Beykent Üniversitesi
Veri Madenciliği Dersi
Dönem Projesi, Ocak 2021

Dönem Projesi


www.beykent.edu.tr

Kongre Oy Veri Seti Üzerinde Denetimli Öğrenme Çalışması

Uzay Peker[1]

[1*] Beykent Üniversitesi, İktisat Fakültesi, Yönetim Bilişim Sistemleri Bölümü, İstanbul, Türkiye, uzaypeker@hotmail.com

(İlk Geliş Tarihi Ocak 2022 ve Kabul Tarihi Ocak 2022)

Öz

Büyük miktarda veri içerisinden, gelecekle ilgili tahmin yapmamızı sağlayacak kuralların ve bağıntıların açığa çıkması için kullanılan Veri Madenciliği; bu özellikleri ile siyasal verilerinin önemini arttırmış ve kullanım yaygınlığı hızla artmaya devam eden bir yöntem haline gelmiştir. Bu çalışmada 1984 Amerika Birleşik Devletleri kongre seçim veri seti bulunmuş, denetimli öğrenme modeli oluşturularak tahminlerde bulunulmuştur. Çalışma Jupyter Notebook üzerinden Python dili ile hazırlanmış ve algoritma olarak K-NN seçilmiştir. Algoritma doğruluğunu ölçmek amacıyla score metodunun yanı sıra hata matrisi de kullanılmıştır. Çalışma sonucunda örnek veriyle siyasal parti sınıfı tahmin edilmiş olup parti kategorisini etkileyen en büyük faktörün doktor ücretleriyle ilgili olan sütun olduğu bulunmuştur.

Anahtar Kelimeler: Veri Madenciliği, Makine Öğrenmesi, Denetimli Öğrenme, Programlama Dilleri.

Implementation of a Supervised Model on US Congre Vote Data

Abstract

Data mining is the search for rules and relations that will enable us to predict the future from large amounts of data. In this study, the 1984 United States congressional vote data set was found, and predictions were made by a supervised machine learning model. The study was prepared with the Python language on Jupyter Notebook and, as an algorithm K-NN was chosen. In order to measure the accuracy of the algorithm, the error matrix and score method used.

Keywords: Data mining, Machine Learning, Supervised Learning, Programming Language.

1. Giriş

Günümüzde internet teknolojilerinin hızlı gelişimi, dünya popülasyonunun veya genel kullanıcıların artması ve şirketlerin gelecekteki tüketici hareketlerini tahmin etme istekleri gibi sebepler ile sahip olduğumuz büyük veriler artmış ve bunlar üzerinden anlamlı bağıntılar çıkarma çalışmaları hız kazanmıştır. Bu yöntemler sadece kâr amacı güden firmalar tarafından değil, sağlık ve eğitim sektörü gibi toplumu şekillendiren organlarda da çokça kullanılmaktadır. Bunlara örnek olarak veri madenciliği sayesinde kanser hastalığının tespiti, tüketicilerin satın alma davranışları ile ilgili veya gelecek bir zamanda salgın hastalıklardan en çok etkilenecek ülkeyi bulmak gibi tahmin modelleri oluşturulabilmekteyiz.

Kullanılacak veri seti içerisindeki bilgilerin metin tipinde olması sebebiyle veri madenciliği ile ilişkili birçok alandan yararlanılmıştır. Veri seti, 1987 yılında David Aha ve UC Irvine yüksek lisans öğrencileri tarafından oluşturulan UCI Machine Learning Repository sitesinden alınmıştır. Veri seti içerisinde 1984 yılında yapılmış olan oylamalar ve bu cevapların demokratlar mı yoksa cumhuriyetçiler tarafından mı verildiği belirtilmektedir.

Bu makalede ilk olarak kullanmış olduğum algoritmayı tanıtacak olup sonrasında ise bu algoritmayı gerçekleştirebilmek

için kullandığım araçlardan bahsedeceğim. Sonrasında araştırmam ile edindiğim sonuçlar ve grafikleri paylaşarak kullanmış olduğum yolları göstereceğim. Son bölümde ulaştığım bilgiler ışığında yaptığım çıkarımları paylaşacağım.

2. Materyal ve Metot

2.1. K-En Yakın Komşu (KNN) Algoritması

KNN algoritması, sınıflandırma veya regresyon için örüntü tanıma ve makine öğrenmesi kapsamında kullanılan klasik yöntemlerden bir tanesidir. KNN, sınıflandırılmak istenen nesnenin ait olduğu kümeyi, en yakınında yer alan K birim nesneden en fala birime ait olanla aynı kümede sınıflandırması mantığına dayanmaktadır. KNN basit olmasına karşın hesaplama açısından yoğun olabilecek bir algoritmadır. Uygulamada eğitim ve test veri kümelerinin çok büyük olması, çalışma ortamı içinde çıktı verme süresini çok uzatarak darboğaza sebep olabilmektedir. K-en Yakın Komşu algoritmasında eğitim olayı olmamaktadır. Avantajlarına rağmen, veri sayısının artmasıyla yüksek bellek alanına ihtiyaç duymakta, işlem yükü ve maliyetin önemli oranda artması algoritma performansının k komşu sayısı gibi parametreye ve özelliklere bağlı olarak etkilenmesi beraberinde dezavantajları getirmektedir.

KNN algoritması için birden fazla hesaplama yöntemi bulunmaktadır. Bunlar; Öklid, Manatthan uzaklığı ve Minkowski uzaklığıdır. Öklid hesaplamasından örnek verecek olursak, nokta ile merkezler arasındaki uzaklık farkları bulunduktan sonra kareleri alınır ve bu sayılar da birbirlerine eklendikten sonra karekök şeklinde sonu yazılır. Bu sonuçlardan küçük olan yani merkez alternatiflerinden en yakın olan komşu seçilir ve o sınıfa ait olduğu varsayılır. Öklid yöntemine dair formül aşağıda gösterilmiştir.

$$\text{Euclidean distance} = \sqrt{\sum_{i=1}^{n} \left(x_i - y_i\right)^2}$$

2.2. Programlama Dili: Python

Python, nesne yönelimli, yorumlamalı, bilimsel ve etkileşimli yüksek seviyeli bir programlama dilidir. Girintilere dayalı basit söz dizimi, dilin öğrenilmesini ve akılda kalmasını kolaylaştırmaktadır. Her geçen gün daha fazla şirket Python kullanmaya başlamakta ve dilin sunduğu ayrıcalıklarla verileri üzerinde analizler gerçekleştirmekte. Bunun gibi birçok sebepten ötürü şu anki çalışma için de Pyton seçilmiş olunup, veri ile ilgili işlemlerin gerçekleştirilmesi adına Pandas, Numpy, Sckit-learn, Seaborn ve Matplotlib kullanılmıştır. Neredeyse her projede kullanılan 2 önemli kütüphane aşağıda anlatılmıştır.

2.2.1. Pandas Kütüphanesi

Pandas paketi, günümüzün Python tabanlı veri bilimcileri ve analistlerinin erişebildiği en güçlüü platformdur. Tüm odak, güçlü makine öğrenimi ve şık görselleştirme yöntemlerine verilebilir. Ancak Pandas, Numpy paketinin üzerine inşa edilmiştir. Yani Pandas birçok Numpy yapısını kullanıyor veya üretiyor.

2.2.2. Numpy Kütüphanesi

Numpy Python dilinin önemli matematik kütüphanesidir. İsim olarak Numeric Python kelimelerinin kısaltmasından oluşur. Çok boyutlu ve toplu dizilerin içindeki objelerle işlem yapmayı kolaylaştırmaktadır. Bazı veri paketlerini kullanmanız için Numpy kütüphanesini çalışma ortamınıza import etmeniz gerekmektedir.

2.3. Jupyter Notebook

Juyter Notebook, bir web tarayıcısı üzerinden notebook belgesi formatındaki kodları düzenlemeyi ve çalıştırmayı sağlayan bir sunucu-istemci uygulamasıdır. Çıktığı ilk zamanlarda IPython Notebook olarak bilinmekteydi. Başlangıçta sadece Python desteklese de zaman içinde gelişerek Julia, Octave, R, Haskel, Ruby gibi dilleri de desteklemeye başlamıştır.

Jupyter Notebook, internet erişimi gerektirmeden çalışabilir, uzaktaki bir sunucuya kurulabilir ve internet üzerinden erişilebilir. Jupyter metin belgeleri görüntülemeye, düzenlemeye ve çalıştırmaya ek olarak, yerel dosyaları gösteren ve notebook belgelerini açmaya veya notebook çekirdeğini kapatmaya izin veren bir ‘Dashboard’ (Kontrol Paneli) içerir.

2.4. Veri Temizleme ve Dönüştürme

Veri setimiz tabloda gözüktüğü gibi boş değerler ve metin türünde bilgiler içermektedir. Yapacağımız KNN uygulamasında veri setimizin boş değer içermemesi gerekmektedir. Bu nedenden ötürü önümüzde iki seçenek belirmektedir. İlki boş veri içeren satırları düşürmek, ikincisi ise bu boş değerleri belli istatistiksel yöntemlerle doldurmak. İlk seçeneğimiz boş verilerimiz model eğitimlerimizi etkilemeyecek sayıda az olduğu zaman seçilebilir ancak burada boş veri içeren satırlarımız veri setinin nerdeyse %50’lik kısmını oluşturmaktadır. Bu neden ikinci seçeneği seçerek işlemlerimi tamamladım. Bu işlemleri Sckit-learn kütüphanesinin imputer metoduyla kolayca yaptım. Imputer metodu boş olan veri noktasına en yakın referans noktası veya noktalarıyla gerekli bilgiyi doldurabilmektedir. Bu yapı kullanacağımız KNN algoritmasına benzer ve uygun bir yöntemdir.

KNN modelimizi kullanabilmek için verilerimizin sayısal değerlerde olması gerekmektedir. Bunu gerçekleştirebilmek adına cumhuriyetçilerin ve olumlu oyların değerleri 1 ile, demokrat ve olumsuz oyların değerleri 0 ile değiştirilmiştir. Bu değerlendirmede siyasal olarak yanlılık bulunmamakta, veri setindeki ilk satıra göre puanlama yapılmıştır. Bu adımlar tamamlandıktan sonra veri setimiz model için hazır konuma gelmiştir.

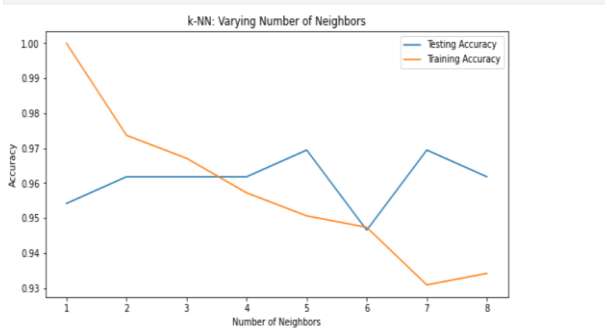
Tablo 1. Kongre Oyları Veri Seti

Class Name	handicapped-infants	water-project-cost-sharing	adoption-of-the-budget-resolution	physician-fee-freeze	11 Sütun Daha
republican	'n'	'y'	'n'	'y'	'n' veya 'y'

3. Araştırma Sonuçları ve Tartışma

3.1. K Değeri İçin Çizgi Grafiği

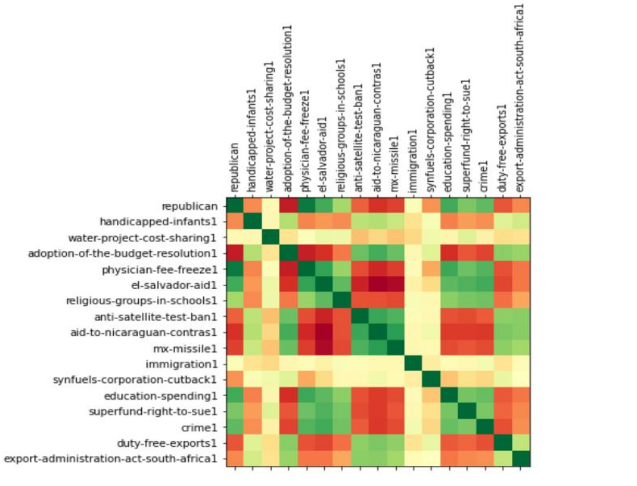
Modelimiz için seçeceğimiz optimum K sayısını bulmak amacıyla for döngüsü içerisinde model skorlarımızın çizgi grafiğini aşağıda oluşturduk. Bu grafik üzerinden K sayımızı 5 olarak belirledik.



Görsel 1. Optimum K Sayısı

3.2. Korelasyon Isı Haritası

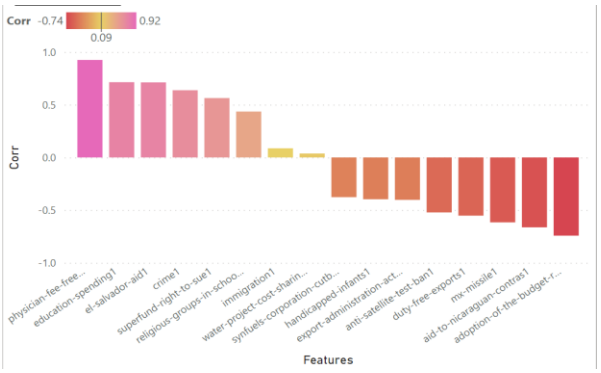
Verilerimizin hedef sütunumuzu hangi önem derecesinde etkilediğini bulmak amacıyla oluşturulmuş ısı grafiğini aşağıda görebiliriz.



Görsel 2. Korelasyon Isı Haritası

3.3. Korelasyon Bar Grafiği

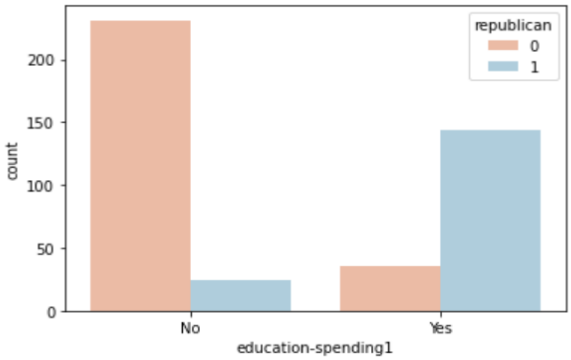
Verilerimizin hedef sütunumuzu hangi önem derecesinde etkilediğini daha iyi görebilmek adına yapılmış bar grafiği aşağıdadır. Bu grafik PowerBI raporlama aracı üzerinden yapıldığı için notebook içinde bulunmamaktadır. PowerBI klasöründeki dosyayı indirerek grafiği inceleyebilir ve değiştirebilirsiniz.



Görsel 3. Korelasyon Bar Grafiği

3.4. Eğitim Verileri ile Bar Grafiği

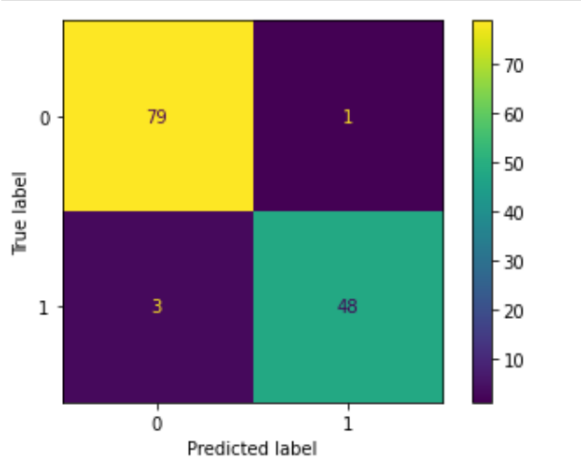
Korelasyon grafiklerinden yüksek değere sahip olan eğitim ile sınıflandırma çıkarımı yapmak amacıyla bar grafiği oluşturuldu.



Görsel 4. Eğitim Sütunu Bar Grafiği

3.5. Karmaşıklık Matrisi

Tahminlerimizin True Positive ve True Negative gibi sonuçlarını incelemek için oluşturulmuş grafik aşağıdadır.



Görsel 5. Karmaşıklık Matrisi

3.6. Sınıflandırma Raporu

KNN algoritması ölçümünde kullanılan önemli indikatörlerin bulunduğu sınıflandırma raporu aşağıda verilmiştir. Burada benim dikkate alacağım metrik F1 skordur. F1 skor, karmaşıklık matrisindeki verilerle ortaya çıkan, 0 ile 1 arasında değer alan ve 1'e yaklaştıkça başarının arttığı anlamına gelen bir metriktir.

	precision	recall	f1-score	support
0	0.96	0.99	0.98	80
1	0.98	0.94	0.96	51
accuracy			0.97	131
macro avg	0.97	0.96	0.97	131
weighted avg	0.97	0.97	0.97	131

Görsel 6. Sınıflandırma Rapor Değerleri

4. Sonuç

Bir önceki bölümde grafiklerin verilmiş olduğu sıra ile çıkarımlarımdan bahsedecek olursam. KNN algoritmasının başlangıç ve önemli bir noktası olan K seçimi her zaman döngüler yardımıyla deneme-yanılma yolu seçilerek bulunmalı. Kullanmış olduğumuz kongre seçim verilerinde parti sınıfını tahmin etmemizi sağlayan en önemli özellikleri bulmak için bar grafiğine bakabiliriz. Buradan en yüksek skora sahip iki özelliğimiz doktor ödemeleri ve eğitimdir. Bu çıkarımı ısı haritasına bakılarak, yeşillik derecesi en koyu olan renk sütunlarını bularak da yapabiliriz.

Modelle ilgili olarak, karmaşıklık matrisi incelendiğinde doğru bilmemiz yani True Positive ve True Negative kısımlarımızın çok yüksek olduğu görülmektedir. Bu da bize modelimizin gayet başarılı olduğu izlenimini vermekte. Modelimizin doğruluğunu bir başka metrikle daha ölçmek için sınıflandırma raporuna bakabiliriz. Buradaki F1 skorumuz iki sınıfımız için de 1'e oldukça yakın durmaktadır. Bu metriğimiz de 1'e ne kadar yakın olursa, model başarımızın da o kadar yüksek olduğunu belirtir. Skor bu kadar yüksek iken modelimizin veri setini ezberlediğini düşünebiliriz ancak ezber yapmasını engellemek amacıyla veri setimizi test ve eğitim olarak ayırmıştık. Son olarak eğitim için oluşturulan grafik, sınıflandırmaya etki eden en yüksek iki özellikten birini içerdiği için eğer tek bir özellikten tahmin yapmak istersek bu görseli kullanarak eğitime evet oyu vermiş kişilerin cumhuriyetçi olduğunu söyleyebiliriz.

5. Teşekkür

İlk olarak bu çalışmayı bize oluşturtan ve araştırmalarımdan yeni şeyler öğrenmemi sağlayan Dr. Öğr. Üyesi Atınç Yılmaz'a sonrasında birçok zorlandığım noktanın çözümünü bana öğreten Patika ekibine ve de son olarak veriyle ilgili çalışmalarda kendimi geliştirmeme olanak sağlayan Microsoft ailesine çok teşekkürler.

Kaynakça

Quansheng Kuang, and Lei Zhao,(2009), A Practical GPU Based KNN Algorithm,Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)

Deniz KILINÇ, Emin BORANDAĞ, Fatih YÜCALAR, Volkan TUNALI, Macit ŞİMŞEK, Akın ÖZÇİFT, (2016), KNN Algoritması ve R Dili ile Metin Madenciliği Kullanılarak Bilimsel Makale Tasnifi,Marmara Fen Bilimleri Dergisi 2016, 3: 89-94

Yun-lei Cai, Duo Ji ,Dong-feng Cai(2010), A KNN Research Paper Classification Method Based on Shared Nearest Neighbor,Proceedings of NTCIR-8 Workshop Meeting, June 15–18

Taşçı E., & Onan, A. (2016). K-en Yakın Komşu Algoritması Parametrelerinin Sınıflandırma Performansı Üzerine Etkisinin İncelenmesi, Akademik Bilişim.

N. Community, NumPy User Guide 1.19. 214, 2020.

McKinney, W. & Team, P. D. Pandas - Powerful Python Data Analysis Toolkit. Pandas - Powerful Python Data Anal. Toolkit. 1625, 2015