Data Preprocessing

Dr. Cahit Karakuş

∴ Veri düzenlemesi

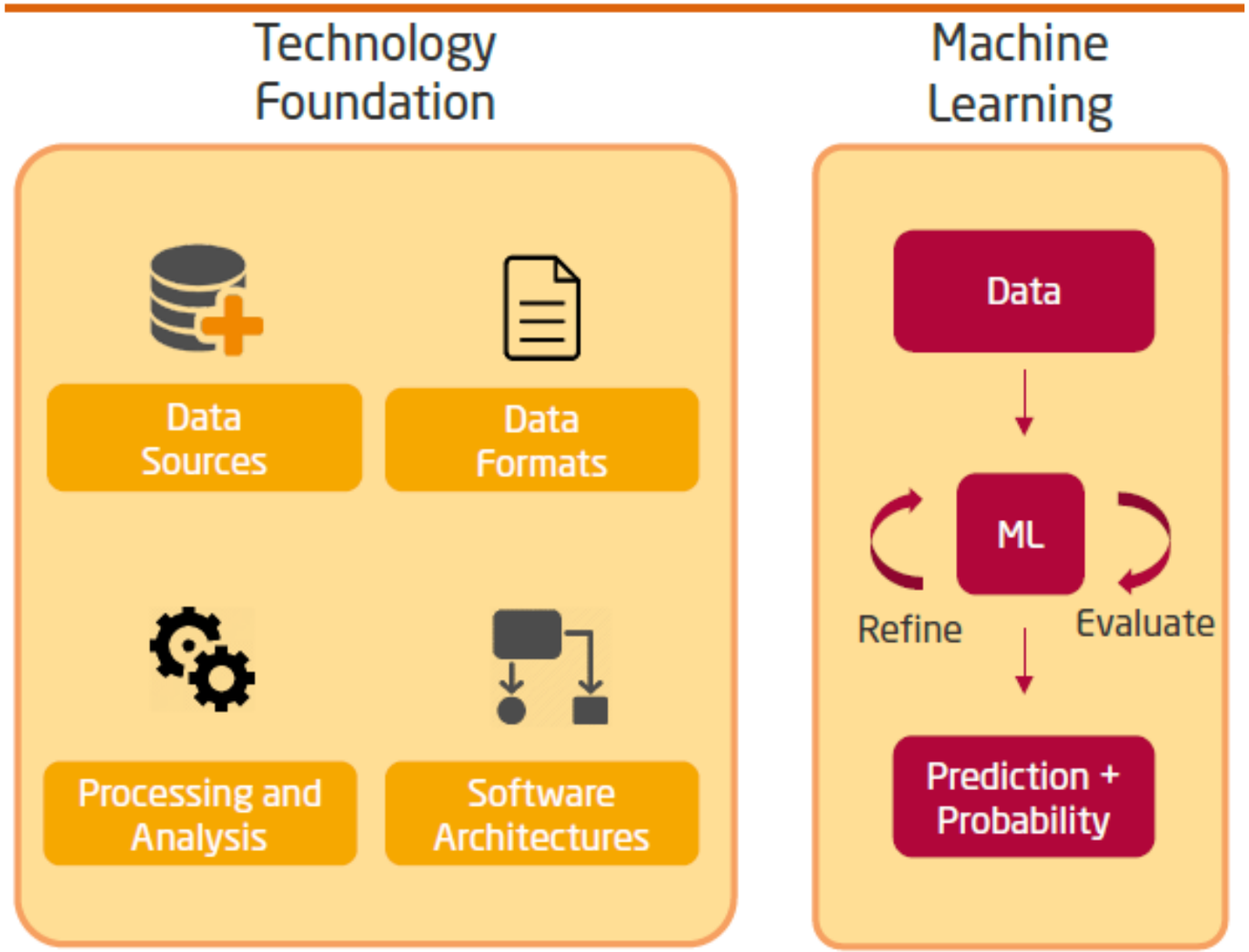
1) Eksen düzlemi değişikliği.
   Kartezyen – Polar.

2) Boyut – 1d, 2, 3
   x(t) $x_{x,y}^{(t)}$ x, y, z

   f[x(t)]
   f(x, y)
   f(x, y, z)

3) Sınıf (Sınıflandırma, kümelendirme)

4) Öğrenmenin başarısı – Performans.


Technology Foundation
Machine Learning
Data Sources
Data Formats
Processing and Analysis
Software Architectures
Data
ML
Refine
Evaluate
Prediction + Probability

Eğitim Verileri

Eğitim Verileri Nedir?

- Makine öğrenimi modelleri, yüksek kaliteli eğitim verilerine erişime dayanır. Verilerinizi etkili bir şekilde nasıl toplayacağınızı, hazırlayacağınızı ve test edeceğinizi anlamak, AI'nın tam değerini ortaya çıkarmanıza yardımcı olur.

Eğitim verileri nedir?

- Daha fazla uygulama ve kullanım için bir temel olarak hareket etmek amacı ile eğitim veri kümesi adı verilen bir başlangıç veri kümesi gerektirir. Bu veri kümesi, programın büyüyen bilgi kitaplığının temelidir. Modelin işleyebilmesi ve ondan öğrenebilmesi için eğitim veri kümesi doğru bir şekilde etiketlenmelidir.

Eğitim Verileri Nedir?

- Makine Öğrenimi algoritmaları verilerden öğrenir.
- Etiketli veri yapıları arasındaki ilişkiler bulurlar, anlayış geliştirirler, kararlar alırlar ve kendilerine verilen eğitim verilerinden güvenlerini değerlendirirler.
- Eğitim verileri ne kadar iyi olursa, model o kadar iyi performans gösterir.
- Aslında, makine öğrenimi eğitim verilerinizin kalitesi ve niceliği, veri projenizin başarısıyla, algoritmaların kendileri kadar ilgilidir.
- İlk olarak, veri kümesi terimiyle ne demek istediğimize dair ortak bir anlayışa sahip olmak önemlidir. Bir veri kümesinin tanımı, her satırın bir gözlem içerdiği hem satırları hem de sütunları olmasıdır.
- Bu gözlem bir resim, bir ses klibi, metin veya video olabilir.

Eğitim Verileri Nedir?

- Şimdi, veri kümenizde çok büyük miktarda iyi yapılandırılmış veri depolamış olsanız bile, aslında modeliniz için bir eğitim veri kümesi olarak çalışacak şekilde etiketlenmemiş olabilir.
- Örneğin, otonom araçlar sadece yol resimlerine ihtiyaç duymazlar, her bir arabanın, yayanın, sokak tabelasının ve daha fazlasının açıklamalı olduğu etiketli resimlere ihtiyaç duyarlar.
- Duygu analizi projeleri, bir algoritmanın birinin argo veya alaycılık kullandığını anlamasına yardımcı olan etiketler gerektirir.
- Chatbotlar, yalnızca ham dile değil, varlık ayıklamaya ve dikkatli sözdizimsel analize ihtiyaç duyar.Başka bir deyişle, eğitim için kullanmak istediğiniz verilerin genellikle zenginleştirilmesi veya etiketlenmesi gerekir.
- Ayrıca, algoritmalarınızı güçlendirmek için daha fazlasını toplamanız gerekebilir. Büyük olasılıkla, sakladığınız veriler makine öğrenimi algoritmalarını eğitmek için kullanılmaya tam olarak hazır değil.

Test kümesi nedir?

- Bir makine öğrenimi algoritması oluşturmak için hem eğitim hem de test verilerine ihtiyacınız var.
- Bir model bir eğitim setinde eğitildikten sonra, genellikle bir test setinde değerlendirilir. Çoğu zaman, bu kümeler aynı genel veri kümesinden alınır, ancak bir algoritmanın güvenini ve doğruluğunu artırmak için eğitim kümesinin etiketlenmesi veya zenginleştirilmesi gerekir.
- Bir veri kümesini test ve eğitim kümelerine nasıl ayırmalısınız?
- Genel olarak, eğitim verileri az çok rastgele bölünürken, önceden bildiğiniz önemli sınıfları yakalamayı garanti eder. Örneğin, çeşitli mağazalardan alınan makbuz görüntülerini okuyabilen bir model oluşturmaya çalışıyorsanız, algoritmanızı tek bir franchise görüntüleri üzerinde eğitmekten kaçınmak isteyeceksiniz. Bu, modelinizi daha sağlam hale getirecek ve fazla takılmasını önlemeye yardımcı olacaktır.

Test kümesi nedir?

- Test verilerimin taraflı olmadığından nasıl emin olabilirim?Şirketler yapay zekayı herkes için daha etik ve etkili hale getirmeye çalışırken bu önemli bir sorudur. Önyargı, AI oluşturma sürecinin birçok aşamasında ortaya çıkabilir, bu nedenle yolun her adımında onu azaltmalısınız.
- Eğitim verilerinizi toplarken, verilerinizin tüm kullanım durumlarını ve son kullanıcıları temsil ettiğinden emin olun. Bu aşamada yanlılık olasılığını azaltmak için, verilerinizi etiketleyen ve model performansını izleyen çeşitli insanlardan oluşan bir grubunuz olduğundan emin olmak isteyeceksiniz. Son olarak, temel performans göstergelerinize ölçülebilir bir faktör olarak önyargıyı dahil edin.

Ne kadar eğitim verisi yeterlidir?

- Ne kadar veriye ihtiyacınız olduğuna dair kesin ve kesin bir kural yoktur. Sonuçta farklı kullanım durumları, farklı miktarlarda veri gerektirecektir.
- Kendinden emin olmak için modelinize ihtiyaç duyduğunuz modeller (kendi kendini süren arabalar gibi) çok miktarda veri gerektirirken, metne dayalı oldukça dar bir duygu modeli çok daha az veri gerektirir.
- Genel bir kural olarak, tahmin ettiğinizden daha fazla veriye ihtiyacınız olacak.

Eğitim verileri ile büyük veriler arasındaki fark nedir?

- Büyük veri ve eğitim verileri aynı şey değildir. Gartner, büyük verileri "yüksek hacimli, yüksek hızlı ve/veya yüksek çeşitlilikte" olarak adlandırır ve bu bilgilerin gerçekten yararlı olması için genellikle bir şekilde işlenmesi gerekir. Eğitim verileri, yukarıda bahsedildiği gibi, AI modellerini veya makine öğrenimi algoritmalarını öğretmek için kullanılan etiketlenmiş verilerdir.

Eğitim Verisine olan ihtiyacın Belirlenmesi

- Ne kadar makine öğrenimi eğitim verisine ihtiyacınız olduğuna karar vermek için oyunda birçok faktör var.
- İlk ve en önemlisi, doğruluğun ne kadar önemli olduğudur. Bir duygu analizi algoritması oluşturduğunuzu varsayalım. Sorununuz karmaşık, evet, ama bu bir ölüm kalım meselesi değil. %85 veya %90 doğruluk sağlayan bir duyarlılık algoritması, çoğu insanın ihtiyaçları için fazlasıyla yeterlidir ve burada yanlış bir pozitif veya negatif, hiçbir şeyi önemli ölçüde değiştirmeyecektir.
- Şimdi, bir kanser tespit modeli mi yoksa kendi kendini süren bir araba algoritması mı? Bu farklı bir hikaye. Önemli göstergeleri kaçırabilecek bir kanser tespit modeli, kelimenin tam anlamıyla bir ölüm kalım meselesidir.
- Tabii ki, daha karmaşık kullanım durumları genellikle daha az karmaşık olanlardan daha fazla veri gerektirir. Yalnızca yiyecekleri tanımlamaya çalışan bir bilgisayar vizyonuna karşı nesneleri tanımlamaya çalışan bir bilgisayar görüşü, genel kural olarak daha az eğitim verisine ihtiyaç duyacaktır. Modelinizin tanımlayabileceğini umduğunuz daha fazla sınıf, daha fazla örneğe ihtiyaç duyacaktır.
- Gerçekten çok fazla yüksek kaliteli veri diye bir şey olmadığını unutmayın. Daha iyi eğitim verileri ve daha fazlası modellerinizi iyileştirecektir. Elbette, daha fazla veri eklemenin marjinal kazanımlarının çok küçük olduğu bir nokta vardır, bu yüzden buna ve veri bütçenize göz kulak olmak istersiniz. Başarı eşiğini belirlemeniz gerekir, ancak dikkatli yinelemelerle bunu daha fazla ve daha iyi verilerle aşabileceğinizi bilin.

Eğitim Verileri Hazırlama

- Gerçek şu ki, çoğu veri dağınık veya eksiktir.
- Örneğin bir resim çekin. Bir makine için bir görüntü sadece bir dizi pikseldir. Bazıları yeşil, bazıları kahverengi olabilir, ancak bir makine, özünde bu piksel koleksiyonunun bir ağaç olduğunu söyleyen bir etikete sahip olana kadar bunun bir ağaç olduğunu bilmez. Bir makine bir ağacın yeterince etiketlenmiş görüntüsünü görürse, etiketlenmemiş bir görüntüdeki benzer piksel gruplarının da bir ağaç oluşturduğunu anlamaya başlayabilir.
- Peki, modelini başarılı olması için ihtiyaç duyulan özelliklere ve etiketlere sahip olması için eğitim verilerini nasıl hazırlarsınız?
- En iyi yol, döngü içinde bir insandır. Ya da daha doğrusu, döngüdeki insanlar.
- İdeal olarak, verileri doğru ve verimli bir şekilde etiketleyebilen çeşitli ek açıklamalardan (bazı durumlarda alan uzmanlarına ihtiyacınız olabilir) yararlanırsınız.
- İnsanlar ayrıca bir çıktıya, örneğin bir görüntünün aslında bir köpek olup olmadığıyla ilgili bir modelin tahminine bakabilir ve bu çıktıyı doğrulayabilir veya düzeltebilir (yani, "evet, bu bir köpek" veya "hayır, bu bir kedi"). Bu, temel gerçeği izleme olarak bilinir ve yinelemeli döngüdeki insan sürecinin bir parçasıdır.
- Eğitim veri etiketleriniz ne kadar doğru olursa, modeliniz o kadar iyi performans gösterecektir. Genellikle zaman alan veri etiketleme süreci için ek açıklama araçları ve kalabalık çalışanlara erişim sağlayabilecek bir veri ortağı bulmak yardımcı olabilir.

Eğitim Verilerini Test Etme ve Değerlendirme

- Tipik olarak, bir model oluştururken etiketli veri kümenizi eğitim ve test kümelerine bölersiniz (ancak bazen test kümeniz etiketlenmemiş olabilir).
- Ve elbette, algoritmanızı birincisi üzerinde eğitir ve ikincisinde performansını doğrularsınız.
- Doğrulama kümeniz size aradığınız sonuçları vermediğinde ne olur? Ağırlıklarınızı güncellemeniz, etiketleri düşürmeniz veya eklemeniz, farklı yaklaşımlar denemeniz ve modelinizi yeniden eğitmeniz gerekecek.Bunu yaptığınızda, aynı şekilde bölünmüş veri kümelerinizle yapmak inanılmaz derecede önemlidir.
- Nedenmiş? Başarıyı değerlendirmenin en iyi yolu budur. Geliştirdiği etiketleri ve kararları ve nerede düştüğünü görebileceksiniz. Farklı eğitim setleri, aynı algoritma üzerinde belirgin şekilde farklı sonuçlara yol açabilir, bu nedenle farklı modelleri test ederken, gerçekten iyileşip iyileşmediğinizi anlamak için aynı eğitim verilerini kullanmanız gerekir.
- Test verileriniz, tanımlamayı umduğunuz her kategoriden eşit miktarda bulunmaz. Basit bir örnek kullanmak gerekirse: bilgisayarlı görme algoritmanız 10.000 köpek ve sadece 5 kedi örneği görüyorsa, kedileri tanımlamada sorun yaşama olasılığı yüksektir. Burada akılda tutulması gereken önemli şey, gerçek dünyada modeliniz için başarının ne anlama geldiğidir. Sınıflandırıcınız sadece köpekleri tanımlamaya çalışıyorsa, kedi tanımlamadaki düşük performansı muhtemelen bir anlaşma kırıcı değildir. Ancak üretimde ihtiyaç duyacağınız etiketlerde model başarısını değerlendirmek isteyeceksiniz.
- İstediğiniz doğruluk düzeyine ulaşmak için yeterli bilgiye sahip değilseniz ne olur? Şansınız, daha fazla eğitim verisine ihtiyacınız olacak. Birkaç bin satır üzerine inşa edilen modeller, genellikle büyük ölçekli iş uygulamaları için başarılı olacak kadar sağlam değildir.

Veri Kaynakları

Veriyi ayıklama

- Yığın içerisinde yinelenenleri bulma ve kaldırma
- Hatalı olanları belirleme
- Gereksiz, anlamsız verileri belirleme
- Sistematik hata kaynakları: Eksik veri, kayıp veri, yanlılık, bilinmezlik, belirsizlik,
- Hata: Önyargı veya sistematik hata, rastgele hatalar, hassasiyet, değişkenlik.
- Mükemmel doğruluk, kesinlik ve belirlilik mümkün asla değildir. Önyargılar genellikle “bilinmeyenlerdir”.
- Güven aralığı önemlidir. Sistematik hataların (önyargıların) yakalanması zordur çünkü genellikle bu hataların farkında olunmaz.

Uygulamalar, Programlar ve Analitik Araçları için Veri Kaynakları

- Sensörler, takipciler, web günlükleri, bilgisayar sistemleri günlükleri ve beslemeleri gibi harici olabilir
- Veri oluşturma programlarından veri sağlayan makineler olabilir.
- Veri kaynakları yapılandırılmış, yarı yapılandırılmış, çok yapılandırılmış veya yapılandırılmamış olabilir.
- Veri kaynakları sosyal medya olabilir (L4 Veri İşleme Katmanı)
- Bir kaynak dahili olabilir.
- Kaynaklar, veritabanı, ilişkisel veritabanı, düz dosya, elektronik tablo, posta sunucusu, web sunucusu, dizin hizmetleri gibi veri havuzları olabilir.

Veri kaynakları

- Metin veya virgülle ayrılmış değerler (CSV) gibi dosyalar olabilir.
- Kaynak, uygulamalar için bir veri deposu olabilir (L4 Veri İşleme Katmanı)

Yapılandırılmış Veri Kaynakları

- SQL Server, MySQL, Microsoft Access veritabanı, Oracle DBMS, IBMDB2, Informix, Amazon SimpleDB veya bir sunucudaki dosya toplama dizini
- Verilere erişim için referanslar sağlayan veri sözlüğü – bir dizi ana arama tablosundan oluşur.

Yapılandırılmamış Veri Kaynakları

- Yüksek hızlı işlemeye ihtiyaç duyan yüksek hızlı ağlar üzerinden dağıtılmış veriler
- Kaynaklar dağıtılmış dosya sistemlerindendir. Kaynaklar, .txt (metin dosyası), .csv (virgülle ayrılmış değerler dosyası) gibi dosya türleridir.
- Veriler, karma anahtar değer çiftleri gibi anahtar değer çiftleri şeklinde olabilir.
- Veriler, e-posta, Facebook sayfaları, twitter mesajları vb. gibi dahili yapılara sahip olabilir.
- Veriler modellemez, ilişkileri, hiyerarşi ilişkilerini veya genişletilebilirlik gibi nesne yönelimli özellikleri ortaya çıkarmaz.

Veri kalitesi

- Referans alınan gerçek dünya yapısını temsil ettiğinde veri kalitesi yüksektir.
- Yüksek kalite, gerekli tüm operasyonları, analizleri, kararları, planlamayı ve bilgi keşfini doğru bir şekilde sağlayan veri anlamına gelir.
- Özellikle yapay zeka uygulamaları için yüksek kaliteli veriler için bir tanım, "şu şekilde beş R'ye sahip veriler: Alaka düzeyi, güncellik, aralık, sağlamlık ve güvenilirlik" olabilir.
- Alaka düzeyi son derece önemlidir.

Veri bütünlüğü

- Kullanılabilir ömrü boyunca verilerde tutarlılık ve doğruluğun korunmasını ifade eder.
- Verileri depolayan, işleyen veya alan yazılım, verilerin bütünlüğünü korumalıdır.

Gürültü

- Gürültü – Veri kalitesini etkileyen faktörlerden biridir.
- Gürültü – doğru (gerçek/gerekli) bilgilerin yanı sıra ek anlamsız bilgiler veren verileri ifade eder.

Aykırılık

- Aykırı Değerler – Kaliteyi etkileyen bir faktör
- Veri kümesine ait değil gibi görünen verileri ifade eder
- Örneğin, beklenen aralığın dışında olan veriler.
- Gerçek aykırı değerlerin veri setinden çıkarılması gerekir, aksi takdirde sonuç küçük veya büyük miktarda etkilenecektir.
- Aykırı değer, eğer gerçekse, hata nedeniyle değil, anormalliği tespit etmede faydalı olabilir, .

Kayıp değerler

- Eksik değer – Veri kalitesini etkileyen bir faktör
- Veri kümesinde görünmeyen verileri ima eder.

Yinelenen Değerler

- Yinelenen değer – Veri kalitesini etkileyen bir faktör
- bir veri kümesinde iki veya daha fazla kez görünen aynı veriyi ifade eder.
- Manipule oyununu belirlemede yinelenen değerler önemli rol oynar. Frekans, sıklık analizi yapılmalıdır. FFT ile eksik data, yanlı data, anomali değerleri belirlenmektedir.

Sapma - Anomali

Hatalar

- Kasdi hatalar. Fark edilmeyen sistematik hatalar. Bireysel kaynaklı hatalar. Yazılım hataları: matematiksel modelleme, algoritma, kodlama; verilerin yalnış girilmesi
- • Sistematik hata : Rasgele, Ölçme hatası, Örnekleme hatası.
- • Eksik Veri
- • Kayıp Veri
- • Yanlılık
- • Bilinmezlik
- • Belirsizlik
- • Hassasiyet
- • Değişkenlik:
- • Önyargı
- • İnterferans: parazit, kaışma egelleme
- • Sapma

Test etme ve Doğrulama

- Test etme, onun hakkında bir şeyler bulmaya çalışmaktır (Colllaborative International Dictionary of English'e göre "kanıtı ortaya koymak; deney yoluyla gerçeği, gerçekliği veya kalitesini kanıtlamak") ve doğrulamak, bir şeyin geçerli olduğunu kanıtlamaktır ("Onaylamak; geçerli kılmak" İşbirlikçi Uluslararası İngilizce Sözlüğü).
- Hem endüstride hem de akademide, dahili süreci geliştirmek için farklı modellerin test edildiği (bir geliştirme seti olarak test seti) ve nihai modelin gerçek kullanımdan önce doğrulanması gereken model olduğu düşünülerek bazen birbirlerinin yerine kullanılırlar. görünmeyen bir veri (doğrulama seti).
- "Makine öğrenimi literatürü genellikle 'doğrulama' ve 'test' kümelerinin anlamını tersine çevirir. Bu, yapay zeka araştırmalarına yayılan terminolojik karışıklığın en bariz örneğidir. Bununla birlikte, korunması gereken önemli kavram şudur: Son küme, ister test ister doğrulama olarak adlandırılsın, yalnızca son deneyde kullanılmalıdır.
- Daha kararlı sonuçlar elde etmek ve eğitim için tüm değerli verileri kullanmak için, bir veri seti tekrar tekrar birkaç eğitim ve bir doğrulama veri setine bölünebilir. Bu işleme çapraz doğrulama olarak bilinir . Model performansını doğrulamak için, normalde çapraz doğrulamadan elde edilen ek bir test veri seti kullanılır.

Why Data Preprocessing?

Veri Ön İşleme'ye Giriş

- Veri Entegrasyonu
- Veri Dönüşümleri-Veri Ayrıklaştırma-Veri Kodlama
- Veri temizleme
- Veri Azaltma

Why Data Preprocessing?

- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data. e.g., occupation=“ ”
  - noisy: containing errors or outliers. e.g., Salary=“ 10”
  - inconsistent: containing discrepancies in codes or names
    - e.g., Age=“42” Birthday=“03/07/1997”
    - e.g., Was rating “1,2,3”, now rating “A, B, C”
    - e.g., discrepancy between duplicate records

Why Data Preprocessing?

- Eksik veriler gelebilir
  - Toplandığında geçerli olmayan veri değerleri
  - Verilerin toplandığı ve analiz edildiği zaman arasında farklılıklar, çelişkiler.
  - İnsan/donanım/yazılım sorunları
- Gürültülü veriler (yanlış değerler) aşağıdakilerden gelebilir:
  - Hatalı veri toplama araçları
  - Veri girişinde insan veya bilgisayar hatası
  - Veri aktarımındaki hatalar
- Tutarsız veriler şunlardan gelebilir:
  - Farklı veri kaynakları
  - İşlevsel bağımlılık ihlali (ör. bazı bağlantılı verileri değiştirin)
- Yinelenen kayıtlar ayrıca veri temizliğine ihtiyaç duyar

Why Is Data Preprocessing Important?

- Kaliteli veri yok ise kaliteli madencilik yok!
  - Kaliteli kararlar kalite verilerine dayanmalıdır. örneğin, mükerrer veya eksik veriler, yanlış ve hattalı verile yanıltıcı istatistiklere neden olabilir.
  - Veri ambarı, kaliteli verilerin tutarlı bir şekilde entegrasyonuna ihtiyaç duyar
- Veri çıkarma, temizleme ve dönüştürme, bir veri ambarı oluşturma işinin çoğunu içerir.

Veri Ön İşlemedeki Başlıca Görevler

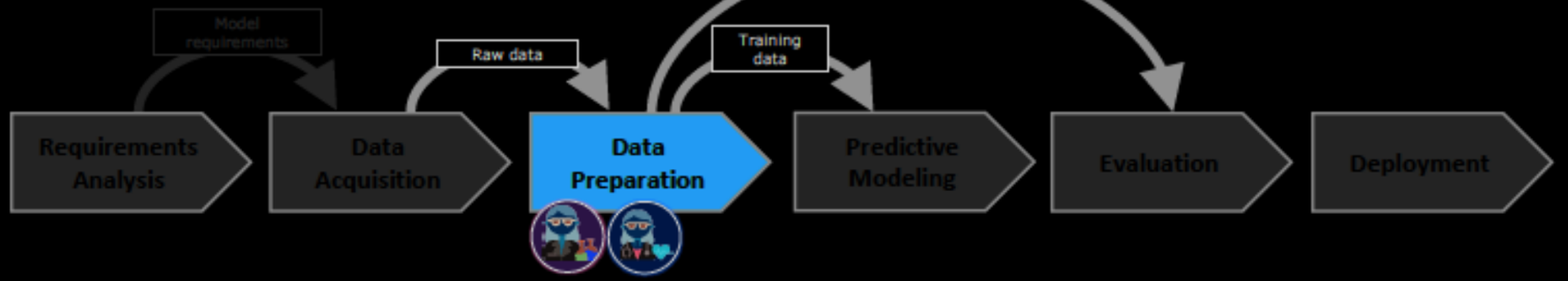
- Veri entegrasyonu: Birden çok veritabanının, veri küpünün veya dosyanın entegrasyonu
- Veri dönüşümü: Normalleştirme ve Toplama. Kodlama ve Binning.
- Veri temizleme: Eksik değerlerin doldurulması, gürültülü verilerin düzeltilmesi, aykırı değerlerin belirlenmesi veya kaldırılması ve tutarsızlıkları çözülmesi.
- Veri azaltma: Hacim olarak azaltılmış temsil elde eder ancak aynı veya benzer analitik sonuçları üretir

Veri Hazırlama

Veri Ön işleme

- Veri İşleme Mimarisindeki L2 alım katmanında önemli bir adım
- Bir Makine Öğrenimi (ML) algoritması ve Analytics çalıştırmadan önce gerekir
- Veriler bir veri deposuna veya bulut hizmetine aktarılmadan önce gerekli
- veri depolamadan, analitik uygulamasından, hizmetten veya buluttan Aktarım Formatları.

Data Preparation



Data Preparation

- Exploration
- Quality assessment
- Cleansing
- Labeling
- Imputation
- Feature engineering

Roles — Data Scientist — Domain Expert — (Data) Engineer

Veri Hazırlama Nedir?

- Veri hazırlama, modelin tahmin yeteneğini artırabilir veya bozabilir de!
- Veri hazırlama, eğitim seti verilerinin eklenmesi, silinmesi veya dönüştürülmesi işlemidir.
- Bazen verilerin ön işlenmesi, model doğruluğunda beklenmeyen gelişmelere yol açabilir.
- Veri hazırlama önemli bir adımdır ve model doğruluğunda istenen bu artışı elde edip edilemeyeceğini görmek için veriler için uygun olan veri ön işleme adımları denemelidir.

Veri Ön İşleme İhtiyaçları

- (i) Aralık dışı, tutarsız ve aykırı değerler
- (ii) Güvenilmez, alakasız ve gereksiz bilgileri filtreleme
- (iii) Veri temizleme, düzenleme, azaltma ve/veya tartışma
- (iv) Veri doğrulama, dönüştürme veya kod dönüştürme
- (v) ELT işleme: Ayıkla, Yükle ve Dönüştür . ELT processing: Extract, Load and Transform.
- (vi) Zenginleştirme, Düzenleme, Tartışma, Azaltma

Veri Hazırlama Adımları

- Verileri nasıl temizlerim? Veri Temizleme
- Doğru verileri nasıl sağlarım? Veri Dönüşümü
- Verileri nasıl dahil eder ve ayarlarım? Veri Entegrasyonu
- Verileri nasıl birleştirir ve ölçeklendiririm? Veri ormalleştirme
- Eksik verileri nasıl halledebilirim? Eksik Veri Düzeltme
- Gürültüyü nasıl algılar ve yönetirim? Gürültü Tanımlama

Veri Hazırlama Süreci

Verileri bir makine öğrenmesi algoritması için hazırlama süreci özetlenebilir:

- 1. Adım: Veri Seçilir
- 2.Adım: Veriler Ön İşleme tabii tutulur.
- 3. Adım: Verileri Dönüştürülür.
- Bu süreci doğrusal bir şekilde takip edilir.

Veri Seçme

- Mevcut olan tüm verileri dahil etmek için her zaman güçlü bir istek vardır, "daha fazlası daha iyidir" özdeyişi tutacaktır.
- Doğru olabilir de olmayabilir de.
- Üzerinde çalıştığınız soruyu veya sorunu ele almak için gerçekte hangi verilere ihtiyacınız olduğu düşünülür.
- Düşünmenize yardımcı olacak sorular:
  - Sahip olduğunuz verilerin kapsamı nedir?
  - Elinizde olmasını istediğiniz hangi veriler mevcut değil?
  - Sorunu çözmek için hangi verilere ihtiyacınız yok?

Verileri Ön İşleme

Daha İyi Veriler > Daha İyi Algoritmalar

- **Biçimlendirme:** Seçilen veriler uygun bir biçimde olmayabilir
- **Temizleme**: Eksik verilerin kaldırılması veya düzeltilmesi
  - Sorunu çözmek için veri taşınmaz ve tamamlanmaz.
  - Hassas bilgiler anonimleştirilir veya kaldırılır.
  - Verilerin eksik, yanlış, hatalı, alakasız kısımları belirlenir.
- **Örnekleme:** Gerekenden daha fazla seçilmiş veri mevcut
  - Algoritmalar için daha uzun çalışma süreleri
  - Daha büyük hesaplama ve bellek gereksinimleri
  - Tüm veri setini değerlendirmeden önce daha küçük temsili örnek alınır.

Verileri Ön İşleme

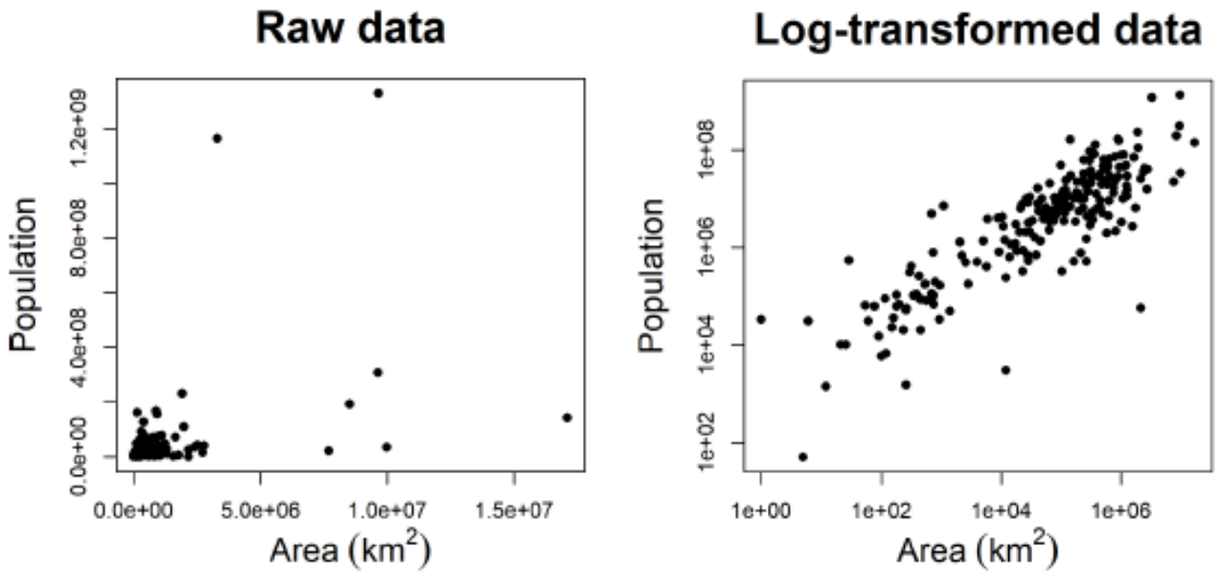
Rasgele (dummy) Değişkenler

- Kategorik (kuşku bırakmayan, açık, kesin) öznitelik sayısal özniteliğe dönüştürülür.
- Her öznitelik 0 veya 1 değerine sahip olacaktır.
- Tam Rasgele Değişkenler: Her düzey için bir değişken olmak üzere n rasgele değişken kullanarak n kategori temsil edilir.
- Referans Gruplu Rasgele Değişkenler: Kategorik değişken n-1 rasgele değişken kullanarak n kategoriyle temsil edilir.
- Referans Gruplu Sıralı Kategorik Değişken için Rasgele Değişkenler: Matematiksel sıralamalar Küçük < Orta < Büyük olarak varsayılır.
- Sıralamayı belirtmek için daha yüksek kategoriler için daha fazla 1'ler kullanılır.

Verileri Ön İşleme

Dönüştürülmüş Nitelikler

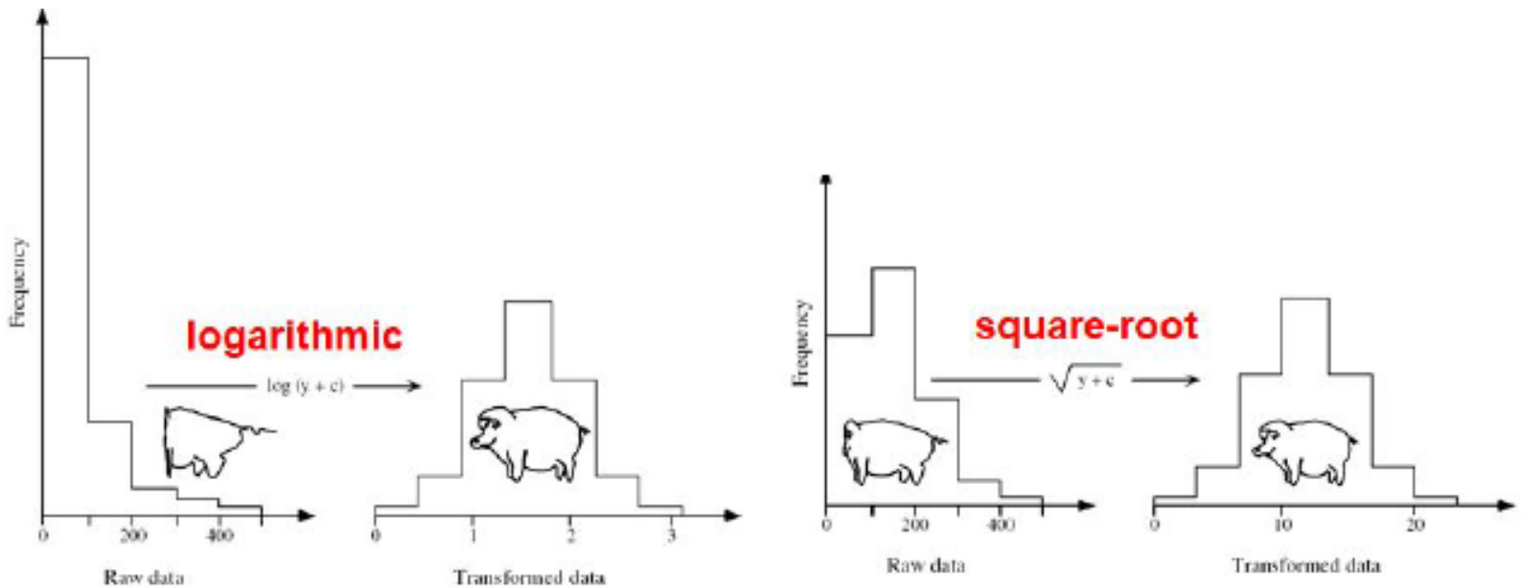
- Veri dönüştürmede, bireysel değerler arasındaki göreli farklılıklar olacak şekilde değiştirilir.
- Dönüşüm türleri:
  - Doğrusal: Sabitler eklenerek veya sabitle çarpılarak
  - Doğrusal olmayan: log dönüşümü, karekök dönüşümü vb. alınır.



Verileri Ön İşleme

Dönüştürülmüş Nitelikler

- log dönüşümü aşırı sağa eğik veriler için uygundur, sqrt dönüşümü biraz sağa eğik veriler için uygundur.



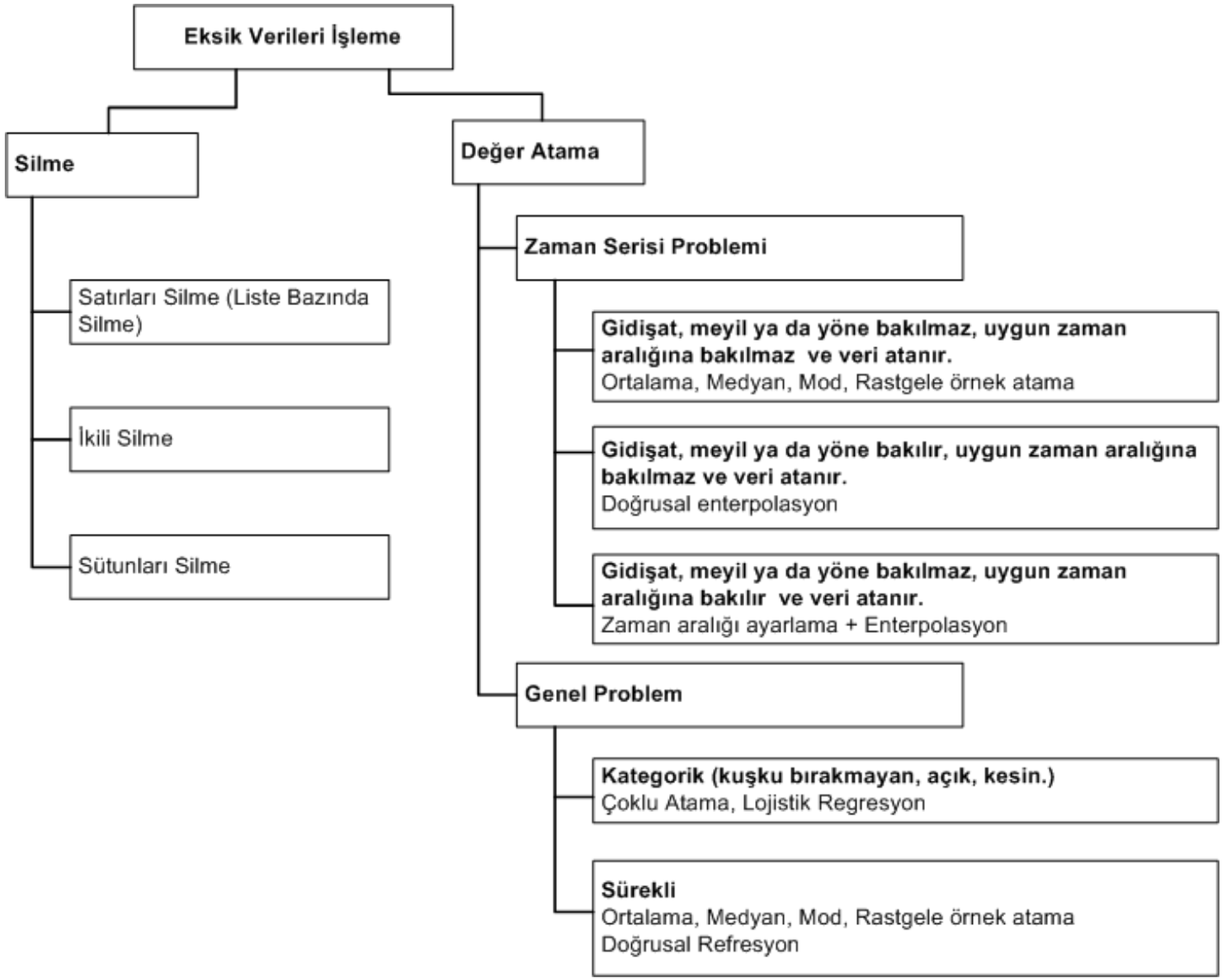
Verileri Ön İşleme

Eksik Veriler Nasıl İşlenir?

- Eksik verilerle başa çıkmanın iyi bir yolu YOKTUR!
- Sorunun türüne bağlı olarak veri ataması için farklı çözümler: Fouier Dönüşümü, Zaman Serisi Analizi, ML, Regresyon vb.
- Genel çözüm yok

Verileri Ön İşleme

Eksik Veriler Nasıl İşlenir?



Varyans - Standart Sapma

- Aritmetik Ortalama: Alınan örnekleme değerlerinden bir ya da iki tanesi çok yüksek ya da düşük olursa aritmetik ortalama davranışın eğilimini yansıtmaz.
- Varyans - Standart Sapma: Standart sapma, değerlerin aritmetik ortalamasından kaynaklanan kök ortalama karesi (RMS) sapmasıdır. Olasılık ve istatistikte, bir olasılık dağılımının standart sapması, rasgele değişken veya popülasyon veya değerlerin yayılmasının bir ölçüsüdür. Genellikle σ harfi ile belirtilir (küçük harf sigma). Standat sapma, varyansın karekökü olarak tanımlanır. Varyans, veriler ile aritmetik ortalama farklarının karlerinin toplamıdır. Ölçülen verilerin ortalamaya yayılmasını ölçer. Standart sapma, aritmetik ortalamadan olan sapmayı verir.
- Veri değerleri aritmetik ortalamaya yakınsa, standart sapma küçüktür. Ayrıca, birçok veri noktası ortalamanın uzağındaysa, standart sapma büyüktür. Tüm veri değerleri eşitse, standart sapma sıfırdır.
- Bir veri dağılımındaki değişimin önemli bir ölçüsü varyanstır. Varyansın karekökü alınarak standart sapma elde edilir.
- Standart sapma dizideki herbir değerin aritmetik ortalamaya yakınlığını gösterir. Standart sapmanın küçük olması ortalamalarda sapmaların ve riskin az olduğunu, standart sapmanın büyük olması ortalamalarda sapmaların ve riskin çok olduğunu gösterir.

Veri Tahmini/Atama (Ortalama/Medyan) Değerleri

- Bir sütundaki eksik olmayan değerlerin ortalaması/medyanı hesaplanır.
- Artıları:
  - Kolay ve Hızlı
  - Küçük sayısal veri kümeleriyle iyi çalışır
- Eksileri:
  - Özellikler arasındaki korelasyonları etkilemez. Yalnızca sütun düzeyinde çalışır.
  - Kodlanmış kategorik özelliklerde kötü sonuçlar verir (kategorik özelliklerde KULLANMAYIN)
  - çok doğru değil
  - Tahminlerdeki belirsizliği hesaba katmaz

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

Veri Tahmini/Atama (En Sık) veya (Sıfır/Sabit) Değerler

- Eksik değerleri yüklemek için en sık kullanılan istatistiksel strateji
- Eksik verileri her sütunda en sık görülen değerlerle değiştirme
- Sıfır veya Sabit atama, eksik değerler sıfır veya belirtilen herhangi bir sabit değerle değiştirilir
- Artıları:
  - Kategorik özelliklerle iyi çalışır

- Eksileri:
  - Ayrıca özellikler arasındaki korelasyonları da etkilemez.
  - Verilerde önyargı oluşturabilir

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

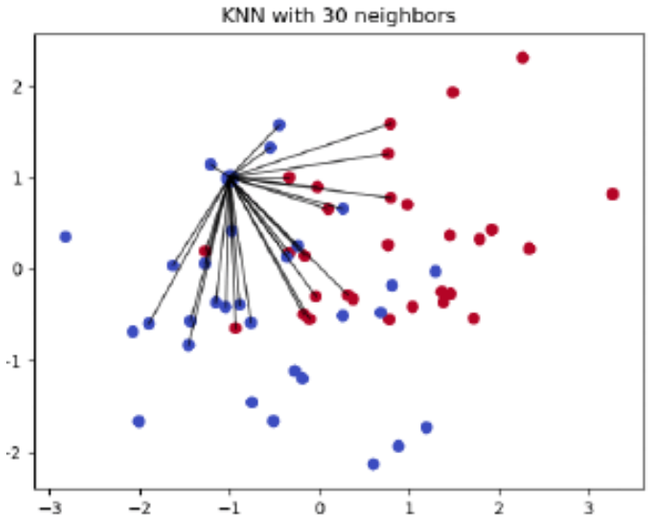
df.fillna(0)

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	0.0
1	9	0.0	9.0	0	7.0
2	19	17.0	0.0	9	0.0

Verileri Ön İşleme

Veri Tahmini/Atama: k-NN

- K-en yakın komşular basit sınıflandırma için kullanılan bir algoritmadır
- Algoritma, herhangi bir yeni veri noktasının değerlerini tahmin etmek için 'özellik benzerliğini' kullanır
- Yeni noktaya, eğitim kümesindeki noktalara ne kadar benzediğine bağlı olarak bir değer atanır.

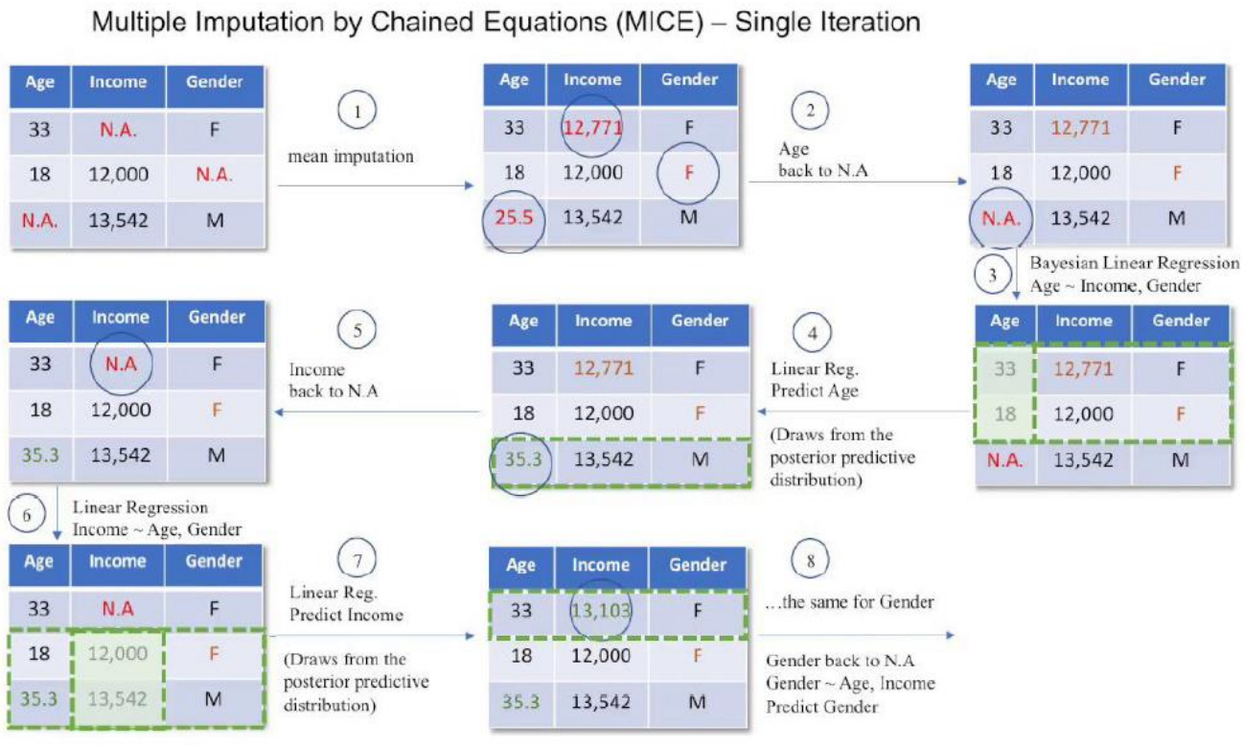


- Artıları:
  - Ortalama, medyan veya en sık kullanılan atama yöntemlerinden çok daha doğru olabilir (Veri kümesine bağlıdır)

- Eksileri:
  - Hesaplamalı olarak pahalı.
  - KNN, tüm eğitim veri setini bellekte saklayarak çalışır.
  - K-NN, verilerdeki aykırı değerlere karşı oldukça hassastır (SVM'den farklı olarak)

Veri Tahmini / Atama: Çok Değişkenli Atama

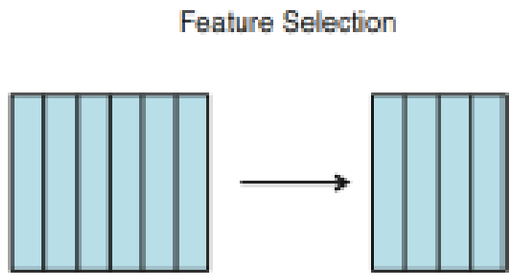
- Eksik verilerin birden çok kez doldurulması.
- Çoklu atamalarlar , eksik değerlerin belirsizliğini daha iyi bir şekilde ölçtüğü için tek bir atamadan çok daha iyidir.
- Zincirli denklemler yaklaşımı da çok esnektir ve farklı veri tiplerinin farklı değişkenlerini işleyebilir.

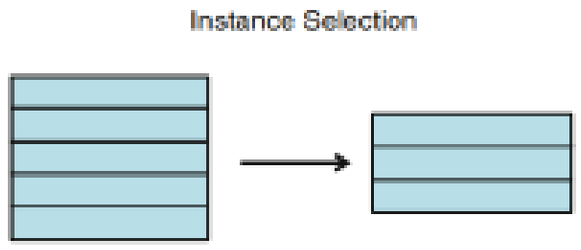


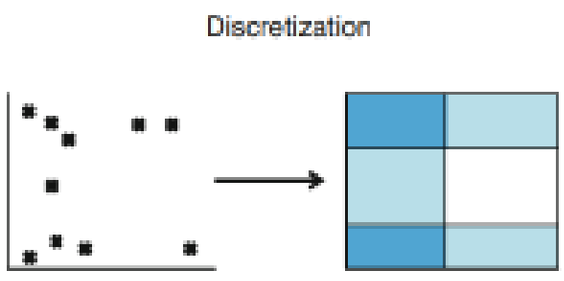
Verileri Ön İşleme

Veri Azaltma

- Verilerin boyutsallığı nasıl azaltılabilir? Özellik Seçimi (Feature Selection - FS)
- Gereksiz ve/veya çelişkili örnekler nasıl kaldırılır? Örnek Seçimi (Instance Selection - IS)
- Bir özniteliğin etki alanı nasıl basitleştirilir? Ayrıklaştırma (Discretization)
- Verilerdeki boşluklar nasıl doldurulur? Özellik Çıkarma ve/veya Örnek Oluşturma (Feature Extraction and/or Instance Generation)

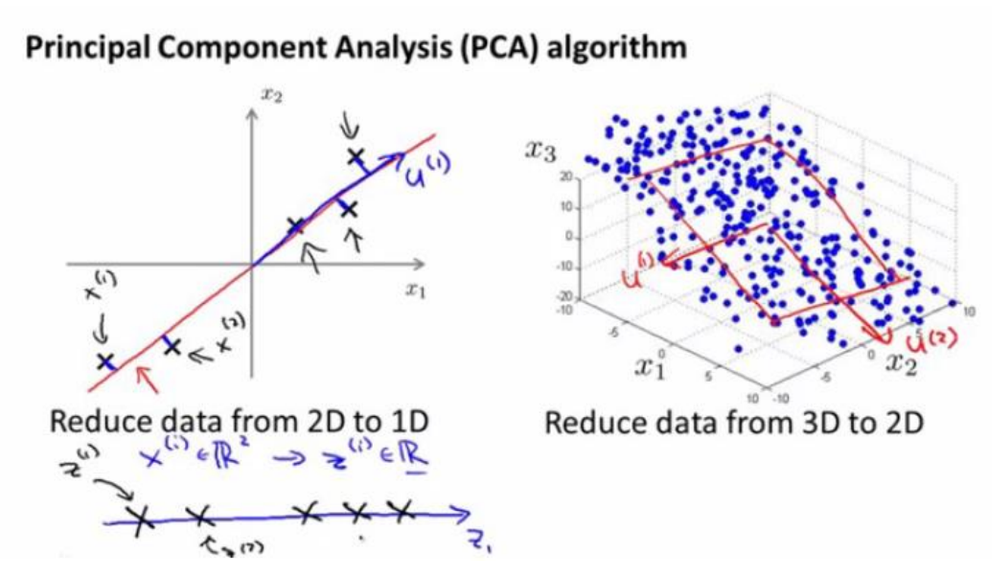






Verileri Ön İşleme
Projeksiyon: Temel Bileşen Analizi (PCA -nPrincipal Component Analysis)

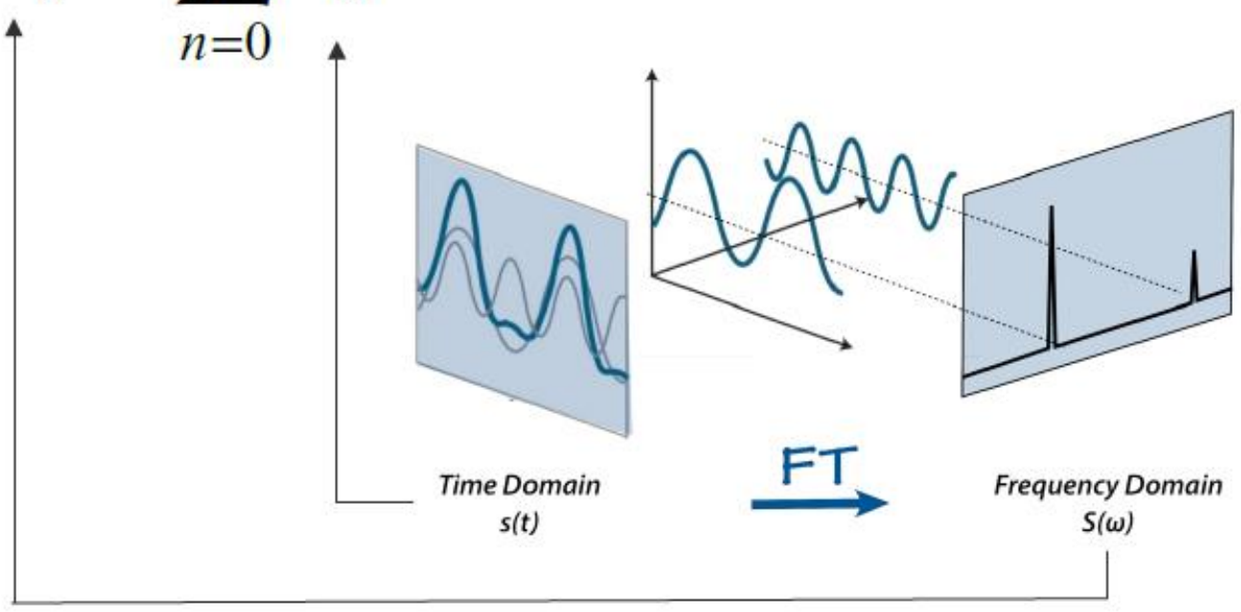
- Dünyada veri miktarı arttıkça, makine öğrenimi geliştirme için kullanılabilen veri kümelerinin boyutu da büyüyor
- Boyutsallık azaltma, herhangi bir önemli bilgiyi kaybetmeden bazı boyutların atılmasını kolaylaştıracak şekilde verilerin yeni boyutlara dönüştürülmesini içerir.
- Büyük ölçekli problemler, görselleştirilmesi çok zor olabilen çeşitli boyutları beraberinde getirir.
- Daha iyi bir görselleştirme için bu boyutlardan bazıları kolayca düşürülebilir.



Verileri Ön İşleme

Discrete Fourier Transform

$$X_k = \sum_{n=0}^{N-1} x_n e^{-i 2\pi \frac{k}{N} n}$$

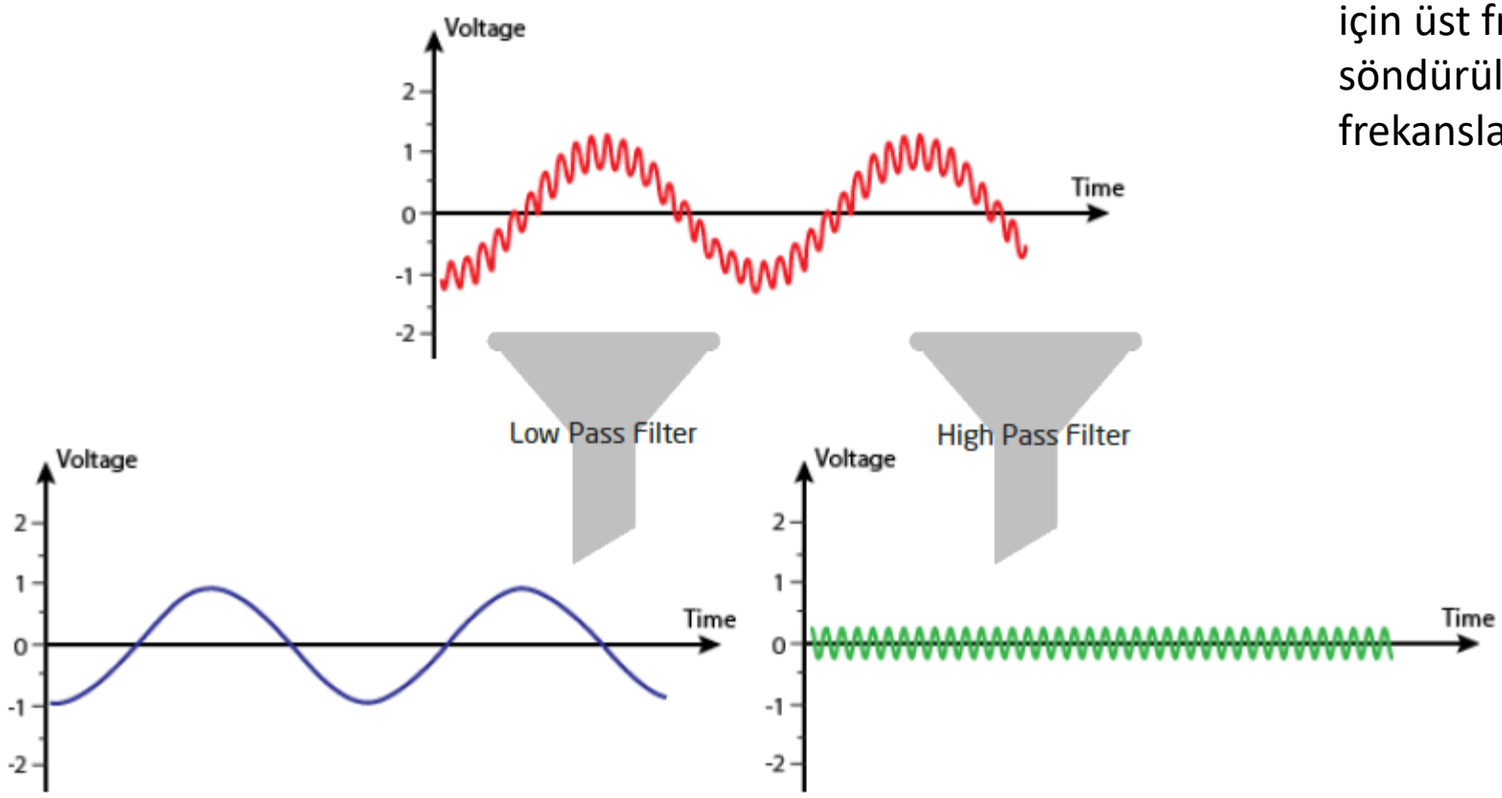


Fourier series in 1822

Verileri Ön İşleme

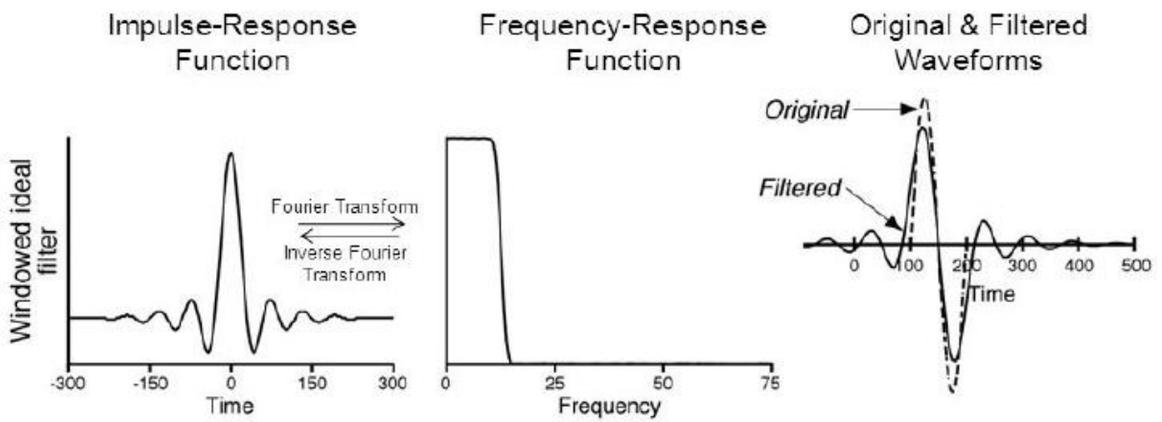
Filter

FFT ile frekans domenindeki değişimleri elde edilir. LPF için üst frekanslar söndürülür. HPF için alt frekanslar söndürülür.



Fourier Transformation

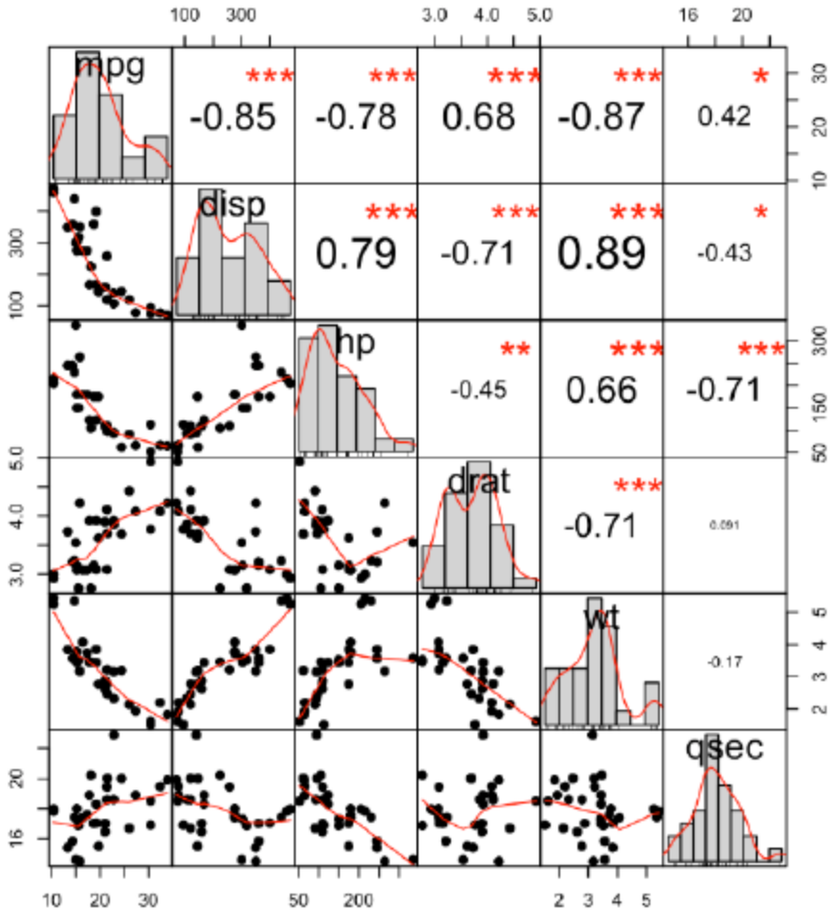
- Önemli sinyal işleme aracıdır.
- Bir sinyali sinüs ve kosinüs bileşenlerine ayrıştırmak için kullanılır.
- Dönüşümün çıktısı, Fourier veya frekans alanındaki sinyali temsil eder.
- Belirli frekans alanlarını çok kolay bir şekilde ortadan kaldırmak için matematiksel işlemleri uygulanır.
- Orijinal zaman sinyalini kurtarmak için ters Fourier dönüşümünü uygulanır.



Verileri Ön İşleme

Korelasyon

- Veri kümenizdeki birden çok değişken ve nitelik arasındaki ilişkiyi anlamanın yolu
- Korelasyonu kullanarak, aşağıdakiler gibi bazı öngörüler elde edebilirsiniz:
  - Bir veya birden fazla nitelik diğerine bağlıdır
  - Bir veya birden fazla nitelik, diğer niteliklerle ilişkilendirilir
- Bir özelliği diğerinden tahmin etmede yardımcı olabilir (eksik değerleri atfetmenin harika bir yolu)
- (bazen) nedensel bir ilişkinin varlığını gösterebilir

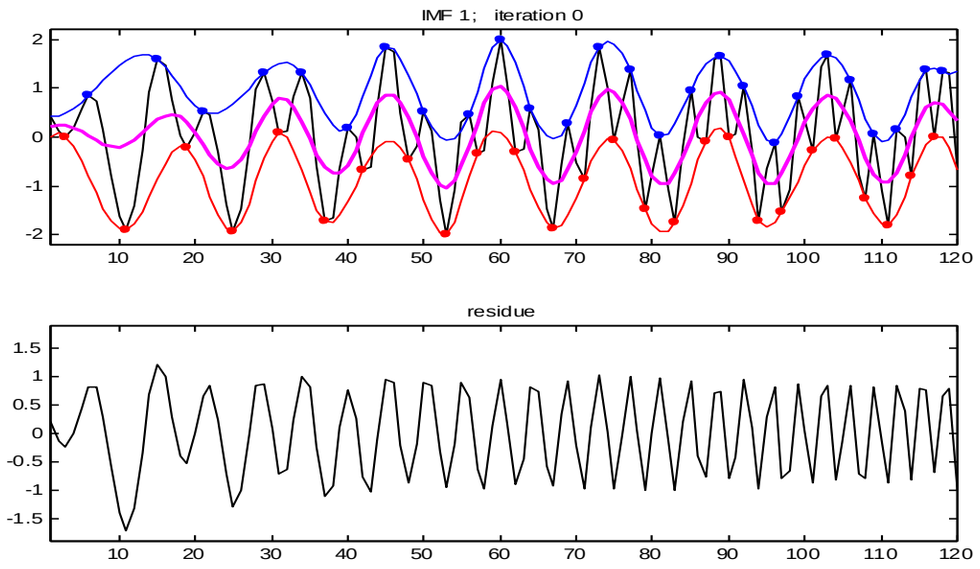


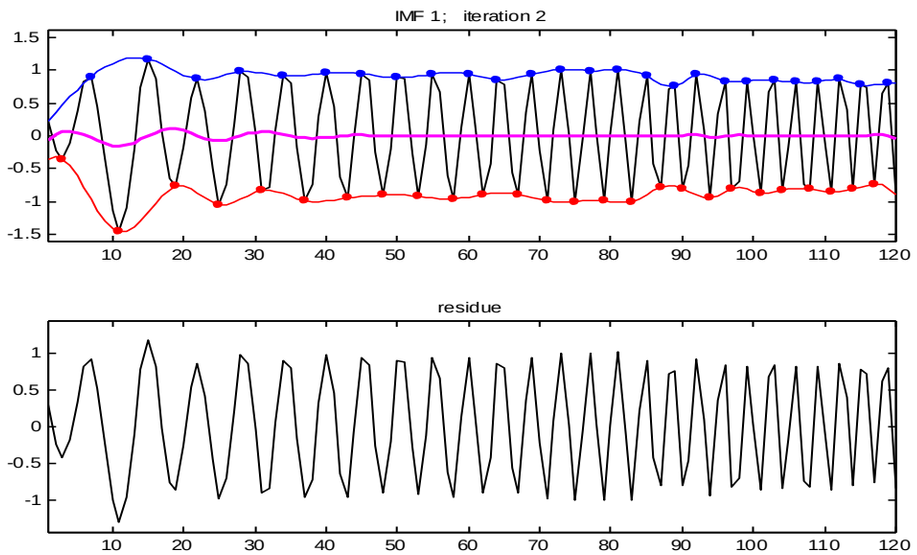
Verileri Ön İşleme

Otokorelasyon

- Zaman serisi analizi ve tahmininde yoğun olarak kullanılır
- Bir zaman serisinin gecikmeli değerleri arasındaki ilişkinin ölçüsü
- Verilerdeki gizli kalıpları ortaya çıkarılır
- Zaman serisi verilerimizdeki mevsimsellik ve eğilim belirlenir

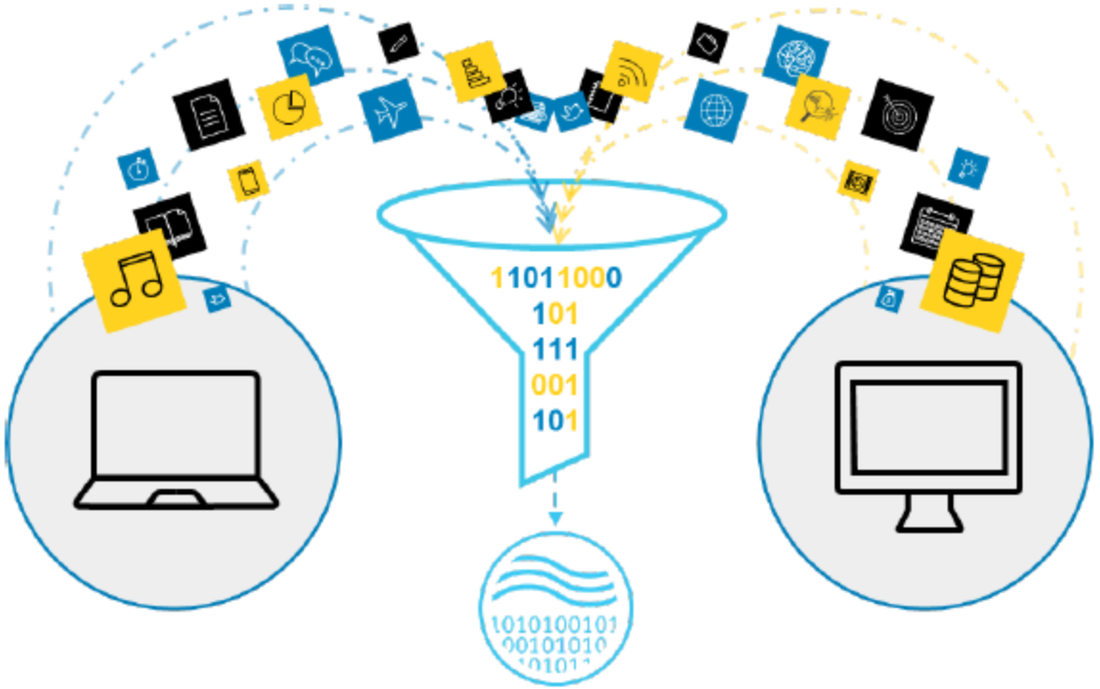
Hilbert Huang Transform





Transform Data

- Scaling: The preprocessed data may contain attributes with a mixtures of scales for various quantities. Many machine learning methods like data attributes to have the same scale
- Decomposition: There may be features that represent a complex concept that may be more useful to a machine learning method when split into the constituent parts, Example -> Date
- Aggregation: There may be features that can be aggregated into a single feature

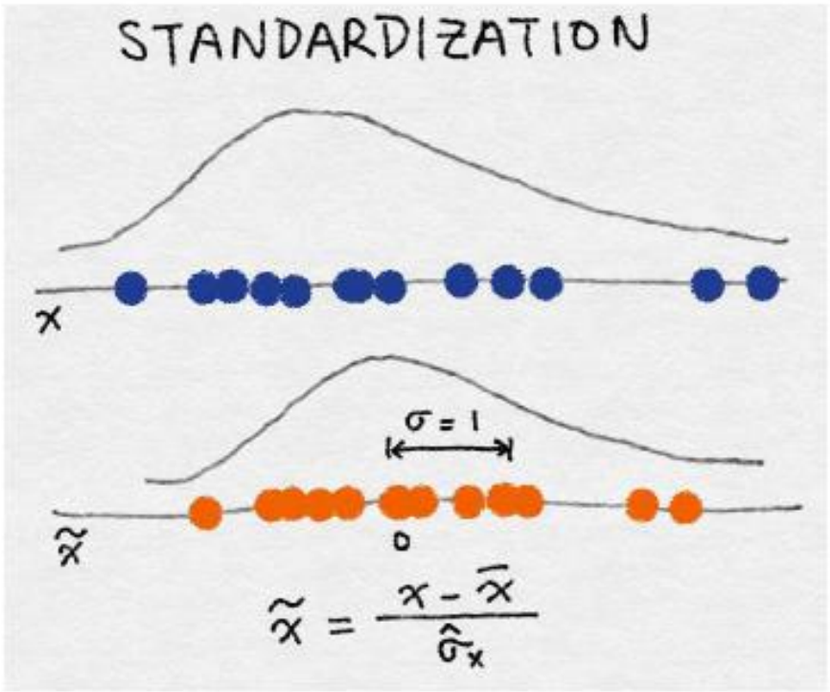


Transform Data

Standardization (Variance Scaling)

- Özelliğin ortalamasını (tüm veri noktalarından) çıkarır ve varyansa böler
- Ayrıca varyans ölçekleme olarak da adlandırılabilir, sonuçta ölçeklenen özelliğin ortalaması 0 ve varyansı 1'dir.
- Orijinal özelliğin bir Gauss dağılımı varsa, ölçeklenen özellik de öyledir

$$\tilde{x} = \frac{x - \text{mean}(x)}{\text{sqrt}(\text{var}(x))}$$

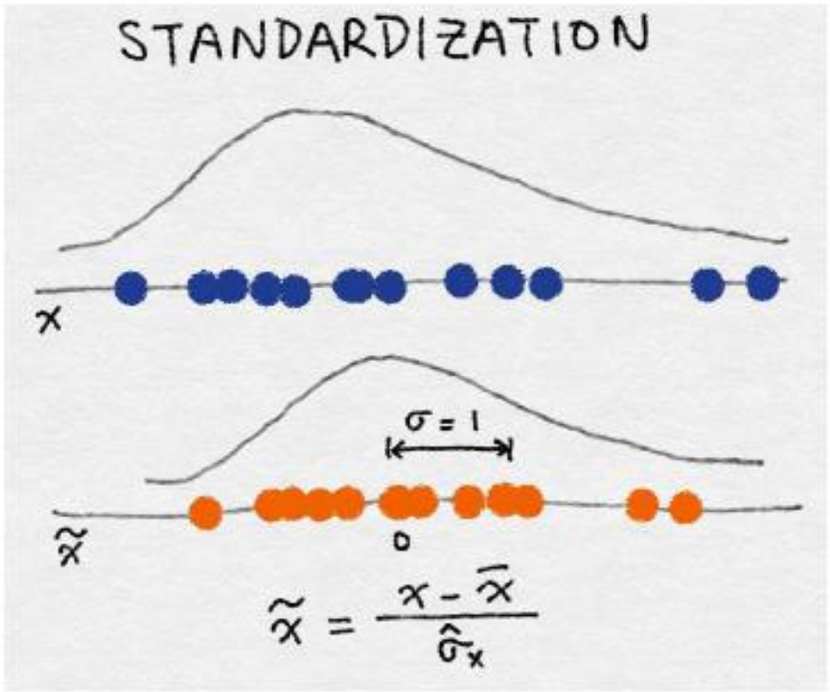


Transform Data

Min-Max Scaling

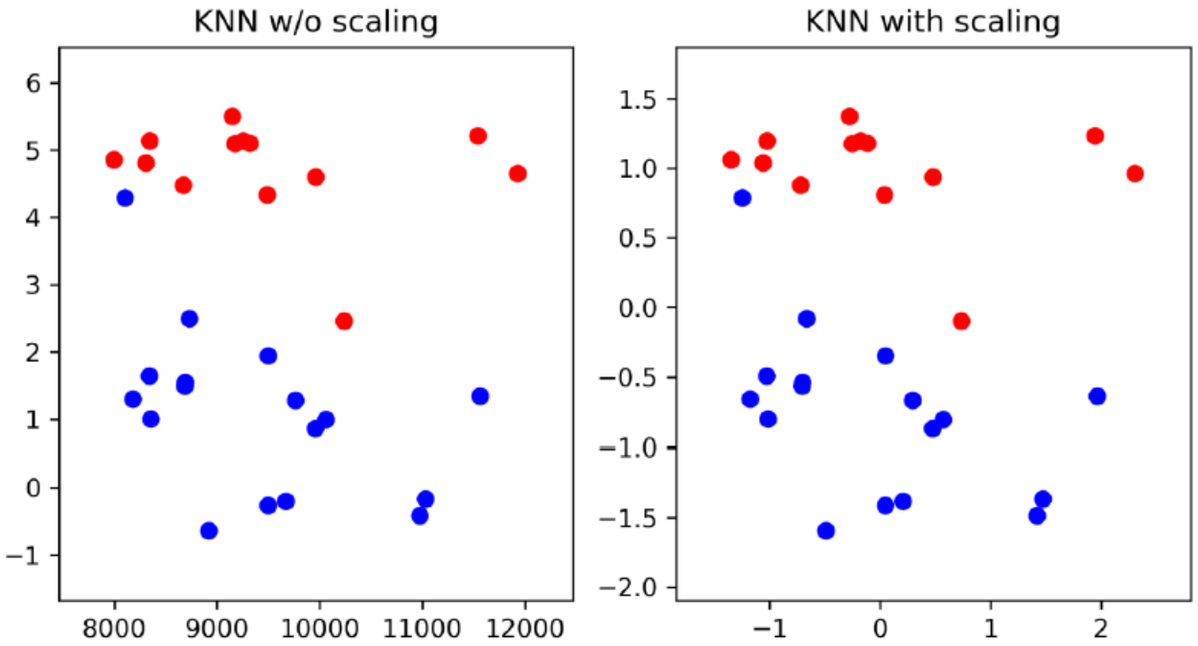
- Let x be an individual feature value (i.e., a value of the feature in some data point)
- min (x) and max (x), respectively, be the minimum and maximum values of this feature over the entire dataset
- Min-max scaling squeezes (or stretches) all feature values to be within the range of [0, 1]

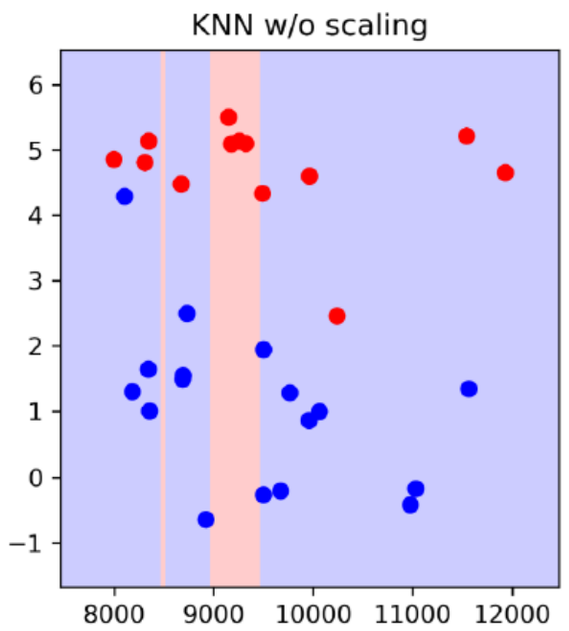
$$\tilde{x} = \frac{x - \text{mean}(x)}{\text{sqrt}(\text{var}(x))}$$

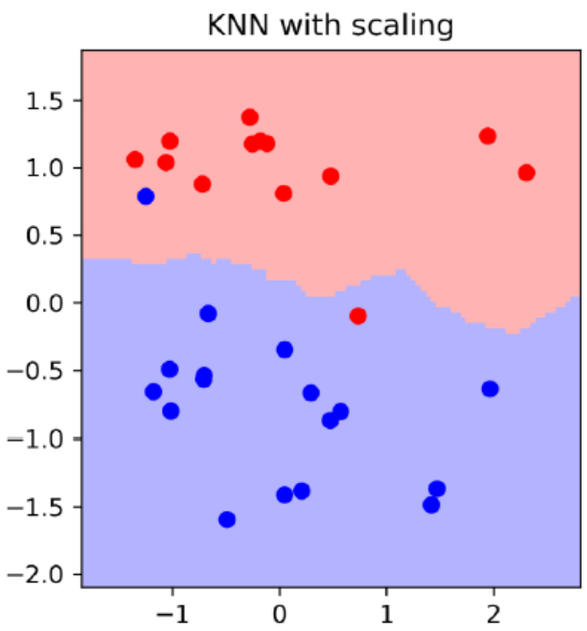


Transform Data

Why Scaling?



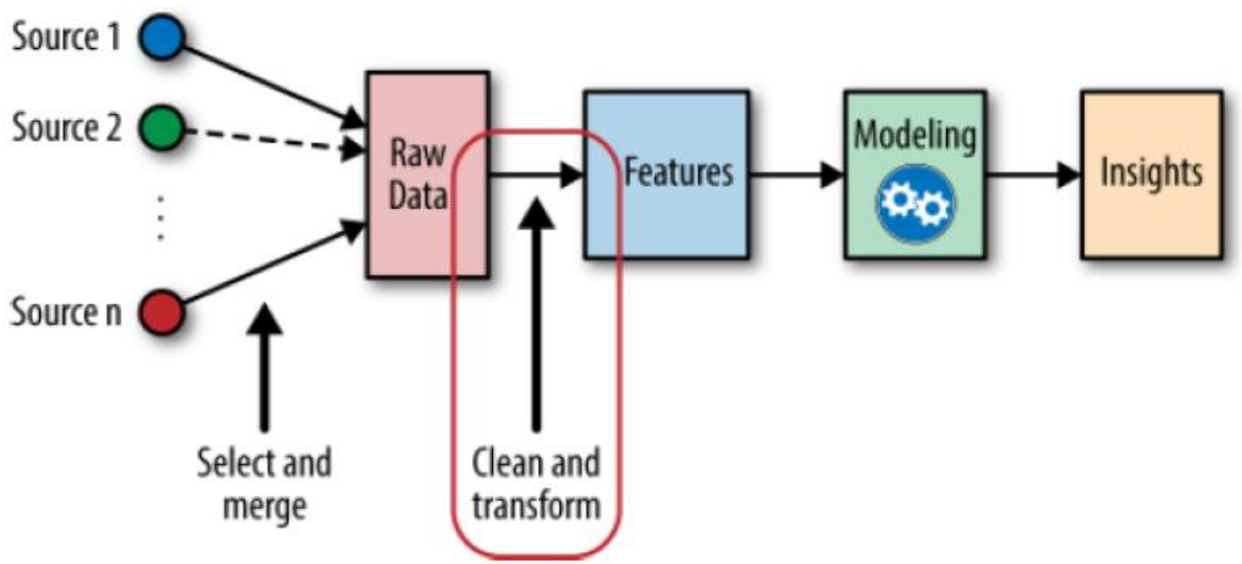




Transform Data

Feature Engineering

- Coming up with features is difficult, time-consuming, requires expert knowledge. "Applied machine learning" is basically feature engineering.
- The features you use influence more than everything else the result. No algorithm alone, to my knowledge, can supplement the information gain given by correct feature engineering.
- Good data preparation and feature engineering is integral to better prediction.



Transform Data

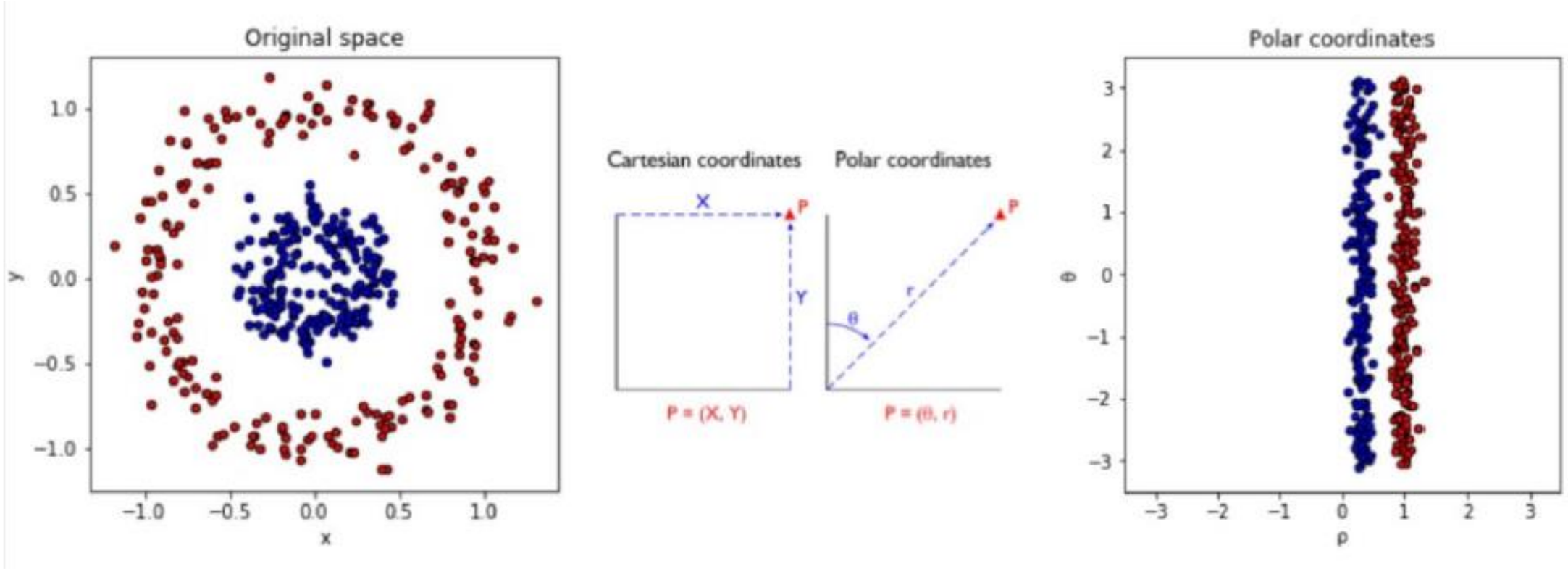
Feature Engineering

- Verilerin anlaşılması ve işlenmesi zor olabilir
- Makine öğrenimi modellerimiz için verilerin okunmasını kolaylaştırmak için özellik mühendisliği yapılır
- Özellik Mühendisliği, verilen verileri yorumlanması daha kolay bir forma dönüştürme işlemidir.
- Genel olarak: Veri ile ilgili arka planı olmayan kişiler için hazırlanan veri görselleştirmesinin daha sindirilebilir olması için özellikler oluşturulabilir.
- Farklı modeller genellikle farklı veri türleri için farklı yaklaşımlar gerektirir.

Transform Data

Feature Engineering

Example: Coordinate Transformation



Transform Data

Özellik Mühendisliğinin Yinelemeli Süreci

- Beyin fırtınası özellikleri: Gerçekten problemin içine girilir, birçok veriye bakılır, diğer problemler üzerinde özellik mühendisliği incelenir ve neler alınabileceği görülür.
- Özellikler geliştir: Sorununa bağlıdır, ancak otomatik özellik çıkarma, manuel özellik oluşturma ve ikisinin karışımı kullanılabilir.
- Özelliklerin seçilmesi: Modellerinizin üzerinde çalışacağı bir veya daha fazla "görünüm" hazırlamak için farklı özellik önem puanlamalarını ve özellik seçim yöntemlerini kullanılır.
- Modellerin değerlendirilmesi: Seçilen özellikleri kullanarak görünmeyen verilerde model doğruluğu tahmin edilir.

Transform Data

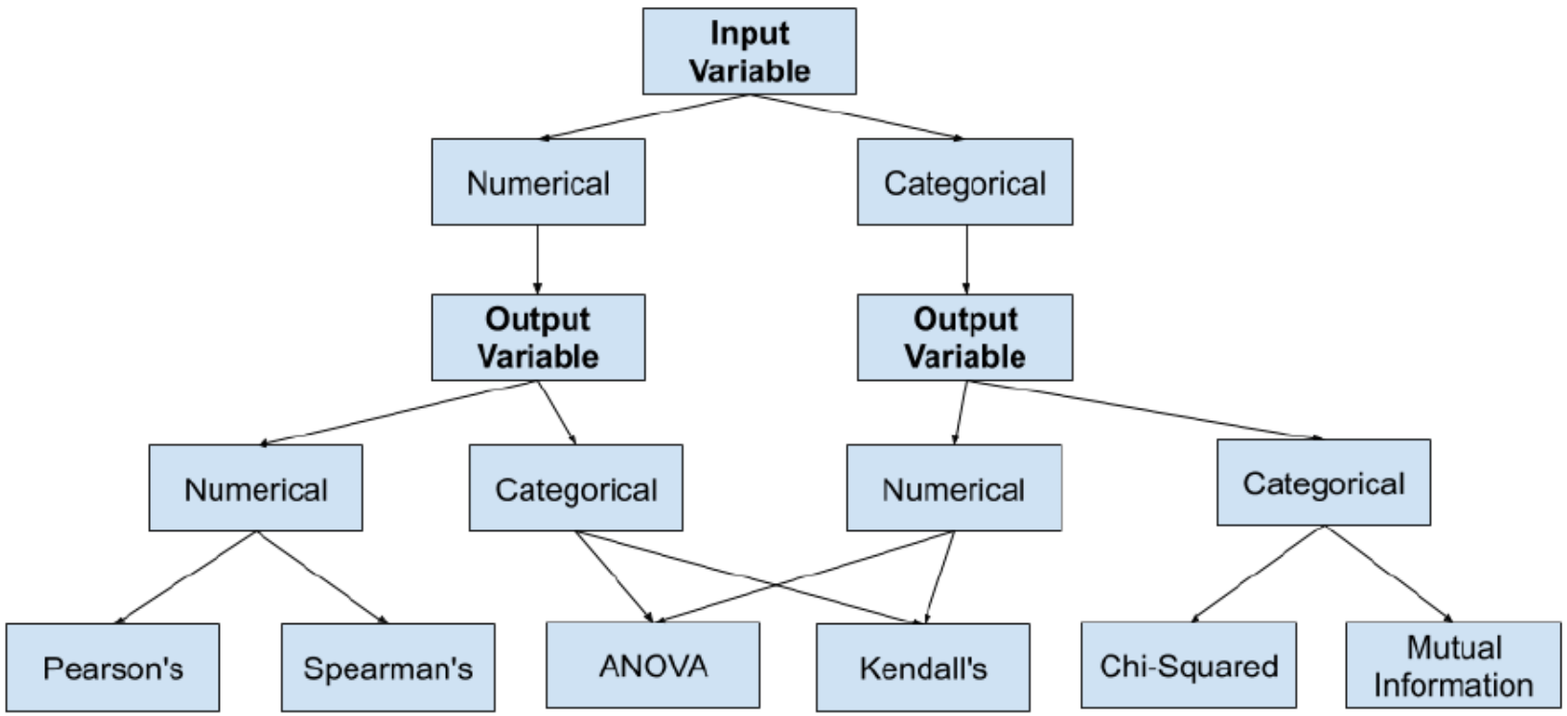
Özellik Mühendisliğinin Yönleri

- Özellik Seçimi: En kullanışlı ve ilgili özellikler mevcut verilerden seçilir.
- Özellik Çıkarma: Mevcut özellikler, daha kullanışlı özellikler geliştirmek için birleştirilir.
- Özellik Ekleme: Yeni veriler toplanarak yeni özellikler oluşturulur.
- Özellik Filtreleme: Modelleme adımını kolaylaştırmak için alakasız özellikler filtrelenir.

Transform Data

Öznitelik Seçimi

- Verilerde, ilgilenilen tahmin değişkenine veya çıktıya en çok katkıda bulunan özelliklerin otomatik olarak seçildiği süreç.
- Verilerde alakasız özelliklerin olması, birçok modelin, özellikle lineer ve lojistik regresyon gibi lineer algoritmaların doğruluğunu azaltabilir.



Usage Notes

- A lot of slides are adopted from the presentations and documents published on internet by experts who know the subject very well.
- I would like to thank who prepared slides and documents.
- Also, these slides are made publicly available on the web for anyone to use
- If you choose to use them, I ask that you alert me of any mistakes which were made and allow me the option of incorporating such changes (with an acknowledgment) in my set of slides.

Sincerely,

Dr. Cahit Karakuş

cahitkarakus@gmail.com