# Metin Madenciliğinin Sınıflandırma ve Kümeleme Uygulamaları

Burak Özdemir Gebze Teknik Üniversitesi

[zdmr.brkk@gmail.com](mailto:zdmr.brkk@gmail.com)

# ÖZET

Veri madenciliğinin önemli bir konusu olan “Metin Madenciliği” bu projede hem “Sınıflandırma” hem de “Kümeleme” metodları açısından değerlendirilmiştir. Sınıflandırma algoritması olarak , istatistiksel olasılık hesabı yapan efektiv ve hızlı bir algoritma olan Naive Bayes algoritmasının 3 türü ele alınmıştır. Kümeleme algoritması olarak ise PAM(Medoidler Etrafında Bölümleme) ı temel alan ve iyileştirilmiş CLARANS(Rastgele Aramaya Dayalı Büyük Uygulamaları Kümeleme) algoritması kullanılmıştır.Proje için çeşitli preprocess ve postprocess işlemleri uygulanıp algoritmalar test edilmiştir.

**Anahtar Kelimeler:** Naive Bayes , Bernoulli , Gaussian , Multinomial , Clarans , Pearson , Kendall Tau , PCA , LDA , Bag of Words , TFIDF

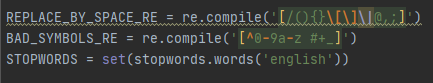
# GİRİŞ

Bu projede veri kümesi olarak “Stackoverflow” platformu üzerinde atılmış commentler ve bu commentlerin hangi konu ile ilişkili olduğunu gösteren tag yani label bilgisi içeren bir dosya kullanılmıştır. Veri algoritmalara ulaşmadan önce metin şeklinde olduğu için bu metinleri sayısal ifadelere dönüştürme gerekliliği vardır. Bag of Words ve TFIDF modelleri metinleri sayısallaştırmak için kullanışmıştır. Tabiki algoritmaların daha iyi sonuç verebilmesi için , en önce metinleri gereksiz karakterlerden , sayılardan temizleme işlemi yapılmıştır. Bu şekilde metinlerin tahminlendirilme yüzdesi artacaktır. Sınıflandırma için kullandığımız Naive Bayes algoritmasının 3 çeşidi kullanılmıştır . Bernouilli algoritması input olarak binary vectorler kullanıyor yani vector içerisinde 1 ve 0 olarak kelimelerin varlğı ve yokluğu ile ilgileniyor. Gaussian ise sınıfın verilerindeki normal dağılımını hesaplayarak snıflandırıyor. Son olarak Multinomial ise kelimelerin dökümanlardaki frekansını kullanarak sınıflandırma yapıyor. Kümeleme için ise CLARANS algoritması veri kümesini medoidler üzerinden bölümleyerek ilerliyor . Bu algoritma Kmedoid ve Clara algoritmasına alternatif ve iyileştirilmiş olarak literatüre katılmıştır.

# İKİNCİ BÖLÜM

Bu bölümde projede yapılanlar , karşılaşılan sorunlar ve sonuçlar detaylı anlatılacaktır. Preprocess , Feature Selection and Extraction , Modeller ve Metrikler bölümlerinden oluşuyor.

**Preprocess**

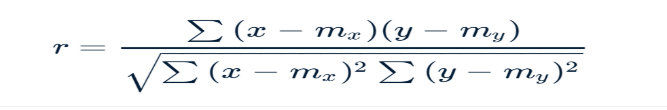
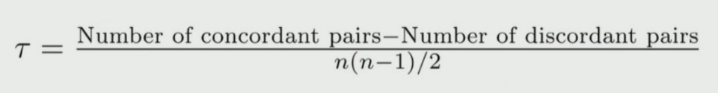
Bu aşamada gelen tüm sampllerı özel

karakterlerden , sayılardan ve

“stopwords” model için bi anlam ifade

etmeyecek ifadelerden temizliyoruz

**Feature Selection and Extraction**

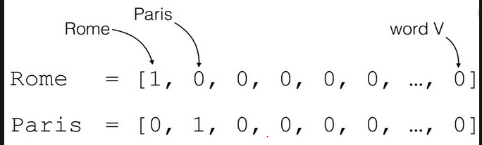
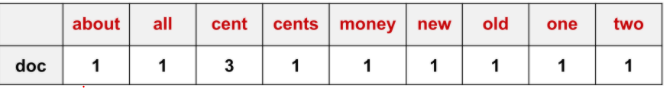
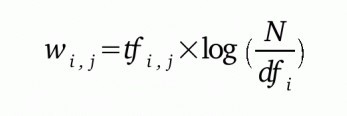


Kendall Tau Formula Pearson Formula

Bu bölümde parametrik korelasyon yöntemi olan Pearson ve Non-Parametric korelasyon olan Kendall Tau yöntemi implement edilmiştir. Label bazında kolonların arasındaki korelasyonlara bakılıp maximum pozitif korelasyonlar seçilmiştir . Projede Gaussian Naive Bayesin performansı düşük olduğu için ona uygulanmıştır

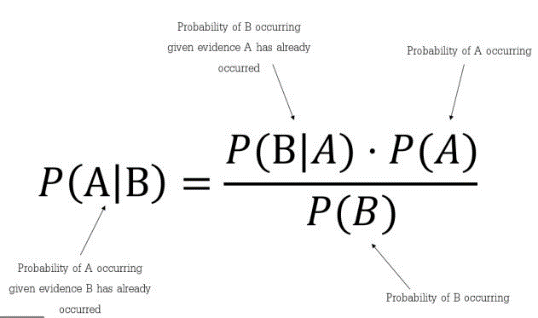
Ayrıca Clarans kümeleme algoritması için “Feature Extraction” metodlarından PCA(Temel Bileşenler Analizi) ve LDA (Doğrusal Diskriminant Analizi) kullanılarak verikümesi algoritma için uygun hale getirilmiştir.

**Modeller**



Bernoulli Vectorization Gaussian Vectorization Multinomial Vectori.

Yukarıdaki şekillerde 3 tip Naive Bayes algoritması için metin vectorization dönüşümlerinde nasıl bir dönüşüm kullanıldığı gösterilmiştir . Bernoulli yaklaşımı temel Naive Bayes teoriminde text verisinin binary representasyonu kullanılmıştır. Gaussian yaklaşımı için ise TfIdf represantasyonu kullanılmıştır. TF(kelimenın dokumanda geçme sayısı) \* IDF( log(Kelimenin bulunduğu döküman sayısı / döküman sayısı)) hesabı ile metin verisi daha anlamlı hale gelmektedir. Multinomial yaklaşımı için ise Bag Of Word yani kelimelerin frekansları ile dönüşüm kullanılmıştır.

Algoritmanın çalışma şekli bir eleman için her durumun olasılığını hesaplar ve olasılık değeri en yüksek olan o verinin tahminlendirilmiş sınıfı olur .

# C:\Users\IS97640\Desktop\datamining\termProject\report\clarans.png

# Bu algoritma veritabanının tümünü almak yerine küçük bir örneklem (sample) kümesini temsilcisi olarak alıp örneklem üzerinde PAM algoritmasını uygular. Veritabanında birden çok örneklem seçerek en iyi sonuç veren örneklemden elde ettiği PAM sonucunu çıktı olarak verir. Örneklem seçimindeki önyargıyı gidermek için sabit bir örneklem yerine her aşamada değişen örneklem kavramını ortaya atmıştır. Rastgele seçilen noktalar çevresi dikkate alınarak bir örneklem oluşturulur. Bu örneklemde k-medoids algoritması ile bulunan merkez noktalar temsil noktası olarak alınır. Bu işlemden sonra, başka rastgele noktalar bulunur ve bu noktalar çevresindeki noktalara yine kmedoids uygulanır ve yeni merkez noktalar bulunarak bu şekilde tüm veritabanının örneklemi tarafsızca oluşturulur. Böylece örneklem seçimi önyargıdan bağımsız hale gelir

# Metrikler

# 

# Sınıflandırma kalitesini ölçmek için F1Score değerleri kullanılmıştır . Bu Score değerlerini çıkartmak için önce Confusion Matrisi oluşturulup ordan çıkarım yapılmıştır . Kesinlik ve Duyarlılık kavramlarının harmonik ortalaması F1Score değerini verir.

# Kümeleme metriği olarak ise Siluet Yöntemi

# kullanılmıştır.Kümeler ve veri kümeleri içinde

# yorumlanması ve tutarlılık doğrulama optimal sayısını

# bulmak için bir yöntemdir.Bir noktanın diğer

# kümelere kıyasla kendi kümesine ne kadar benzediğini

# ölçen bir yaklaşımdır.

# TEST

**Sınıflandırma**

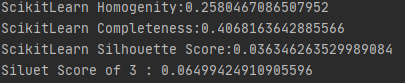
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset\Kriter | Yapı | Bernoulli | Gaussian | Multinomial | LogisticReg |
| postAndTags | 20L,40kS | **%70** | **%0.5** | **%72** | **%79** |
| postAndTags2 | 20L,10kS | **%68** | **%0.5** | **%69** | **%78** |
| postAndTags3 | 20L,1kS | **%53** | **%0.5** | **%55** | **%67** |
| postAndTags5 | 5L,2.6kS | **%85** | **%20** | **%88** | **%91** |
| postAndTags6 | 7L,14kS | **%53** | **%14** | **%85** | **%92** |
| postAndTags7 | 20L,15kS | **%69** | **%0.5** | **%71** | **%85** |

Yukarıda değişik özellikte verikümeleri ile algoritmalar test edilmiştir. Veri kümesindeki sample sayısının fazlalığı algoritmalara pozitif bir yarar sağlamıştır. Label sayısının az olmasıda pozitif katkı sağladığı tespit edilmiştir. Bernoulli ve Multinomialin aksine Gaussian çok kötü performans sergilemiştir çünkü metinlerden oluşturulan veri kümesinde “Normal Dağılım” gözlenmemektedir. Metin sınıflandırmada kütüphanede hazır algoritma olarak Logistic Regresyo kullanılmıştır.

**Sınıflandırma – Feature Selection (Gaussian)**

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset\Method | Raw | KendallTau | Pearson |
| postAndTags3 (20L,1kS) | **%0.5** | **%32** | **%28** |

Özellik seçimi olarak 2 metod uygulanmıştır. Bunlar KendallTau (Non-Parametric) ve Pearson (Parametric) metodlarıdır. KendalTau hem dikey hem yanal korelasyon taraması yaptığı yaptığı için süre bakımından daha yavaş çalışıyor ama score bakımından biraz daha Pearson a göre etkili bir yöntem olduğu gözlenmiştir.

**Kümeleme**

Kümeleme işleminde homogenity , completeness

ve silhoutte scoreları değerlendirilmiştir. Yandaki

çıktılarda gözüktüğü gibi orjinal K değerine

yaklaşıldığı zamanlarda scorelar artıyor ve orjinal

K değerinden uzaklaşıldığı zaman değerler

azalıyor. Kendi implementasyonunu yaptığım

Silhouette skor metodumdan güzel sonuçlar

alamadım.Ekstra olarak litareturde kullanılan

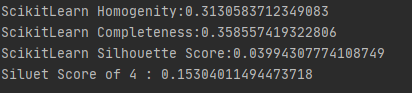
Kmeans algoritması ile karşılaştırma işlemleri

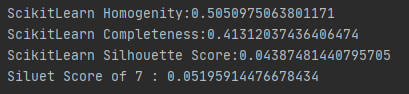
Aşağıda mevcuttur. Ek olarak bu

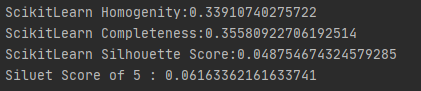
“postAndTags9.csv” verikümesinde optimum

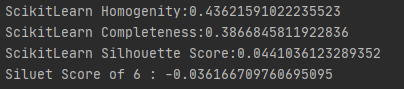
parametreler numLocal:10 , maxNeigh:90

değerleridir.



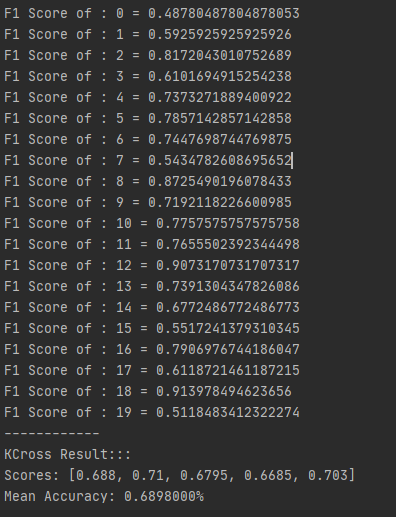


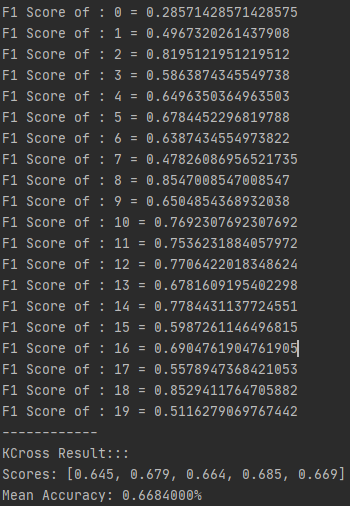
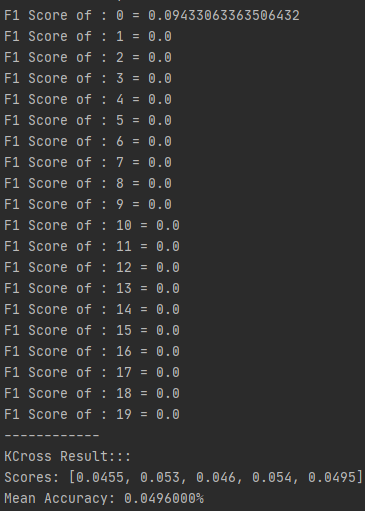




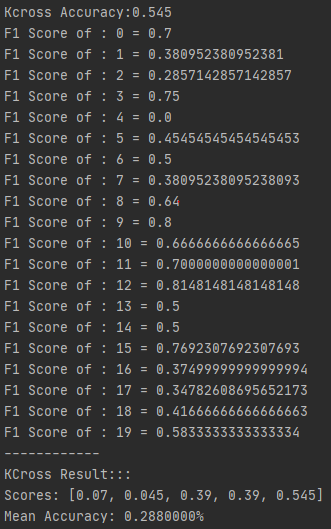
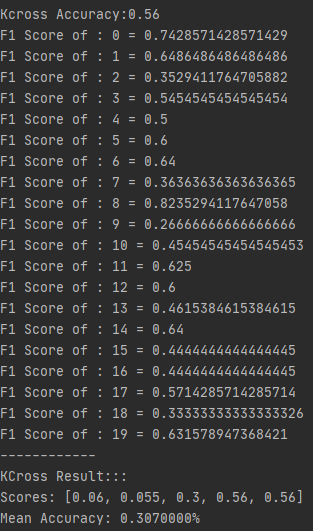
**Bazı Ekran Çıktıları**

**Sınıflandırma**



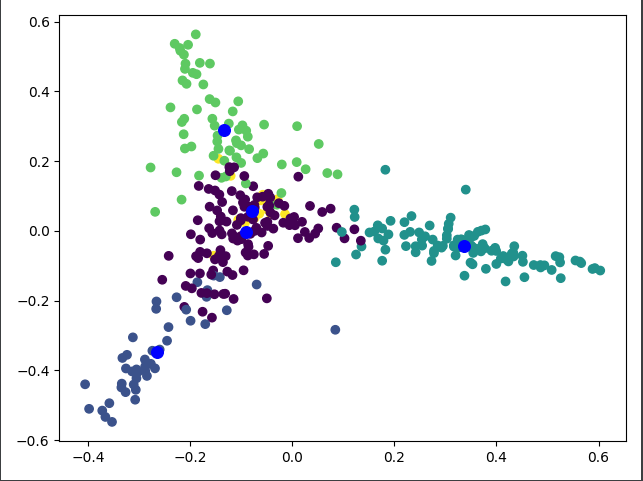


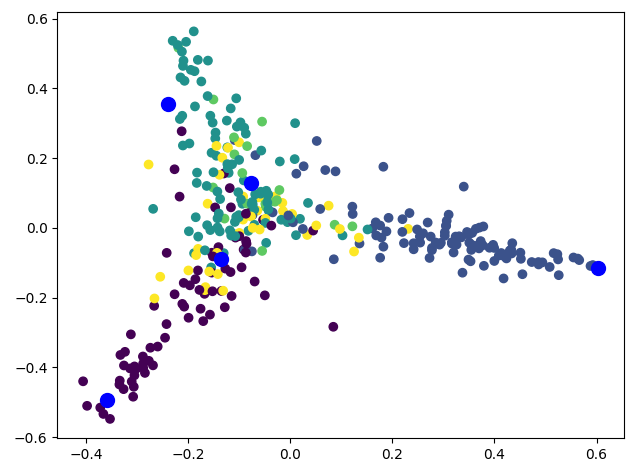
File2 – Bernoulli File2 – Gaussian File2 - Multi

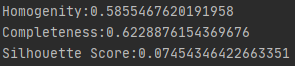


File3 – Gaussian – Kendall File3 – Gaussian – Person

**Kümeleme**

****

****

**C:\Users\IS97640\Desktop\datamining\termProject\report\claransFile9Scores.PNG**

**Kmeans – K:5 Clarans – K:5**

**Kaynaklar**

1. <https://en.wikipedia.org/wiki/Principal_component_analysis#:~:text=Principal%20component%20analysis%20(PCA)%20is,components%20and%20ignoring%20the%20rest>.
2. <http://www.cs.ecu.edu/dingq/CSCI6905/readings/CLARANS.pdf>
3. <https://web.stanford.edu/class/cs124/lec/naivebayes.pdf>
4. <https://hanj.cs.illinois.edu/bk3/bk3_slidesindex.htm>
5. <http://fizyka.umk.pl/publications/kmk/07-Pearson.pdf>