HACETTEPE
ÜNİVERSİTESİ

Mayıs 2024

BÜYÜK VERİ ANALİTİĞİ FİNAL ÖDEVİ:

DUYGU ANALİZİ

Doç. Dr. Duygu İÇEN

Eda Yaren ÖZEL
2200329007

1 BÜYÜK VERİ VE İSTATİSTİK

İstatistik, veri toplama, analiz etme, veriyi anlama ve ilgili belirsizlikleri hesaba katma bilimidir. Bu nedenle, fiziksel, doğal ve sosyal bilimlere; halk sağlığına; tıbba; iş dünyasına ve politikaya nüfuz eder. Büyük Veri, hacim, çeşitlilik ve bazı durumlarda toplanma hızı açısından karmaşık olan veri setlerinin toplanması ve analiz edilmesidir. Büyük Veriler özellikle zorludur çünkü bazıları belirli bir bilimsel soruyu ele almak için toplanmamıştır[1]. Modern teknoloji, sağlık verilerinden sosyal medya ölçümlerine kadar büyük, karmaşık veri setlerinin neredeyse gerçek zamanlı olarak sunulmasına olanak tanıyor. 'Büyük veri' terimi, yapılandırılmış veya yapılandırılmamış bu veri kümelerini tanımlamak için kullanılır ve genellikle geleneksel yöntemler kullanılarak analiz edilemeyen verileri ifade eder. Büyük verinin boyutu ve karmaşıklığı, uzman yazılımların kullanılmasını gerektirir; bu da önemli miktarda işlem gücü ve depolama kapasitesi gerektirebilir . Büyük veriyi benimsemek maliyetli olsa da kuruluşların güçlü içgörüler elde etmesine ve rekabet avantajı kazanmasına olanak tanır[2].

Büyük veri sorunları, doğaları gereği genellikle çok disiplinli ekipler gerektirir. En azından, tipik olarak konu alanı (domain) uzmanları, hesaplama uzmanları, makine öğrenimi uzmanları ve istatistikçiler gerektirir. İstatistik, Büyük Veri'den anlamlı ve doğru bilgilerin çıkarılmasını sağlamak için temel öneme sahiptir. Aşağıdaki konular çok önemlidir ve Büyük Veri ile daha da kötüleşmektedir:

- Veri kalitesi ve eksik veriler,
- Verilerin gözleme dayalı olması, dolayısıyla müdahalelerin karşılaştırılması gibi nedensel soruların karışıklığa maruz kalabilmesi,
- Tahminlerin, öngörülerin ve modellerin belirsizliğinin ölçülmesi,

Bilimsel istatistik disiplini, bu konulara gelişmiş teknikler ve modeller getirmektedir. İstatistikçiler bilimsel sorunun istatistiksel bir soruya dönüştürülmesine yardımcı olurlar; bu da veri yapısının, veriyi oluşturan temel sistemin (model) ve değerlendirmeye çalıştığımız şeyin (tahmin etmek istediğimiz parametre veya parametreler) veya tahminin dikkatlice tanımlanmasını içerir[1].

Büyük Veri'de istatistik bilimleri ve alan bilimleri her zamankinden daha fazla iç içe geçmiştir ve istatistiksel metodoloji çıkarım yapmak için kesinlikle kritik öneme sahiptir. Büyük Veri, genellikle düşük boyutlu ve daha az karmaşık ortamlarda çalışan "hazır" yöntemler veya kara kutu hesaplama araçları tarafından iyi bir şekilde sunulmayacaktır ve bu nedenle özel istatistiksel yöntemler gerektirir. İstatistikçiler, önyargıyı değerlendirme ve düzeltme; belirsizliği ölçme; çalışmaları ve örnekleme stratejilerini tasarlama; verilerin kalitesini değerlendirme; çalışmaların sınırlamalarını sayma; eksik veriler ve diğer örnekleme dışı hata kaynakları gibi sorunlarla başa çıkma; karmaşık veri yapılarının analizi için modeller geliştirme; nedensel çıkarım ve karşılaştırmalı etkinlik için yöntemler oluşturma; gereksiz ve bilgilendirici olmayan değişkenleri ortadan kaldırma; birden fazla kaynaktan gelen bilgileri birleştirme ve etkili veri görselleştirme tekniklerini belirleme konusunda beceriklidir[1].

1.1 Veri Bilimi ve İstatistik Alanlarının Karşılaştırılması

İstatistik, nicel verileri toplamayı ve yorumlamayı amaçlayan, matematiksel temelli bir alandır. Buna karşılık veri bilimi, çeşitli biçimlerdeki verilerden bilgi çıkarmak için bilimsel yöntemleri, süreçleri ve sistemleri kullanan çok disiplinli bir alandır. Veri bilimcileri istatistik dahil birçok disiplinden yöntemler kullanır. Ancak alanlar, süreçleri, incelenen problem türleri ve diğer bazı faktörler bakımından farklılık gösterir.

Model Oluşturma ve Karşılaştırma Süreci: Birçok veri bilimi sorunu, modelin tahmin doğruluğuna odaklanan bir modelleme süreciyle ele alınır. Veri bilimcileri bunu, farklı makine öğrenimi yöntemlerinin tahmin doğruluğunu karşılaştırıp en doğru modeli seçerek yaparlar. İstatistikçiler modellerini oluşturma ve test etme konusunda farklı bir yaklaşım benimserler. İstatistikte başlangıç noktası genellikle basit bir modeldir (örneğin doğrusal regresyon) ve verilerin bu modelin varsayımlarıyla tutarlı olup olmadığı kontrol edilir. Modelde ihlal edilen varsayımlar ele alınarak model iyileştirilir. Tüm varsayımlar kontrol edildiğinde ve hiçbir varsayım ihlal edilmediğinde modelleme süreci tamamlanır. Veri bilimi, en iyi makine öğrenimi modelini oluşturmak için birçok yöntemi karşılaştırmaya odaklanırken, istatistik bunun yerine verilere en iyi şekilde uyacak tek ve basit bir modeli geliştirir.

Belirsizliğin Ölçülmesi: İstatistikçiler belirsizliği ölçmeye veri bilimcilerden çok daha fazla odaklanırlar. İstatistiksel model oluşturma sürecinin bir parçası, her bir tahminci ile tahmin edilen sonuç arasındaki kesin ilişkiyi ölçmektir. Bu ilişkiyle ilgili herhangi bir belirsizlik de ölçülür. Bu süreç makine öğreniminde nadiren gerçekleşir.

Büyük Veri: Veri bilimcileri genellikle tek bir bilgisayarda depolanamayacak kadar büyük veritabanlarıyla uğraşırlar. Bu tür veriler bazen istatistiklerde yer alsa da, bu bir normdan çok istisnadır. Tarihsel olarak istatistiğin odak noktası daha çok çok küçük miktarlardaki verilerden ne öğrenilebileceği üzerine olmuştur. Küçük verilere bu şekilde odaklanılması, istatistiklerdeki belirsizliğin ölçülmesinin neden önemli olduğunu açıklamaktadır. Yalnızca az miktarda veriye sahip olduğunuzda, sinyali gürültüyle karıştırmak kolaydır. Veri bilimi tarafından sıklıkla incelenen verilerin büyüklüğü, aynı zamanda veri bilimcilerin varsayımları kontrol etmesinin pratik olmamasının da nedenidir.

İncelenen Problem Türleri: Veri bilimi sorunları genellikle tahminlerde bulunmak ve büyük veritabanlarında aramayı optimize etmekle ilgilidir. Bunun tersine, istatistiklerin incelediği problemler daha çok dünya geneli hakkında sonuçlar çıkarmaya odaklanıyor. Bu, verilerin en iyi şekilde nasıl toplanacağı ve ölçümlerin yapılacağı ve bu ölçümlerle ilgili belirsizliğin nasıl ölçüleceği üzerinde çalışmayı içerir. İstatistiksel analizin nihai amacı genellikle belirsizliğin niceliğine dayalı olarak neyin neye sebep olduğu hakkında bir sonuca varmaktır. Buna karşılık, veri bilimi analizinin nihai hedefi daha çok belirli bir veritabanı veya tahmine dayalı modelle ilgilidir.

Alanlar, modelleme süreçlerine, verilerinin boyutuna, incelenen problem türlerine, alandaki kişilerin geçmişine ve kullanılan dile göre farklılık gösterir. Ancak alanlar birbiriyle yakından ilişkilidir. Sonuçta hem istatistik hem de veri bilimi verilerden bilgi çıkarmayı amaçlamaktadır[3].

2 VERİ KÜMESİ

Table "default" - Rows: 199898 | Spec - Columns: 2 | Properties | Flow Variables

Row ID	S text	S humor
Row0	Joe biden rules out 2020 bid: 'guys, i'm not running'	False
Row1	Watch: darvish gave hitter whiplash with slow pitch	False
Row2	What do you call a turtle without its shell? dead.	True
Row3	5 reasons the 2016 election feels so personal	False
Row4	Pasco police shot mexican migrant from behind, new autopsy shows	False
Row5	Martha stewart tweets hideous food photo, twitter responds accordingly	False
Row6	What is a pokemon master's favorite kind of pasta? wartortellini!	True

Şekil 1: Mizah analizi için kullanılacak veri seti

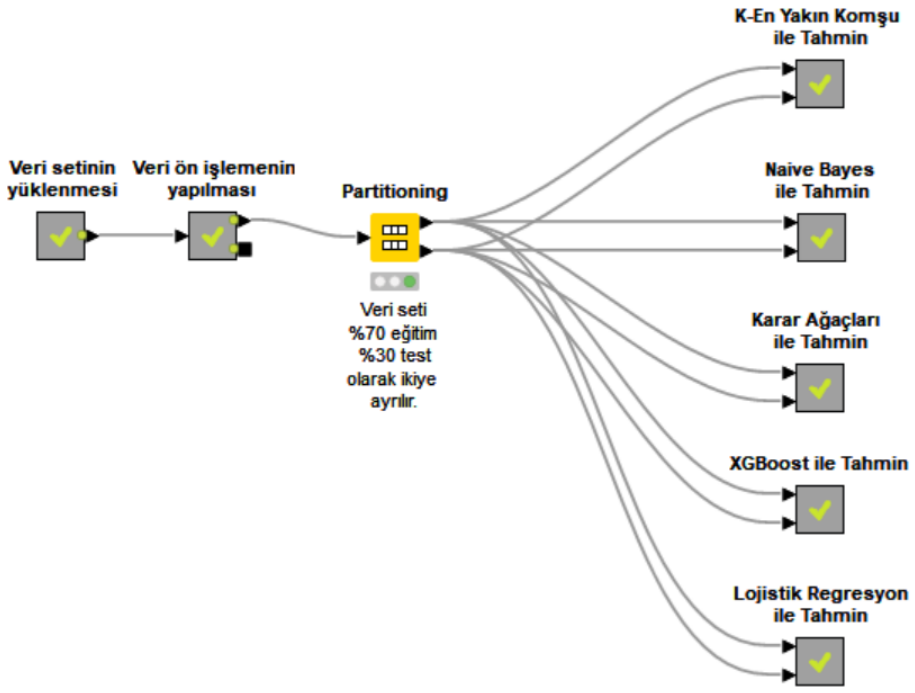
Ödev kapsamında kullanılan veri seti [4] Tablo 1'deki, 200.000 tane benzersiz metinin bulunduğu "text" ve bu metinlerin mizah olup olmadığını içeren "humor" sütunlarından oluşmaktadır.

Tablo 1: Verisetine İlişkin Değişkenler

Değişkenler
Metin
Mizah (True/False)

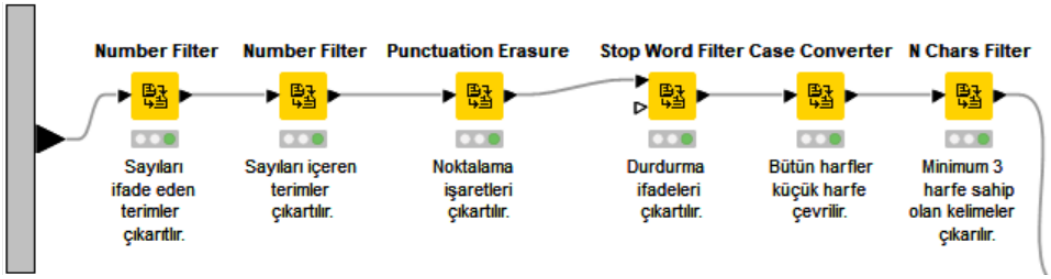
NOT: Ödev kapsamında kullanılan veri kümesi üzerinden **5000** tane rasgele veri seçilerek analizler seçilen veriler üzerinden gerçekleştirilmiştir.

3 KNİME UYGULAMASI



Şekil 2: Mizah analizi için oluşturulan knime akışı

Kaggle üzerinden bulunan veri seti üzerinde mizah analizi gerçekleştirmek için ilk olarak veri setine çeşitli ön işleme adımları uygulanmıştır. Bu adımlar, veriyi temizleyip analiz için uygun hale getirmeyi içermektedir. Daha sonra, veri seti eğitim ve test olmak üzere ikiye ayrılmıştır. Eğitim ve test kümeleri oluşturulduktan sonra, veri madenciliği algoritmalarından K-NN, Naive Bayes, Karar Ağaçları, XGBoost ve Lojistik Regresyon modelleri kullanılarak metinlerin mizah tahminleri yapılmıştır.



Şekil 3: Veri yükleme adımları

- **CSV Reader:** CSV dosyasından veri okumak için kullanılan bir düğümdür. Bu düğümde, dosyanın konumu ve hangi sütunların okunacağı seçilebilir. Ayrıca, sütun türlerini, ayırıcı karakterleri ve veri türü tanımlamalarını yapılandırmak için seçenekler sunar.
- **Row Sampling:** Veri setinden rastgele veya belirli bir oranda satır örneklemeleri yapmak için kullanılır. Örnekleme yöntemleri arasında sabit bir sayıda satır seçme veya veri setinin belirli bir yüzdesini seçme gibi seçenekler bulunur. Bu düğüm, büyük veri setlerini daha küçük ve yönetilebilir alt kümelere bölmek için kullanışlıdır.
- **Row Filter:** Veri setindeki satırları belirli koşullara göre filtrelemek için kullanılır. Bu düğüm, belirli bir sütun değeri veya değeri aralığına dayalı olarak satırları dahil etmek veya hariç tutmak için yapılandırılabilir.

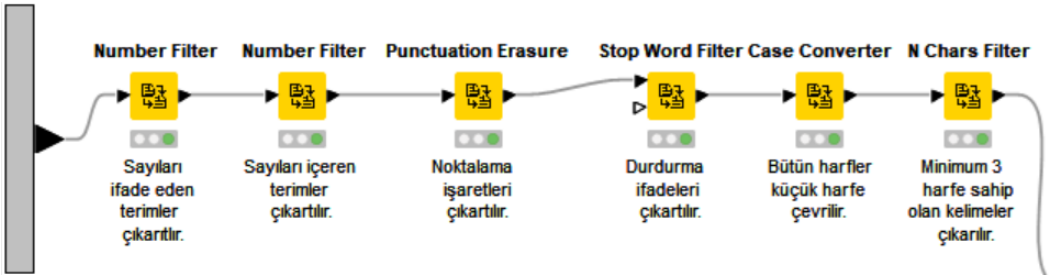
- **String to Document:** Metin verilerini belge formatına dönüştürmek için kullanılan bir düğümdür. Metin verilerini daha sonra analiz etmek üzere uygun belge formatına dönüştürerek, kelime frekansı, duygu analizi ve diğer metin tabanlı işlemler için hazır hale getirir.
- **Document Viewer:** Belgeleri görselleştirmek ve içeriğini incelemek için kullanılan bir düğümdür. Bu düğüm, metin verilerini belge formatında görüntüler ve kullanıcıların metin içeriğini detaylı bir şekilde incelemesine olanak tanır.
- **Column Filter:** Veri setindeki sütunları filtrelemek veya seçmek için kullanılır. Bu düğüm, belirli sütunları dahil etmek veya hariç tutmak için yapılandırılabilir. Analiz için sadece gerekli sütunları seçerek veri setini sadeleştirir ve performansı artırır.

Table "default" - Rows: 5000 | Spec - Column: 1 | Properties | Flow Variables

Row ID	Document
Row5	"Martha stewart tweets hideous food photo, twitter responds accordingly"
Row114	"How to build muscle: proven strength lessons from milo of croton"
Row174	"'gotham' actor donal logue's missing child has been found and is safely home"
Row246	"Why did the chicken cross the mobius strip. to get to the same side."
Row278	"What do tampons and white women have in common? they are both stuck up cunts."
Row286	"Civilian death toll mounts as iraqi forces push on in mosul"

Şekil 4: Veri yükleme işlemlerinin sonucu

Düğümler sonucunda, veri seti okunduktan sonra 5000 veriden oluşan bir örneklem çekilmiştir. Bu örneklem üzerinde öncelikle kayıp satırlar çıkarılmış, ardından sütunlar belge formatına dönüştürülmüş ve analizde kullanılmak üzere metin sütunu diğer sütunlardan filtrelenmiştir.



Şekil 5: Veri ön işleme adımları

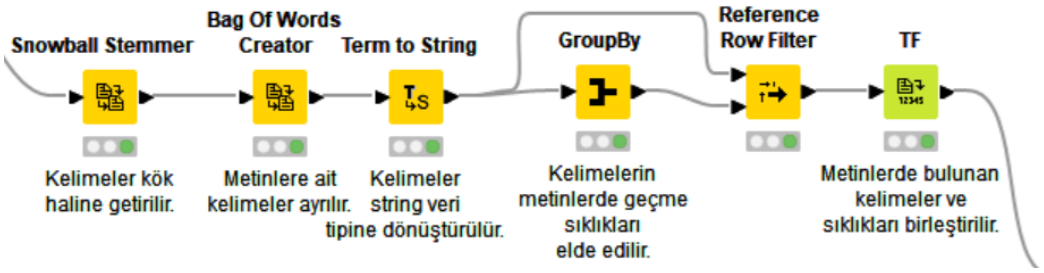
- **Number Filter:** Metin verilerinden sayıları filtrelemek için kullanılır. Metin verileri içindeki sayıları, metinden çıkarmak veya belirli bir eşik değerinden büyük veya küçük olanları filtrelemek için kullanılabilir.
- **Punctuation Erasure:** Metin verilerinden noktalama işaretlerini (örneğin, virgül, nokta, ünlem işareti vb.) kaldırmak için kullanılır. Metin verileri içindeki noktalama işaretlerini temizlemek veya değiştirmek için kullanılabilir.
- **Stop Word Filter:** Metin verilerinde bulunan sık kullanılan kelimeleri (durdurma kelimeleri) filtrelemek için kullanılır. Durak kelimeleri genellikle anlamı olmayan veya analiz için önemsiz olan kelimelerdir.
- **Case Converter:** Metin verilerindeki harf büyüklüğünü (büyük/küçük) değiştirmek için kullanılır.
- **N Chars Filter:** Metin verilerinde belirli bir karakter uzunluğundan daha uzun veya daha kısa olan kelimeleri filtrelemek için kullanılır.

Table "default" - Rows: 5000 | Spec - Columns: 2 | Properties | Flow Variables

Row ID	Document	Preprocessed Document
Row5	"martha stewart tweets hideous food photo twitter responds accordingly"	"Martha stewart tweets hideous food photo, twitter responds accordingly"
Row114	"build muscle proven strength lessons milo croton"	"How to build muscle: proven strength lessons from milo of croton"
Row174	"gotham actor donal loguemissing child found safely home"	"'gotham' actor donal logue's missing child has been found and is safely home"
Row246	"chicken cross mobius strip"	"Why did the chicken cross the mobius strip. to get to the same side."
Row278	"tampons white women common stuck cunts"	"What do tampons and white women have in common? they are both stuck up cunts."

Şekil 6: Veri ön işleme sonucu

Düğümler sonucunda, analizde kullanılacak metinlerden sayılar, sayıları ifade eden terimler, noktalama işaretleri, durdurma ifadeleri ve üç harften daha kısa kelimeler çıkartılmıştır. Ayrıca, metinlerde yer alan tüm kelimeler küçük harfle başlatılmıştır.



Şekil 7: Veri ön işleme adımları

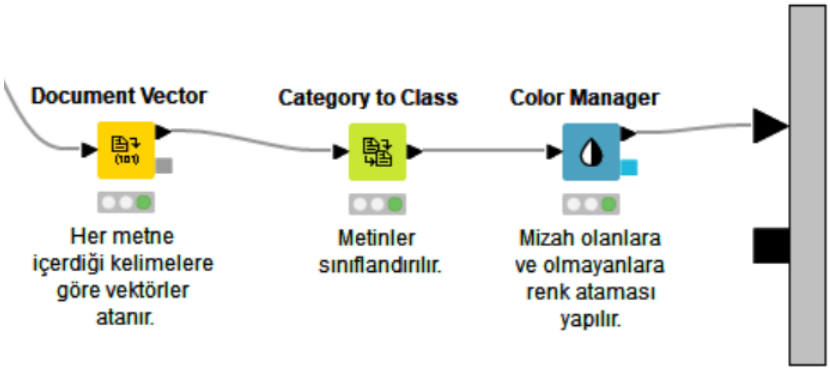
- **Snowball Stemmer:** Metin verilerindeki kelimeleri köklerine (stem) indirgemek için kullanılır. Snowball stemmer, dilbilgisi kurallarını kullanarak kelimeleri kök formlarına dönüştürür.
- **Bag Of Words Creator:** Metin verilerinden bir kelime torbası (bag of words) oluşturmak için kullanılır. Kelime torbası, belirli bir metin koleksiyonunda geçen tüm kelimelerin bir listesini ve her kelimenin bu metinlerdeki sıklığını içerir.
- **Term to String:** Sayısal terimleri metin tabanlı terimlere dönüştürmek için kullanılır.
- **GroupBy:** Veri setindeki satırları belirli bir sütuna göre gruplamak için kullanılır. Gruplama işleminden sonra, her bir grup üzerinde toplu işlemler gerçekleştirilebilir.
- **References Row Filter:** Belirli bir referans sütunundaki değerlere göre satırları filtrelemek için kullanılır.
- **TF:** Belirli bir metin veri kümesindeki terim frekansını (TF) hesaplamak için kullanılır. TF, bir terimin bir belgedeki görünme sıklığını ifade eder.

Table "default" - Rows: 28479 | Spec - Columns: 4 | Properties | Flow Variables

Row ID	Document	Term	Term as String	TF abs
Row0	"martha stewart tweet hideou food photo twitter respond accordingli"	martha[]	martha	2
Row1	"martha stewart tweet hideou food photo twitter respond accordingli"	stewart[]	stewart	2
Row2	"martha stewart tweet hideou food photo twitter respond accordingli"	tweet[]	tweet	2
Row3	"martha stewart tweet hideou food photo twitter respond accordingli"	hideou[]	hideou	2
Row4	"martha stewart tweet hideou food photo twitter respond accordingli"	food[]	food	2
Row5	"martha stewart tweet hideou food photo twitter respond accordingli"	photo[]	photo	2
Row6	"martha stewart tweet hideou food photo twitter respond accordingli"	twitter[]	twitter	2
Row7	"martha stewart tweet hideou food photo twitter respond accordingli"	respond[]	respond	2
Row8	"martha stewart tweet hideou food photo twitter respond accordingli"	accordingli[]	accordingli	2
Row9	"build muscl proven strength lesson milo croton"	build[]	build	2

Şekil 8: Veri ön işleme sonucu

Düğümler sonucunda, analizde kullanılacak metinlerdeki kelimeler kök haline getirilmiş ve her metinde yer alan kelimeler ayrıştırılarak kelimelerin metinlerde geçme sıklıkları elde edilmiştir. Daha sonra elde edilen kelimeler metin formatına çevrilerek her kelimenin metinlerde bulunma sıklığı ve kelime bir araya getirilmiştir.



Şekil 9: Veri ön işleme adımları

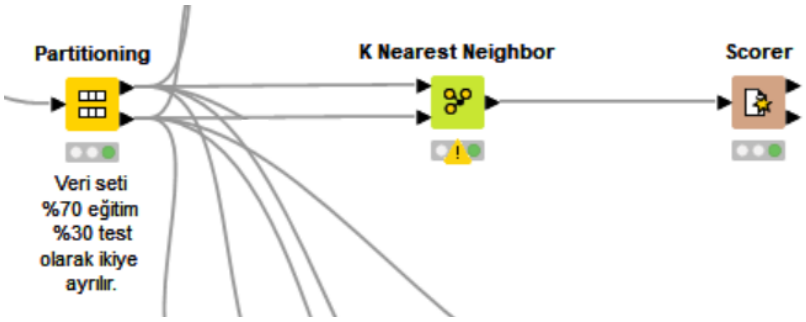
- **Document Vector:** Metin verilerini sayısal vektörler olarak temsil etmek için kullanılır. Metin belgeleri genellikle kelime torbası (bag of words) temsili veya TF-IDF (terim frekansı-ters belge frekansı) temsili gibi tekniklerle vektörlere dönüştürülür.
- **Category to Class:** Kategorik verileri (örneğin, sınıflandırma etiketleri) sayısal olarak temsil etmek için kullanılır.
- **Color Manager:** Renkleri belirlemek ve yönetmek için kullanılır.

Table "default" - Rows: 4998 | Spec - Columns: 7871 | Properties | Flow Variables

Row ID	Document	a-sid	jon	chattman	brian
Row0	"a-sid jon chattman brian wilson summer"	1	1	1	1
Row1	"aboard air forc bush document note"	0	0	0	0
Row2	"abort clinic blue close"	0	0	0	0
Row3	"abort murder condom kidnap"	0	0	0	0
Row4	"abort pro"	0	0	0	0
Row5	"abort train arriv station abort choo choo"	0	0	0	0
Row6	"abram return write direct star war episod"	0	0	0	0

Şekil 10: Veri ön işleme sonucu

Düğümler sonucunda, analizde kullanılacak metinlere içerdikleri kelimelere göre vektörler atanmış ve metinlerin mizah içerip içermediğine bağlı olarak renklendirme yapılmıştır.



Şekil 11: K-en yakın komşu modelinin adımları

- **Partitioning:** Veri setini eğitim ve test kümelerine ayırmak için kullanılır. Veri seti, genellikle belirli bir oranda eğitim ve test verisi olarak bölünür. Bu şekilde, model eğitilirken kullanılan veriler ayrıdır ve modelin performansını değerlendirmek için kullanılan veriler de ayrıdır.
- **K Nearest Neighbor:** K en yakın komşu (KNN) algoritmasını uygulamak için kullanılır. KNN, sınıflandırma veya regresyon problemleri için kullanılan basit bir makine öğrenimi algoritmasıdır. Bir noktanın sınıfını veya değerini belirlemek için, o noktaya en yakın K komşusunun sınıflarının veya değerlerinin ortalaması alınır.

- **Scorer:** Modelin performansını değerlendirmek için kullanılır. Özellikle, sınıflandırma veya regresyon modellerinin doğruluğunu, hassasiyetini, geri çağırma (recall) değerini, F1 skorunu veya diğer metrikleri değerlendirmek için kullanılabilir. Bu düğüm, modelin tahminlerini gerçek etiketlerle karşılaştırarak farklı performans ölçütlerini hesaplar.

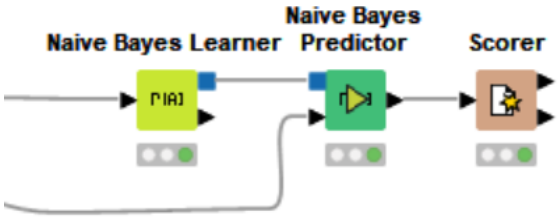
Row ID	False	True
False	51	706
True	31	712

Şekil 12: Knn confusion matrix

Row ID	Precision	Recall	F-measure	Accuracy
False	0.622	0.067	0.122	?
True	0.502	0.958	0.659	?
Overall	?	?	?	0.509

Şekil 13: Knn başarı metrikleri

Düğümler sonucunda, uygulanan knn modelinin başarı oranı %50.9 olarak bulunmuştur.



Şekil 14: Naive bayes modelinin adımları

- **Naive Bayes Learner:** Naive Bayes sınıflandırma modelini eğitmek için kullanılır. Naive Bayes algoritması, sınıflandırma problemleri için yaygın olarak kullanılan bir olasılık temelli bir makine öğrenimi algoritmasıdır. Model, Bayes teoremi ve "naive" varsayımı kullanılarak hesaplanır, yani özellikler arasındaki bağımsızlık varsayılır.
- **Naive Bayes Predictor:** Eğitilen Naive Bayes sınıflandırma modelini kullanarak yeni veri örneklerini sınıflandırmak için kullanılır. Model, önceden hesaplanmış sınıf olasılıklarını kullanarak yeni örneklerin sınıflarını tahmin eder.

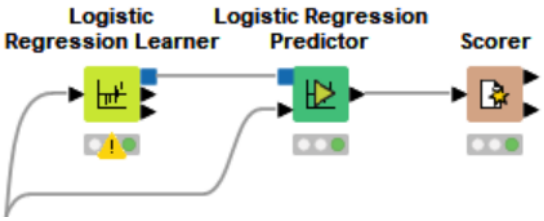
Row ID	True	False
True	743	0
False	757	0

Şekil 15: Naive bayes confusion matrix

Row ID	Precision	Recall	F-measure	Accuracy
True	0.495	1	0.663	?
False	?	0	?	?
Overall	?	?	?	0.495

Şekil 16: Naive bayes başarı metrikleri

Düğümler sonucunda, uygulanan naive bayes modelinin başarı oranı %49.5 olarak bulunmuştur. Modelin mizah olan metinleri tahmin etmeden çok başarılı olduğu fakat mizah olmayan metinleri tahmin edemediği gözlemlenmiştir.



Şekil 17: Lojistik regresyon modelinin adımları

- **Logistic Regression Learner:** Lojistik regresyon modelini eğitmek için kullanılır. Lojistik regresyon, sınıflandırma problemleri için yaygın olarak kullanılan bir makine öğrenimi algoritmasıdır. Model, giriş değişkenlerinin lineer kombinasyonunu alarak ve bu kombinasyonu bir sigmoid fonksiyonuna (logit fonksiyonu) geçirerek hesaplanır. Bu model, sınıflandırma problemlerinde iki veya daha fazla sınıf arasında ayrım yapmak için kullanılabilir.

- **Logistic Regression Predictor:** Eğitilen lojistik regresyon modelini kullanarak yeni veri örneklerini sınıflandırmak için kullanılır. Model, önceden hesaplanmış katsayıları kullanarak yeni örneklerin sınıflarını tahmin eder.

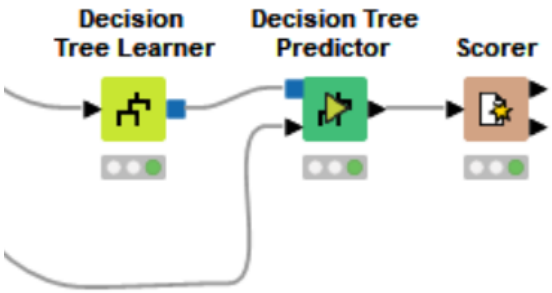
Row ID	False	True
False	603	154
True	136	607

Şekil 18: Lojistik regresyon confusion matrix

Row ID	Precision	Recall	F-measure	Accuracy
False	0.816	0.797	0.806	?
True	0.798	0.817	0.807	?
Overall	?	?	?	0.807

Şekil 19: Lojistik regresyon başarı metrikleri

Düğümler sonucunda, uygulanan lojistik regresyon modelinin başarı oranı %80.7 olarak bulunmuştur.



Şekil 20: Karar ağaçları modelinin adımları

- **Decision Tree Learner:** Karar ağacı sınıflandırma veya regresyon modelini eğitmek için kullanılır. Karar ağacı, sınıflandırma veya regresyon problemlerini çözmek için kullanılan bir makine öğrenimi algoritmasıdır. Model, veri setindeki özelliklerin değerlerine göre bir dizi kararlar alarak ve bu kararlar sonucunda veri örneklerini sınıflara veya değerlere ayıran bir ağaç yapısı oluşturur.
- **Decision Tree Predictor:** Eğitilen karar ağacı modelini kullanarak yeni veri örneklerini sınıflandırmak veya değerlendirmek için kullanılır. Model, önceden belirlenmiş karar kurallarını kullanarak yeni örneklerin sınıflarını veya değerlerini tahmin eder.

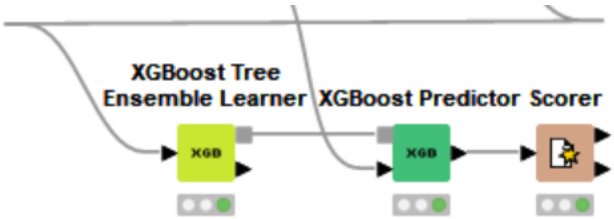
Row ID	False	True
False	490	267
True	128	615

Şekil 21: Karar ağaçları confusion matrix

Row ID	Precision	Recall	F-measure	Accuracy
False	0.793	0.647	0.713	?
True	0.697	0.828	0.757	?
Overall	?	?	?	0.737

Şekil 22: Karar ağaçları başarı metrikleri

Düğümler sonucunda, uygulanan karar ağaçları modelinin başarı oranı %73.7 olarak bulunmuştur.



Şekil 23: XGBoost modelinin adımları

- **XGBoost Tree Ensemble Learner:** XGBoost (eXtreme Gradient Boosting) algoritmasını kullanarak ağaç ansamblı modellerini eğitmek için kullanılır. XGBoost, sınıflandırma ve regresyon problemleri için güçlü bir makine öğrenimi algoritmasıdır. Model, çok sayıda zayıf öğreniciyi (genellikle karar ağaçlarını) bir araya getirerek ve bunları yavaş yavaş geliştirerek oluşturulur.

- **XGBoost Predictor:** Eğitilen XGBoost modelini kullanarak yeni veri örneklerini sınıflandırmak veya değerlendirmek için kullanılır. Model, önceden hesaplanmış karar kurallarını ve ağırlıkları kullanarak yeni örneklerin sınıflarını veya değerlerini tahmin eder.

Row ID	False	True
False	468	289
True	94	649

Şekil 24: XGBoost confusion matrix

Row ID	Precision	Recall	F-measure	Accuracy
False	0.833	0.618	0.71	?
True	0.692	0.873	0.772	?
Overall	?	?	?	0.745

Şekil 25: XGBoost başarı metrikleri

Düğümler sonucunda, uygulanan XGBoost modelinin başarı oranı %74.5 olarak bulunmuştur.

4 SONUÇ VE TARTIŞMA

Ödev kapsamında, Kaggle üzerinden elde edilen bir veri kümesi üzerinde mizah analizi yapmak için KNIME platformu kullanılarak çeşitli makine öğrenimi modelleri uygulanmıştır. 5000 adet rastgele seçilmiş metin ve bu metinlerin mizah içerip içermediğini belirten etiketler içeren veri seti üzerinde K-Nearest Neighbor (KNN), Naive Bayes, Karar Ağaçları, XGBoost ve Lojistik Regresyon modelleri kullanılarak metinlerin mizah tahminleri yapılmıştır.

Tablo 2: Modeller ve Doğruluk oranları

Model	**Doğruluk Oranı**
K Nearest Neighbor	%50.9
Naive Bayes	%49.5
Lojistic Regression	%80.7
Decision Tree	%73.7
XGBoost Tree Ensemble	%74.5

Tablo 2 incelendiğinde, kullanılan modeller arasında Lojistik Regresyon modelinin en yüksek başarı oranına (%80.7) ulaştığı görülmektedir. Bu sonuç, Lojistik Regresyon modelinin mizah içeren ve içermeyen metinleri doğru şekilde sınıflandırmada daha başarılı olduğunu göstermektedir. Diğer yandan, K-Nearest Neighbor (KNN) ve Naive Bayes modellerinin doğruluk oranları sırasıyla %50.9 ve %49.5 ile en düşük seviyede kalmıştır. Bu düşük performans, bu modellerin metin sınıflandırma problemlerinde etkili olamayabileceğini işaret etmektedir. Özellikle KNN modelinin performansının düşük olması, metin verisinin yüksek boyutlu olması ve KNN'in genellikle daha düşük boyutlu veri setlerinde daha etkili olmasıyla açıklanabilir. Karar Ağaçları ve XGBoost modelleri ise sırasıyla %73.7 ve %74.5 doğruluk oranlarıyla Lojistik Regresyon modeline kıyasla daha düşük performans sergilemişlerdir. Karar Ağaçları, aşırı uyum (overfitting) sorununa yatkın olabileceği için, bu durum modelin genel doğruluk oranını olumsuz yönde etkileyebilir. XGBoost modelinin ise daha iyi performans göstermesi, bu modelin ağaç topluluğu (ensemble) yöntemleri kullanarak genellikle daha iyi genelleme kapasitesine sahip olmasından kaynaklanmaktadır.

Sonuç olarak, lojistik regresyon modeli, mizah tahmini için en etkili model olarak öne çıkmıştır. Bu durum, lojistik regresyonun doğrusal olmayan ilişkileri ve sınıflandırma problemlerini iyi bir şekilde modelleyebilme kabiliyetiyle açıklanabilir. Gelecek çalışmalar için, model performanslarının artırılmasına yönelik çeşitli hiperparametre optimizasyonları, daha fazla veri kullanımı ve farklı ön işleme tekniklerinin denenmesi önerilebilir. Ayrıca, derin öğrenme modelleri gibi daha karmaşık algoritmaların da değerlendirilmesi, mizah analizi için potansiyel iyileştirmeler sağlayabilir.

Kaynaklar

[1] "Statistics and big data." https://higherlogicdownload.s3.amazonaws.com/AMSTAT/UploadedImages/49ecf7cf-cb26-4c1b-8380-3dea3b7d8a9d/BigDataOnePager.pdf [Accessed: 11.05.2024].

[2] P. Taylor, "Statista - big data - statistics & facts." https://www.statista.com/topics/1464/big-data/#topicOverview [Accessed: 11.05.2024].

[3] "Statistics vs data science: What's the difference?." https://www.displayr.com/statistics-vs-data-science-whats-the-difference/ [Accessed: 11.05.2024].

[4] https://www.kaggle.com/datasets/deepcontractor/200k-short-texts-for-humor-detection/data. [Accessed: (09.05.2024)].