DİYABET ÜZERİNE MAKİNE ÖĞRENMESİ

EREN DEMİR
HTTPS://GİTHUB.COM/ERENCODE27

PROJEYE BAŞLAMAM

Size biraz başta bu projeye nasıl ve neden başladığımı anlatmak istiyorum. Gönüllü staj yaptığım yerde bana staj boyunca bir proje yapmamız gerektiği söylendi. Bu bana söylenmeden birkaç hafta önce makine öğrenmesine yeni giriş yapmıştım. İstatistik ve makine öğrenmesi alanında teorik olarak kendimi geliştiriyordum. Scikit-learn gibi kaynaklardan bizzat kodları, kütüphaneleri, fonksiyonları, parametrelerini araştırıyordum. Ve araştırdıkça keyif aldığım ve içine daldığım bir alan olduğunu fark ettim. Kaggle üzerinden bir proje bulmuştum ve dediğim kütüphane sayfalarının yanında o proje üzerinden kendimi geliştiriyordum. Proje şu;

PROJE HEDEFIM

Titanic veriseti üzerine bir makine öğrenmesi projesi. En temelden anlatan, mantık ve yapılan işi öğreten bir projeydi. O proje sayesinde çok şey öğrendim. Birçok şeyi araştırmam gerekti. Öğrenmek öğrenmeyi doğurur ne de olsa. Projede üzerinden geçilen meseleleri bazen oturup saatlerce araştırmam da gerekti ve bu saatler alan araştırmanın, damıtılmış halini hazırladığım notebookta bulabilirsiniz. Bence makine öğrenmesinde bilinmesi gereken birçok olaydan,neden makine öğrenmesi kullandığımızdan ve bir proje yapılırken projeden bir sonuç nasıl çıkartılır, bu olayların mantığından ve nasıl bir yol haritası oluşturulur bunlardan bahsediyor.

Örnek olarak aldığım kodu diyabet verisi üzerinde kullandım çünkü sağlık ve enerji sektörüne çok büyük bir ilgim var. Bu alanlarda makine öğrenmesinin öneminin inanılmaz bir şekilde artıcağına inanıyorum bu yüzden ve ilgim olduğundan dolayı diyabet verisini seçtim.

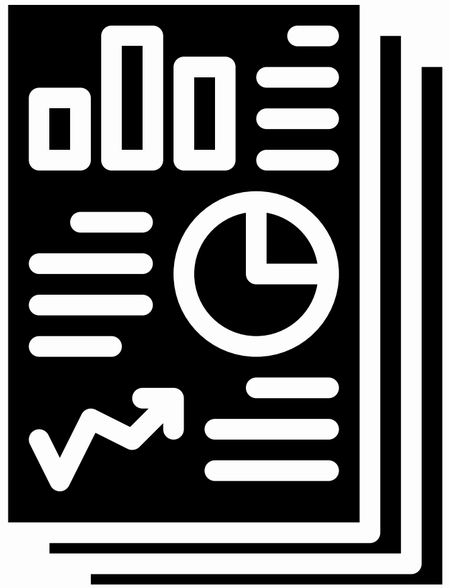
VIZYON

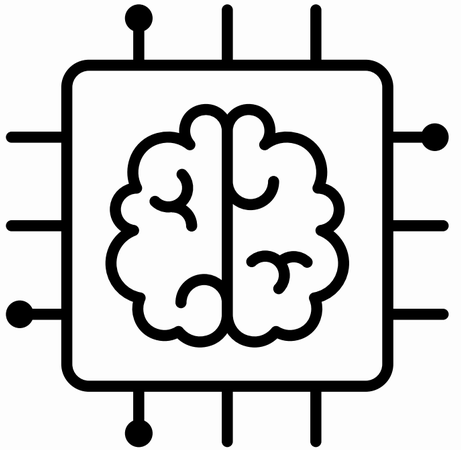
Ama projemi anlatmaya başlamadan önce biraz veri nedir? İstatistik nedir? Neden makine öğrenmesine ihtiyacımız var? Bunlardan bahsetmek istiyorum.

İstatistik en temel ve en kritik şekilde matematiği hayatımıza uyarlama bilimidir.

Direkt olarak matematiğin hayatımıza dokunduğu alandır. İnsan olarak algılayamayacağımız rastlantısal örüntüleri veya hiç rastlantısal olmayan örüntüleri anlamamızı sağlar. İlk olarak örüntüyü görürüz. Bir olayı etkileyen bir faktör olduğunu keşfederiz, ardından o faktör üzerinde araştırma yaparak veya düşünerek o faktörün neden etkilediğini öğreniriz.

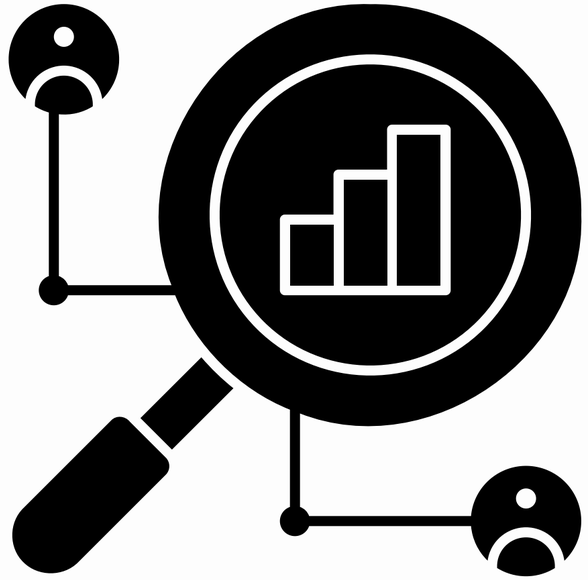
Yani ilk olarak ilişkinin varlığını sonra nedenini ve nasılını öğretir bize istatistik. İstatistik ile almayacağınız bir ürünün hangi koşullarda alındığını tespit edebiliriz. Bir ürün için gerçek hayat örneklemesi yapalım.





BIR ÖRNEK

Mesela bir tatlı üzerinden örnek verelim. Siz tatlıyı akıllı insanlara pek satamazsınız. Hayatında disiplin olan sağlıklı yaşam tarzını benimsemiş insanlara satamazsınız, veya istediğiniz kadar satamazsınız.



Problem Tespiti

Ama sürekli olarak tatlı tüketen insanlara satabilirsiniz. O yüzden burda yemeksepeti gibi uygulamalarda puanı çok yüksek olan insanların verisi çok büyük önem taşıyor. Çünkü size hazır, kategorize edilmiş, ayıklanmış bir insan profili verir. Veya sosyal medyadan, bu büyük firmaların algoritmalarını ele alalım.

Kategorizasyon

Yapay zeka ile insanların tükettiği içerikler direkt olarak etiketleniyor. Veya sosyal medyadan, bu büyük firmaların algoritmalarını ele alalım. Yapay zeka ile insanların tükettiği içerikler direkt olarak etiketleniyor.

Analiz ve Sonuçlar

Mesela yeni bir ayrılık yaşamış bir insan, genel olarak mutsuz, sağlıksız bir hayat yaşıyan birinin tükettiği içerikler etiket isimlerini bilmemekle beraber sonuç olarak kategorize ediliyor. Ve bu etiketlerin verildiği kullanıcıların karşısına daha fazla basit dopamin reklamları çıktığını biliyoruz; yemek, tatlı vs gibi.

PROBLEM GİDEREK BÜYÜYOR

Günümüzde böyle büyük paraların döndüğü piyasalar olan ama küçük, vizyonsuz marketlerde kullanılsa da aslında bu inanılmaz bir gücü ve potansiyelin varlığına işarettir. Bunu sadece bu kadar basit bir finansal alanda kullanmak ve sadece ona yatırım
yapmak çok büyük bir vizyonsuzluk olurdu. Bu teknolojinin yani bu veri biriktirme, kategorize etme, analiz etme ve öğrenme paternlerinin sağlık, enerji, nüfus ve ülke planlaması, şehir planlamacılığı, dünya kaynakları gibi ve daha birçok önemli şey için kullanıldığını düşünün. İnanılmaz bir gelişme sağlayabilir. Ve bu alanlar keşfedilmeyi bekliyorlar, önlerinde hiçbir engel olmaksızın. Makine öğrenmesi bu analizlerin çok daha uç seviyelerde yapılması demektir. Çünkü artık elde ettiğimiz veriler bizim kavrayabileceğimiz boyutların astronomik katları.Bunu şöyle anlatabiliriz: Bir insan yaklaşık olarak ortamala 200 kelime okur, bu yaklaşık saatte 72kb veri demektir.

VERI MIKTARININ NE KADAR ARTIGINI GÖREBILMEK İÇİN BAZI ÖRNEKLER

01 Amazon yağmur ormanlarında tahminen 700 milyar ağaç bulunmakta ve hepsini kağıda çevirip üzerine yazı yazsak bile bu bilgi 2TB veri olarak bilgisayarlarda tutulabilir.

02 80 yaşındaki bir insanın doğumundan itibaren her saniyesinin HD kalitesinde kaydedilmesi 1TB boyutunda veri yapar.

03 2005 yılına kadar insanlığın toplam ürettiği veri 130 exabyte.
2010 yılına kadar 1200 exabyte
2015 yılına kadar 7900 exabyte
2020'de tahminen 40900 exabyte veri üretildi.

04 İnsan genomu 1GB boyutunda veridir. Bir jet motoru 7 saat içersinde 30 terabyte veri üretebiliyor.

GENEL BİLGİLENDİRME

Uzun lafın kısası problemler inanılmaz bir hızla ve boyut olarakta artan bir hızla büyüyerek artıyor. Bu kadar veriyle başa çıkabilmemiz için. Bu verilerden anlamlı sonuçlar çıkarıp globalleşen dünyanın sorunlarını çözebilmemiz için artık insan kapasitesi yetersiz.

Bizlerin artık makine öğrenmesine, derin öğrenmeye ve yapay zekaya ihtiyacımız var. Her alanda ve olabilecek her şekilde.

Makine öğrenmesi dediğimiz konsept 3 şekilde gerçekleşebilir;

-Denetimli
-Denetimsiz
-Güçlendirilmiş

Denetimli(supervised) makine öğrenmesi ile sonuçlar makine öğrenmesine verilir. Gerçek sonuçlardan model öğrenir ve performansını değerlendirir.

Denetimsiz(unsupervised) yöntem ile sonuç verileri modele verilmez. Sonuçlar için yaptığı tahminler ona göre değerlendirilir.

Güçlendirilmiş(reinforced)öğrenme ile sonuçlar başta verilmez, tahminlerden sonra verilir. Sonuç ona göre değerlendirilir. İki yöntemin karışımı gibidir.

PROJEYE GİRİŞ

SUNUMUN BU KISMINDAN İTİBAREN ÖNÜNÜZDE KODLARIN VE DİĞER DOSYALARIN AÇIK OLARAK OKUNMASINI TAVSİYE EDERİM. BU ŞEKİLDE NELER YAPTIĞIMIZIDAHA İYİ ANLAYABİLİRSİNİZ

Verimiz 9 sütundan ve 769 satırdan oluşuyor. 8 bağımsız ve 1 bağımlı değişkenimiz var(Target). Bu sütunlar sırasıyla; Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age ve Outcome.

İlk olarak bizim için gerekli olan kütüphaneleri importluyoruz. Bunlar:
sys, pandas, matplotlib, numpy, scipy, IPython, sklearn, random, time. Kütüphaneleri bunların neden importlandığı ve kullanım amaçlarını detaylı olarak çıkarttığım raporda anlattım.

Ardından üzerinde çalışacağımız directory'i ayarlıyoruz. ve bazı kullanıcağımız modelleri kütüphanelerden çekiyoruz.

Daha sonra verimiz üzerinde daha rahat çalışabilmek adına verimizin birkaç tane kopyasını farklı değişken adlarıyla alırız. Bunların alınmasının sebepleri; temiz bir çalışma alanı oluşturmak(müsvedde oluşturmak gibi), verilerde oluşturduğumuz değişiklerin verinin orjinal halini bozsun istemiyoruz çünkü.

Ardından verimiz hakkında ön bilgi sahibi oluyoruz ve modelimizi oluşturup eğitmeye başlıyoruz.

[6] Devamında ortalama değer bazlı olarak sürekli değişkenli verimizi binary değişkenli bir veri haline dönüştürüyoruz. Bu sayede kategorizasyon işlemlerini ve dolayısıyla çıkarım yapmamızı kolaylaştırıyoruz.
Sonrasında verimizdeki 0 değerleri mod ve duruma göre medyan değerleriyle dolduruyoruz. Verimizde hiç missing value bulunmuyor ama yine de nasıl yapıldığını görmek ve bilmek önemli. Ne zaman mod ne zaman medyan kullanılıyor? Aradaki farkı bilmek gibi.

Ardından bazı tutarlılık (Accuracy) değerleri elde ederiz.

[10] Burda ise modelin performansını değerlendirmek için confusion matrix'ini kullanırız.

[11] Ardından kullanılan confusion matrix'i görselleştiririz ve beraberinde başka birmodel doğruluğu kontrolü olan ROC eğrisini çizdiririz.

[14] Bazı coefficient ve korelasyon değerlerini elde ederiz ve bu değerler üzerine yorumlar yapabiliriz.

[16] Burda ise direkt olarak bir korelasyon grafiği görüyoruz.

[20] Outcome sonuçlarına göre bir görselleştirme yapıyoruz. Burda yaptığımız şey biraz veri analizi.

[21-22] Burda ise ilk olarak bir aralıklandırma yapıp bu aralıklandırma baz alınıp bir görselleştirme yapılıyor. 22. hücrede ise Z tabanlı örneklem boyutu hesaplaması yapılarak bir örneklem boyutu belirlenip ardından aralıklandırma işlemleri yapılıyor.

Ardından değişkenlerin outcome üzerindeki etkileri görselleştiriliyor ve burdan yorumlar çıkartılıyor.

29 da tüm ilişkileri görebileceğimiz bir pair plot elde ederiz. ardından correlation heatmap ile tüm değişkenler arasındaki korelasyon ilişkilerini gözlemleriz.

Ve modelimizi eğitiriz. Burda birsürü modeli bir liste içine aktarırız ve hepsine aynı veriyi verip performans ölçümü yaparız.

yaptığımız işlemi bir yazı tura örneğiyle basitleştirilmiş şekilde gösteriyoruz.

[35] elimizle bir decision tree modeli üretiyoruz.

Ardından bazı analizler, model değerlendirmeleri, korelasyon haritası he son olarak hard voting, soft voting gibi işlemleri yapıyoruz.

İZLEDİĞİNİZ İÇİN TEŞEKKÜRLER..

CONTACT US

E-mail erenndemir27@gmail.com

Github: https://github.com/Erencode27

Phone +90 553 973 7269