GENİŞLETİLMİŞ ÖZET

Eğitim Verilerinin Makine Öğrenmesi Algoritmaları Kullanılarak Modellenmesi

Ayşe İlknur DİLEK

2000006380

GENİŞLETİLMİŞ ÖZET

Çalışmanın Amacı

Ülkemizde akademik başarının önemi her geçen gün artmakla birlikte akademik başarıyı etkileyen faktörler çeşitlilik göstermektedir. Bu çeşitlilik; farklı alanlarda, farklı faktörlerle olmakla birlikte bu değişkenlerin bir arada değerlendirilmesinin ve bunun sonucunda tahmin algoritmaları kullanılarak akademik başarıyı yordayan değişkenlerin kendi içlerinde birbirlerini etkileme ve hedef değişken olan akademik başarıyı etkileme gücü problemin konusunu oluşturmuştur. Bu çalışmada amaç; Lise öğrencilerinde akademik başarıyı etkileyen demografik, sosyoekonomik, tutum, sosyal aktivite, motivasyon, sağlık ve spor, akademik başarı kategorilerinde yer alan anket soruları yardımı ile akademik başarının çalışmanın büyük çoğunluğunda hedef değişken olarak yer alması ve bu faktörlerin akademik başarı hedef değişkenini etkileme derecesinin tespit edilip hangi makine öğrenmesi modellerinin bu gücü anlamlı bir şekilde yorumlayabildiği değerlendirilmesi amaçlanarak bu çalışmanın sonucunda akademik başarıyı etkileyen faktörlerin ve etkileme derecelerinin belirlenerek eğitim sistemine, özellikle öğrenciye, katkı getirmesi amaçlanmıştır.

Araştırma Soruları

Akademik başarıyı etkilediği varsayılan faktörler olan demografik, sosyoekonomik, tutum, motivasyon, sosyal aktivite, sağlık ve spor kategorisinde yer alan soruların kendi kategorisi içerisinde her birinin akademik başarıyı etkileme gücü, etkileme derecesi nedir? Akademik başarıyı etkileyen faktörlerin birbirlerini etkileme derecelerini hesaplayınız? Denetimli öğrenme modellerinden olan Regresyon çeşitlerinden Multilineer Regresyon, Ridge ve Lasso regresyonlarının başarı oranları ve değerlendirilmesi nedir? Denetimli öğrenme modellerinden Sınıflandırma algoritması modellerinden olan Karar ağacı, Rastgele orman, En yakın Komşular, Destek vektör makinaları algoritmalarından hangileri başarılıdır, başarı oranları nelerdir, değerlendirilmesi nedir? Kolektif öğrenme modellerinin başarı oranları nelerdir,değerlendirilmesi nedir? Derin Öğrenme modellerinden olan Yapay sinir ağları modelini değerlendiriniz.

Akademik başarının artırılmasına yönelik çalışmalar her geçen gün artmakla birlikte teknolojinin gelişmesi ile birlikte bilgisayar bilimleri, akademik başarıyı etkileyen faktörlerin değerlendirilmesinde büyük katkılar sağlamaktadır. Makine öğrenmesi algoritmaları kullanılarak eğitim verilerinin modellendirilmesi ve veri madenciliği ve Yapay zekanın birleşimiyle verilerin sınıflandırma, tahmin ve kümeleme çalışmaları yapılmaktadır.

Çalışmaların ulusal ve uluslararası düzeyde sürekli gelişerek artması bu konudaki akademik araştırmaların niteli ve niceliğini geliştirerek bilgiye kolay ulaşılabilinmesine de katkıda bulunmuştur. Bu çalışma yapılırken Ulusal tez merkezi, uluslararası düzeydeki tezler, çeşitli branşlarda olmak şartıyla makaleler (özellikle sosyal bilimlerdeki makaleler çok fazlasıyla taranmıştır.).Dergide yayınlanan makaleler, dergi köşe yazıları incelenmiştir. Kütüphane ziyaretleri yapılarak kaynaklara direkt ulaşım sağlanılmakla birlikte online yayınlar ve online makalelere, çevrimdışı verilere uzaktan eğitim kapsamında erişim sağlanılmıştır. Konu ile ilgili adı geçen sözcükler detaylı bir şekilde incelenmiştir. Aşağıda incelenen tezler arasında konu kapsamı, içerik, kullanılan algoritmalar açısından bu çalışmaya benzer 3 çalışmadan bahsedilmiştir.

Türkiye'de Yalova ilinde 3 farklı ortaokulda uygulanan anket sonucunda öğrencilere demografik, sosyaekonomik, sağlık, spor, sosyal, aktivite, not başarı durumları ile ilgili sorular yöneltilmiş . Türkçe, Matematik ve dönem sonu not ortalamaları hedef değişken alınarak sınıflandırma ve regresyon kullanılarak tahmin algoritmaları sonucunda yordama gücü öznitelik seçiminin de uygulanması ile birlikte anlamlı sonuçlar elde edilmiştir. (Makine Öğrenmesi Yöntemleri İle Akademik Başarının Tahmin Edilmesi Murat GÖK1, * 1Yalova Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 77100, YALOVA)

Portekizde, 2005 2006 yılları arasında, iki devlet okulunda yapılan araştırmada öğrenci dağılımı 9 yıllık temek eğitim sonrasındaki gruptur. Matematik ve Portekizce notları ülkedeki eğitim sistemleri 3 aşamada değerlendirilmiş olup G1,G2,G3 olarak isimlendirilmiştir.G3 final notudur. Bu değişkenler hedef değişken olmakla birlikte Karar ağaçları ,Rastgele Orman ,yapay sinir ağları ve Destek Faktör makinaları olmak üzere farklı sınıflandırma algoritmaları kullanılmış tahmin yapılmıştır. Özellikler arasında kullanılan algoritmalarla anlamlı tahminler çıkarılabilmekle birlikte daha az etkileyen değişkenlerin var olduğu da gözlenmiştir. Ayrıca ANN ve SVM yöntemlerinin gürültülü girdilere, aykırı değerlere değişkenlerine karşı daha hassas yöntemler oldukları gözlenmiştir.

İncelenen üçüncü çalışma Kaggle platformundan hazır data kullanmış ve Karar ağacı, Rastgele orman ile sadece Lojistik regresyon kullanarak tahmin çalışmaları yapmıştır. Bu çalışmada 395 ve 245 öğrenci sayıları olmak üzere iki farklı veri seti kullanılmıştır. Tüm özellikler bu veri seti için aynıdır. En iyi doğruluk oranı Karar Ağacı algoritmasına aittir. Data setleri ayrı ayrı değerlendirilmekle birlikte 649 öğrenci total olarak da değerlendirilmiş.3 farklı veri seti seti kullanıldığı zaman ise en fazla sayıda öğrenci sayısıyla en yüksek doğruluk değeri yine Karar ağacına aittir.

Yöntem

Bu çalışmada öğrenmeyi etkileyen faktörler farklı kategorilerde olmak koşulu ile ayrıntılı bir şekilde açıklanmıştır. Makine öğrenme modellerinden Denetimli, Denetimsiz, öğrenme kavramları açıklanmıştır. Makine öğrenmesi Denetimli öğrenme modellerinde sınıflandırma algoritmaları olan Karar Ağacı,Rassal Orman,K-en yakın komşular,Lojistik regresyon ,Destek vektör makinaları,Regresyon algoritmaları olan Multilineer regresyon ,Ridge ve Lasso regresyonları ,Kolektif öğrenme modelleri ve Derin öğrenme modellerinden yapay sinir ağları modelleri açıklanmıştır

Kaggle'dan edinilen veri ilk önce kullanılabilir olacak şekilde hazır hale getirilmiştir. Makine öğrenme algoritmaları ile Denetimli öğrenme modellerinden olan Sınıflandırma, Regresyon,Kolektif öğrenme modelleri uygulanmış ve başarılı sonuçlar elde edilmiştir. Derin öğrenme modeli olan Yapay Sinir Ağları modelleri veri setine uygulanmıştır. Tahmin,

sınıflandırma ve kümeleme çalışmaları sonucunda model performansı sınıflandırma algoritmaları için doğruluk değerleri ve çeşitleri, Roc eğrisi, karmaşıklık matrisi kullanılarak değerlendirilmiştir. Regresyon modelleri olan Multilineer regresyon, Ridge ve Lasso regresyon modelleri eğitim ve test seti sonuçlarına göre değerlendirildiğinde sonuç değerlerinin aynı olduğu gözlemekle birlikte Ortalama kareler hata katsayısına göre değerlendirildiğinde en iyi çalışan regresyon modelinin Ridge Regresyon olduğu kararına varılmıştır. . Derin öğrenme algoritması olan Yapay sinir ağ modelinde perceptron kulanılarak başarılıbir sonuç elde edilmiştir .

Sonuç ve Değerlendirme: Regresyon modelleri kendi içerisinde, sınıflandırma modelleri kendi içerisinde değerlendirilerek en iyi performansla çalışan modeller değerlendirildiğinde; Regresyon modelleri içerisinde Multilineer Regresyon, Lasso Regresyon ,Ridge Regresyon modelllerinin eğitim ve test sonuçları (her üçünün de aynı) sırasıı ile 0.87 ve 0.77 dir. Ortalama kareler hata katsayısı değerleri incelendiğinde içerisinde Multilineer Regresyon 6.40, Ridge Regresyon 6.41, Lasso Regresyon 6.39 ortalama kareler hata katsayısına sahiptir. Regresyon modellerinde değerlendirme yapıldığında diğerlerinden açık ara fark olmamak üzere skorlar değerlerine bakılarak en iyi performansla çalışan sınıflandırma modeli Lasso Regresyon olmuştur. Sınıflandırma modelleri kendi içlerinde değerlendirildiğinde;

Karar Ağacı algoritması değerlendirildiğinde Doğruluk değeri : 0.89 Roc eğrisi altında kalan alan değeri :0.97

Rassal Orman algoritması değerlendirildiğinde Doğruluk değeri : 0.91 Roc eğrisi altında kalan alan değeri :0.97

Destek Vekör Makinası algoritması değerlendirildiğinde Doğruluk değeri : 0.92 Roc eğrisi altında kalan alan değeri :0.97

XgBoost algoritması değerlendirildiğinde Doğruluk değeri : 0.90 Roc eğrisi altında kalan alan değeri :0.97

AdaBoost algoritması değerlendirildiğinde Doğruluk değeri : 0.86 Roc eğrisi altında kalan alan değeri :0.95

Bagging algoritması değerlendirildiğinde Doğruluk değeri : 0.92 Roc eğrisi altında kalan alan değeri :0.97

Lojistik regresyon algoritması değerlendirildiğinde Doğruluk değeri : 0.94 Roc eğrisi altında kalan alan değeri : 0.97

K- En yakın komşular algoritması değerlendirildiğinde Doğruluk değeri : 0.80 Roc eğrisi altında kalan alan değeri :0.84 sonuçlarına ulaşılmıştır.

Yapay Sinir Ağları algoritması değerlendirildiğinde Doğruluk değeri : 0.94 Roc eğrisi altında kalan alan değeri :0.89

Bu çalışmanın sonunda ; Türkiye'de farklı okul türleri, farklı sınıf düzeyleri, farklı bölgelerden oluşan geniş bir örneklemle öğrenmeyi etkileyen faktörler farklı kategorilerde ve geniş bir şekilde yer almak şartı ile ,öğrenmeyi etkileyen faktörlerin başarılı algoritmalar ve modeller ile birlikte toplanan veri setine uygulanması ve bu çalışmada anlamlı sonuçlar veren geliştirdiğimiz model ve algoritmaları uygulayarak ülkemizde eğitime katkı sağlamaktır.

Anahtar Kelimeler: Makine öğrenmesi, Derin Öğrenme, Yapay zeka, Yapay sinir ağları, Çoklu Lineer regresyon, Polinomsal regresyon, Lojistik regresyon, Lasso and Ridge regresyonları, Karar ağacı, Rastgele Orman,Destek Vektör Makinaları, En yakın K komşuları,

Yapay sinir ağları, K ortalama algoritmaları,Topluluk öğrenmesi,

Bilim Dalı Sayısal Kodu : 20515

University : Istanbul Kültür University
Institute : Institute of Graduate Education
Department :Mathematics and Computer Science
Programme : Mathematics and Computer Science
Supervisor : Asst.Dr.Mehmet Fatih UÇAR
Degree Awarded and Date : MA – June 2022

ABSTRACT

MODELING EDUCATIONAL DATAS WITH MACHINE LEARNING METHODS

Ayşe İlknur DİLEK

In our country, the effect of the academic success of the student, especially in the secondary education period, on the stage of choosing the profession he will have in the future and on the academic career goal is an undeniable reality. Academic success is affected not only by the data belonging to the academy, but also by many different categories. It is affected by many factors, especially methodological, and this diversity increases with individual differences. Regression and Classification from supervised learning models and Clustering algorithms from unsupervised learning models were applied to the data set. Multiple linear regression, polynomial regression, Lasso and Ridge regressions,Decision Tree, Random Forest, Support Vector Regression as regression methods, Decision Tree, Random Forest, Support Vector Machine, Logistic regression, K Nearest Neighbors methods were used as classification methods. As Clustering methods we are used K means algorithms, hierarchical method as unsupervised learning methods. In addition Artifical Neural Network, a deep learning algorithm, were applied to the data set. In the study, these factors and sub-factors were evaluated categorically and machine learning was used. Various determinations were made with estimation algorithms by establishing relations that predict the academic achievement target variable . By evaluating the data results, it is aimed to determine which factors affecting success are significant according to the sample group studied, which variables affect success individually and categorically, and the degree of influence, and as a result, it is aimed to contribute to education.

Keywords: Machine Learning, Deep Learning, Artificial intelligence, Artificial Neural Networks, Multiple linear regression , Polynomial regression, Logistic regression, Lasso and Ridge regressions, Decision tree, Random Forest, Support Vector Machine, Artifical Neural Network,Bagging,XgBoost,AdaBoost

Science Code : 20515