Bölüm

10

Model Seçimi Doğrulama

Model Seçimi

Model Seçimi

- Model Seçimi Nedir?
- Model Seçiminin Önemi
- Model Seçme Kriterleri

Model Seçimi Nedir?

- Makine öğrenmesinde model seçimi, bir problem için en uygun modeli belirleme sürecidir.
- Bu süreç, veri setinin özelliklerine, problemin doğasına ve hedeflenen sonuçlara bağlı olarak yapılır.

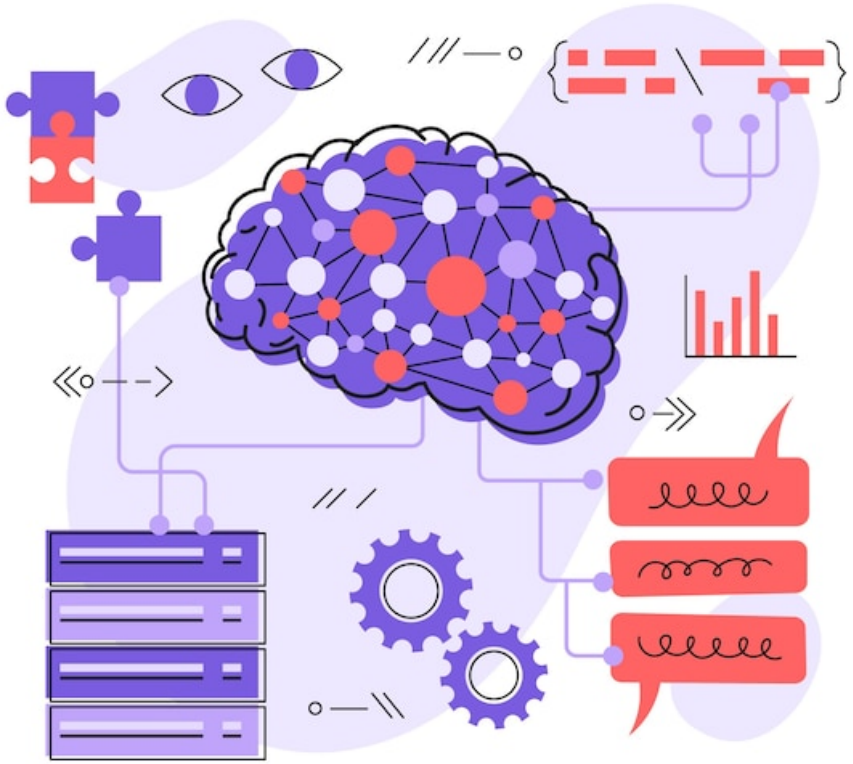
Model Seçiminin Önemi

- Makine öğrenmesinde model seçimi önemlidir.
  - Performansı Etkiler
  - Genelleme Yeteneğini Etkiler
  - Hesaplama Kaynaklarını Etkiler
  - Yorumlanabilirlik ve Açıklanabilirlik
  - Uygulanabilirlik

Model Seçme Kriterleri

- En uygun modeli seçmek, çeşitli yöntemlerin ve alan uzmanlığının bir karışımını gerektirir.
- Basit veya karmaşık bir model, geleneksel veya derin öğrenme modeli, doğrusal veya doğrusal olmayan bir model arasındaki seçim, birkaç faktöre bağlıdır.
  - Problem Türü
  - Veri Erişilebilirliği ve Kalitesi
  - Bilgisayar Kaynakları
  - Açıklanabilirlik
  - Ölçeklenebilirlik
  - Çoklu Girişleri İşleme Yeteneği
  - Zaman Kısıtları
  - Alan Bilgisi

Model Seçme Yöntemleri

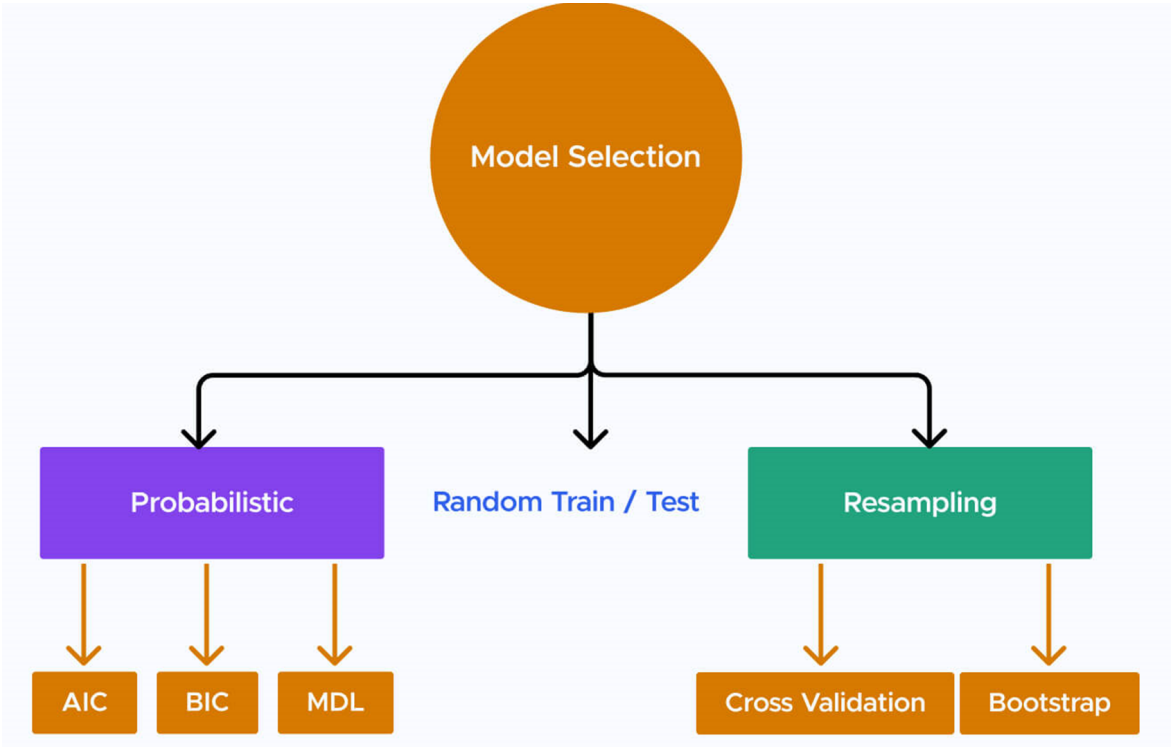


Model Seçme Yöntemleri

- Model Seçme Yöntemleri
  - Olasılıksal
  - Rastgele Bölünme
  - Örnekleme

Model Seçme Yöntemleri

- Makine öğrenmesinde model seçimi, bir problem için en uygun modeli belirleme sürecidir.
- Bu süreç, veri setinin özelliklerine, problemin doğasına ve hedeflenen sonuçlara bağlı olarak yapılır.



https://censius.ai/blogs/machine-learning-model-selection-techniques#blogpost-toc-0

Olasılıksal

- Bilgi Kriteri, istatistiksel prosedürlerin etkinliğini değerlendirmek için kullanılan bir tür olasılıksal ölçümdür.
- Yöntemleri, En Büyük Olabilirlik Tahmini'nin (MLE) log-olabilirlik çerçevesini kullanarak en etkili aday modelleri seçen bir puanlama sistemini içerir.
- Örnekleme sadece model performansına odaklanırken, olasılıksal modelleme hem model performansına hem de karmaşıklığına odaklanır.
- Karmaşıklık derecesini ve belirli bir modelin veri kümesine ne kadar iyi uyduğunu hesaplamak için üç istatistiksel yöntem vardır:
  - Akaike Bilgi Kriteri (Akaike Information Criterion - AIC)
  - Minimum Açıklama Uzunluğu (Minimum Description Length - MDL)
  - Bayesian Bilgi Kriteri (Bayesian Information Criterion - BIC)

Akaike Bilgi Kriteri

- Amaç, bir modelin ne kadar iyi bir şekilde verileri açıkladığını ve ne kadar karmaşık olduğunu dengelemektir.

$$AIC = -2\ln(L) + 2k$$

Minimum Açıklama Uzunluğu

- Karmaşıklık ve uyum arasındaki dengeyi sağlar.
- MDL, veri setinin kodlanması ve modelin karmaşıklığı arasındaki toplam bit sayısını en aza indirmeyi amaçlar.

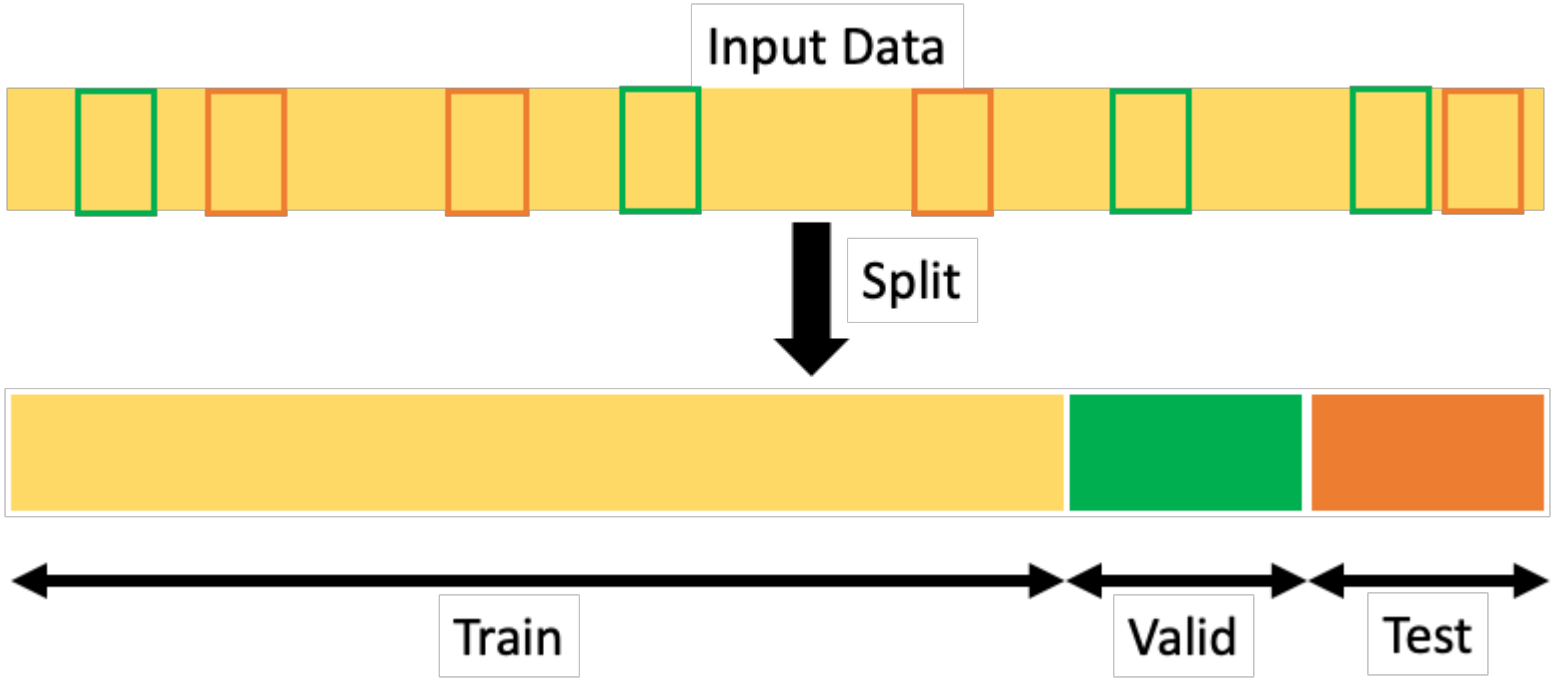
Bayesian Bilgi Kriteri

- AIC gibi, model karmaşıklığı ve uyum arasındaki dengeyi sağlamak için kullanılır.

$$\mathrm{BIC} = k \ln(n) - 2 \ln(\widehat{L})$$

Rastgele Bölünme

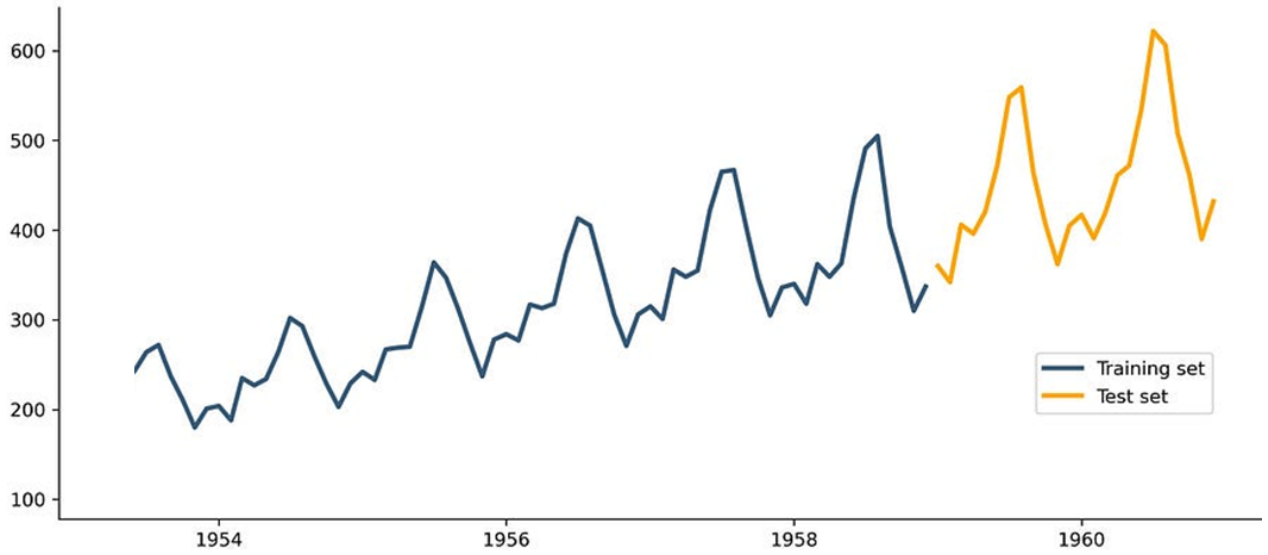
- Bölme işlemi rastgele veya zamana dayalı olabilir.
- Rastgele bölme yöntemi, eğitim verisini eğitim, test ve doğrulama kümelerine rastgele böler.
- Bu işlem, modelin performansını çeşitli test kümelerinde kontrol etmek ve güvenilirliğini gözlemlemek için tekrarlanır.



https://towardsdatascience.com/how-to-split-data-into-three-sets-train-validation-and-test-and-why-e50d22d3e54c

Rastgele Bölünme

- Zaman tabanlı bölme genellikle zaman bileşeni içeren veriler için yapılır, örneğin hava durumu veya borsa verileri gibi.



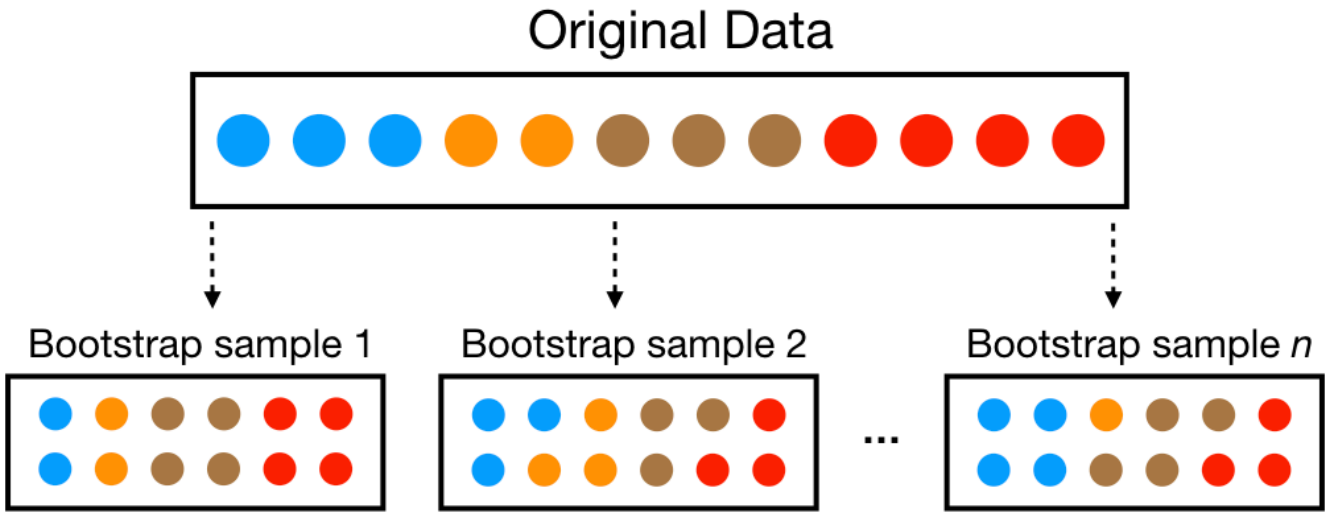
https://towardsdatascience.com/time-series-from-scratch-train-test-splits-and-evaluation-metrics-4fd654de1b37

Örnekleme Yöntemleri

- Örnekleme yöntemleri, modelin eğitim almadığı veri örnekleri üzerinde performansını görmek için veri örneklerini yeniden düzenlemenin basit yöntemleridir.
- Başka bir deyişle, örnekleme, modelin genelleme yeteneğini belirlememizi sağlar.
- Örnekleme yöntemleri:
  - Bootstrap
  - Çarpraz Doğrulama
    - K-Fold Çapraz Doğrulama
    - Leave-One-Out (LOO) Çapraz Doğrulama

Bootstrap

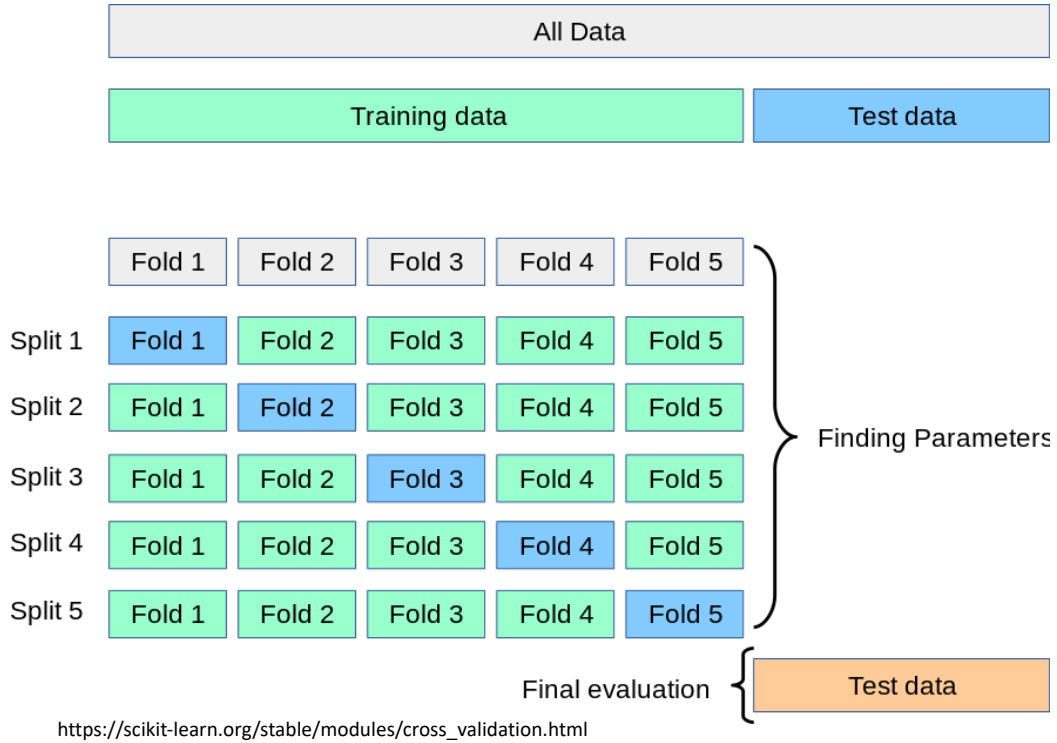
- Bu teknik, bir veri setinden rastgele örneklemler çekmek suretiyle tekrarlanan örnekleme yaparak, popülasyon üzerinde istatistiksel sonuçların tahmin edilmesini sağlar.
- Bootstrap yöntemi, genellikle küçük veri setleri veya sınırlı veriye sahip durumlarda parametrik olmayan istatistiklerin güven aralıklarını ve standart hatalarını hesaplamak için kullanılır.
  - Veri setinden rastgele örneklemler çekilir.
  - Örneklemler üzerinde istatistikler hesaplanır.
  - Bu adımlar birçok kez tekrarlanır.
  - Elde edilen istatistiklerin dağılımı incelenir.



https://bradleyboehmke.github.io/HOML/process.html

K-Fold Çapraz Doğrulama

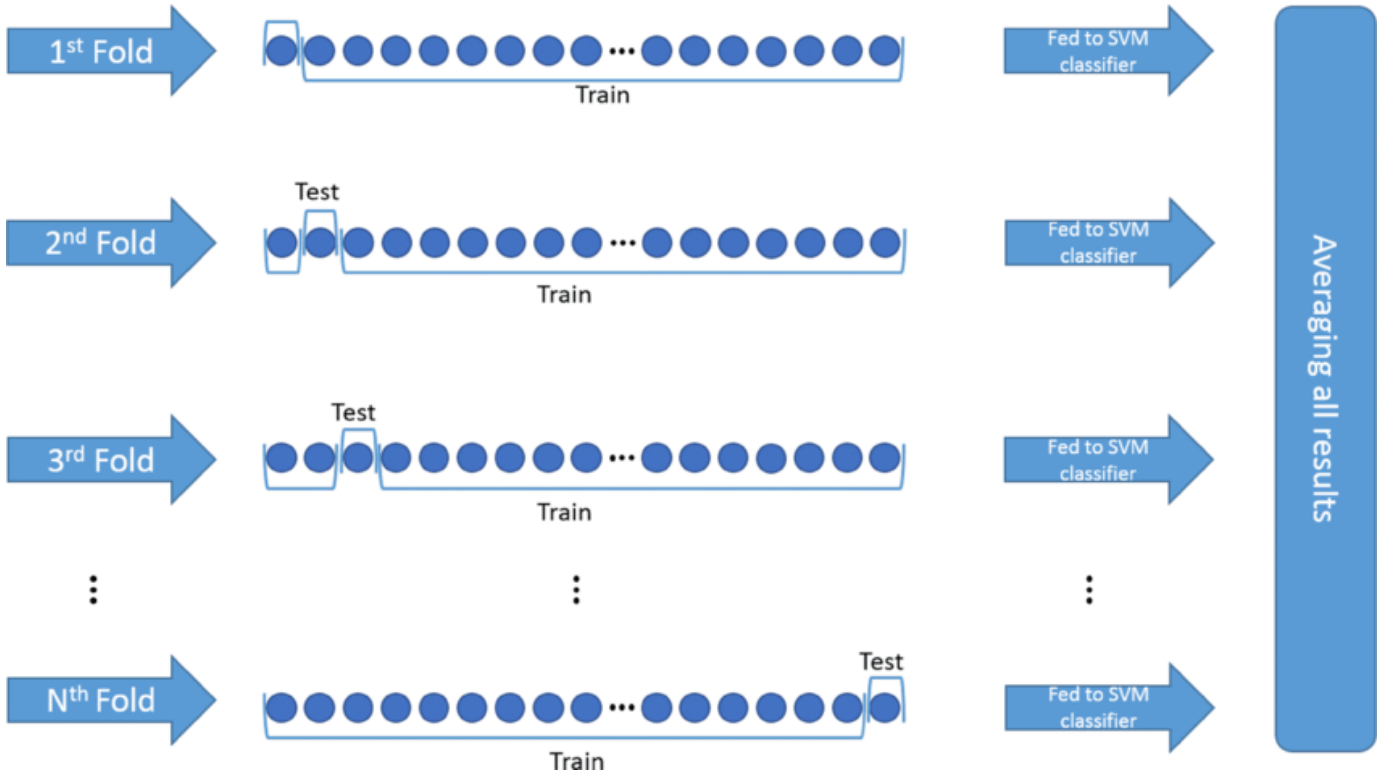
- Veri seti k eşit parçaya bölünür (katman).
- Ardından, her bir katman sırayla test seti olarak kullanılırken, diğer katmanlar birleştirilerek eğitim seti oluşturulur.
- Bu işlem kere tekrarlanır ve her bir katmanın test seti üzerindeki performansı değerlendirilir.
- Sonuçlar genellikle ortalamalar alınarak rapor edilir.



https://scikit-learn.org/stable/modules/cross_validation.html

Leave-One-Out (LOO) Çapraz Doğrulama

- Her bir veri noktası sırayla test seti olarak ayrılırken, geri kalan veri noktaları eğitim seti olarak kullanılır.
- Bu işlem veri noktalarının tamamı için tekrarlanır ve her bir veri noktasının test seti üzerindeki performansı değerlendirilir.



https://www.researchgate.net/figure/The-schematic-of-data-division-in-Leave-one-out-cross-validation-methodology-in-SVM_fig1_332425167