün, İrfan. Alpkoçak, Adil (2019). «Türkçe Metinlerde Duygu Analizi için Farklı Makine Yöntemlerinin Karşılaştırılması. » , *Yüksek Lisans Tezi*
- McCullouch, Pitts. 1943. «*The bulletin of mathematical.* » <https://home.csulb.edu/~cwallis/382/readings/482/mccolloch.logical.calculus.ideas.1943.pdf>
- Molina R., P´erez de la Blanca, Nicholas, Taylor, C.C. 1994. «*Modern statistical techniques.* » In Michie, D., Spiegelhalter , D.J., Taylor,
- Nilsson, N. 2009. « A HISTORY OF IDEAS AND ACHIEVEMENTS. »
- Nizam, Hatice ve Akın, Saliha Sıla. 2014. « Sosyal Medyada Makine Öğrenmesi İle Duygu Analizinde Dengeli Ve Dengesiz Veri Setlerinin

- Performanslarının Karşılaştırılması. » , *XIX Türkiye’de İnternet Konferansı Bildirileri 27-29 Kasım İzmir*.
- Yavuz, Ömer Çağrı. Karaman, Ersin. (2021). « Astronomik Gözlem Kalitesi Tahmininde Makine Öğrenmesi Algoritmalarının Kullanımı. » , *Uluslararası Yönetim Bilişim Sistemleri ve Bilgisayar Bilimleri Dergisi*, 2021, 5(1):12-19
- Onur Seveli ve Ali Tezcan Sarızeybek. 2019. « Makine Öğrenmesi Yöntemleri İle Banka Müşterilerinin Kredi Alma Eğiliminin Araştırma Analizi. » , *Zeki Sistemler Teori ve Uygulamaları Dergisi* 5(2) 2022 137-144
- Ustalı, Nesrin Koç. Tosun. Nedret ve Tosun, Ömer. 2020. « Makine Öğrenmesi Teknikleri İle Hisse Senedi Fiyat Tahmini. » , *Osmangazi Üniversitesi İİBF Dergisi*, 16(1), 1 – 16.
- Öztemel ,Ercan. 2006. «Yapay Sinir Ağları » , *Papatya Yayıncılık, İstanbul*
http://papatyabilim.com.tr/PDF/yapay_sinir_aglari.pdf
- Şenel, F. A. (2020). «Makine Öğrenmesi Algoritmaları Kullanılarak Kayısı İç Çekirdeklerinin Sınıflandırılması. » , *Yüksek Lisans Tezi*
- Tekin, Mahmut. 2008. «Sayısal Yöntemler: Bilgisayar Çözümlü Alıştırmalar.» , *(Güncellenmiştir 6. Baskı) , Konya: Günay Ofset*.
- Uzun ,Erdoğan. 2022. Retrieved from Erdoğan Uzun
https://erdincuzun.com/makine_ogrenmesi/model-degerlendirme-siniflandirma
- Yavuz, Erol. (2020). «Konutlarda enerji tüketimi kestirimi için derin öğrenme ve makine öğrenme yöntemlerinin karşılaştırılması. » , *Yüksek Lisans Tezi*
- Yavuz, Ünal. E. 2011. « Üniversite Öğrencilerin Ortak Zorunlu Derslerdeki Başarılarının K-Means Algoritması İle İncelenmesi. » *NWSA: Engineering Sciences*, 6(1), 342-347.
- Zach. 2021, 9 30. « Retrieved from Statology» <https://www.statology.org/mse-vs-rmse/>

EKLER

Ek-1. Anket

Bu çalışmada veri setini oluşturmak için kullanılan anket soruları.

1-Öğrenci cinsiyeti :

☐ Erkek ☐ Kız

2-Haftalık ortalama kaç saat ders çalışıyorsunuz?

3-Günde kaç saat uyuyorsunuz?

4-Özel ders alıyor musunuz?

☐ Evet ☐ Hayır

5-Günlük bilgisayarınızda kaç saat oyun oynuyorsunuz?

6-Aile geliriniz nedir ?

7-Kardeşiniz var mı?

☐ Evet ☐ Hayır

8-Kendinize ait odanız var mı ?

☐ Evet ☐ Hayır

9-Anne-babanız ayrı mı ?

☐ Evet ☐ Hayır

10-Anneniz yada Babanız sağ mı?

☐ Evet ☐ Hayır

11-Anneniz yada Babanız üvey mi ?

O Evet O Hayır

12-En keyif aldığınız dersler : (İlk 3 sırada belirtin.)

13-Kitap okuyor musunuz ? Haftanın kaç saatini kitap okumaya ayırıyorsunuz?

14-8.sınıf Türkçe notunuz :