Twitter verisi üzerinden lojistik regresyon yöntemi kullanılarak cinsiyet tahmini üretilmesi

Özgür Öney[1]
[1]Gazi University, Central Bank of the Republic of Turkey
ozgur.oney@gazi.edu.tr

***Abstract*—** Usage of Online Social Networks (OSN) are become widely popular and that is lead to production of large scale user data that is publicly available. By using these data, predictions can make about the users which can be used for personalization, advertising and law enforcement. This project will classify the Twitter users according to their genders by using the different attributes provided by user profiles on Twitter. For classification weight of each attribute is calculated via logistic regression. The attributes that are believed to be most reliable to make a prediction with high accuracy are selected in the project. The data set will divide into two subsets one for training and one for testing. After learning the weight of each attribute on the training subset, results of the classification will be evaluated on the test subset. The results of our framework can extended furthermore by adding new extracted features and can be used in recommendation systems, marketing.

Index terms - *social network analysis, social media, gender prediction, Twitter, machine learning, estimation, regression.*

Özet - Çevrimiçi Sosyal Ağların (OSN) kullanımı yaygın bir şekilde popüler hale gelmiştir ve bu, halka açık olan büyük ölçekli kullanıcı verilerinin üretilmesine yol açmaktadır. Bu veriler kullanılarak; kişiselleştirme, reklam ve kanun yürütmeleri için kullanılabilecek kullanıcılar hakkında tahminler yapılabilir. Bu projede, Twitter kullanıcıları Twitter'daki farklı özellikler kullanılarak cinsiyetlerine göre sınıflandırılacaktır. Sınıflandırma için her bir özelliğin ağırlığı, lojistik regresyon ile hesaplanır. Projede yüksek doğrulukla bir tahminde bulunmak için en güvenilir olduğuna inanılan özellikler el ile seçilmiştir. Veri seti, biri eğitim için diğeri test için iki alt gruba ayrılacaktır. Eğitim alt kümesindeki her bir özelliğin ağırlığını öğrendikten sonra, sınıflandırma sonuçları test alt kümesinde değerlendirilecektir. Çerçevemizin sonuçları, yeni çıkarılan özellikler ekleyerek daha da genişletilebilir ve tavsiye sistemlerinde, pazarlamada kullanılabilir.

Endeks terimleri - *sosyal ağ analizi, sosyal medya, cinsiyet tahmini, Twitter, makine öğrenmesi, tahmin, regresyon.*

I. Giriş

Günümüzde yaygın olarak kullanılan birçok Çevrimiçi Sosyal Ağ (OSN) bulunmaktadır ve bunun sonucunda çeşitli büyük ölçekli veriler üretilmektedir. Halka açık bu veriler hem araştırma alanı hem de uygulamaların geliştirilmesi için fırsatlar sunar. Ham veriler hiçbir bilgiyi kendiliğinden sağlamamaktadır, ancak verileri analiz etmek ve onlardan özellik çıkarmak, pazarlamada, kişiselleştirmede ve hatta yasal soruşturmada bile kullanılabilecek kullanıcılar hakkındaki bilgiyi verir. Bu nedenle, son yıllarda büyük veri ve büyük verinin işlenmesine dair atılımlar hızla akademi, sanayi ve hatta dünyadaki hükümetlerin dikkatini çeken bir sıcak nokta haline gelmiştir [1].

Cinsiyet bilgisi genel olarak OSN'ler[1] tarafından toplanmamakta veya toplansa bile kamuoyu ile paylaşılmamaktadır, tahminen bu durumun sebebi yukarıda belirtildiği şekilde birçok alanda kullanılabilecek kıymetli bir ticari bilgi içerdiği içindir. Bu durumda cinsiyet hakkında bilgi, kullanıcı profillerinde mevcut olan diğer bilgiler tarafından tahmin yönteminin yardımıyla elde edilebilir. Bu projede, kullanıcının profili tarafından sağlanan cinsiyet harici bilgiler kullanılarak en popüler sosyal ağlardan Twitter'da kullanıcının cinsiyetini tahmin edilmeye çalışılmaktadır. Cinsiyet, kullanıcı adı, kenar çubuğu ve profilin bağlantı rengi, kullanıcının oluşturduğu içerik olan tweetlerinin konusu, kullanıcının yarattığı tweet sayısı ve kullanıcının görüntüsü, kullanıcının cinsiyetini tahmin etmek için kullanılmaktadır. Bu özelliklerin -ya da bilgilerin- katkısı, bir başka deyişle toplam tahminin sonucunun hesaplanması için kullanılacak ağırlık, lojistik regresyon ile olacaktır. Kısım V'de ayrıntılı olarak açıklanan farklı algoritmalar, kullanıcının cinsiyetini her bir özellikten tahmin etmek için kullanılır. Çalışmanın şu ana kadar yapılan diğer çalışmalardan temel farkı ise tahmin algoritmasıdır; bu algoritma çoğunlukla metin analizine dayanan [2] önceki çalışmalardan farklı şekilde kullanıcı adı, kenar çubuğu ve profilin bağlantı rengi, kullanıcının tweetlerinin konusu, kullanıcının tweet sayısı ve kullanıcının imajı gibi birçok özelliği göz önünde bulundurur ve kullanır.

II. Yazılım ve donanım araçları

Projenin uygulanmasında MatLab programlama dili kullanılmıştır, çünkü bu programlama dili yukarıda belirtilen sorun için uygun ve faydalı çözümler sunmaktadır. Bu uygun ve faydalı çözümler arasından, özellikle MatLab'ın İstatistik ve Makine Öğrenimi Araç Kutusu (MatLab's Statistics and Machine Learning Toolbox) ve Eğri Uydurma Araç Kutusu (Curve Fitting Toolbox) tercih edilmiştir. Kullanıcı adına göre cinsiyet çıkarımı yapılabilmesi için ise Gender API kullanılır. [3] Kenar çubuğuna göre cinsiyet tahminlemesi, bağlantı rengine ilişkin algoritmalar ve kullanıcının tweetlerinin irdelenmesi de yine MatLab kullanılarak hayata geçirilmektedir. Kullanıcının tweetlerinin konularını çıkarmak için LDA algoritması kullanılacaktır. Görüntü işlemede, cinsiyetten profil çıkarımı elde etmek için bir yüz tanıma aracı olan MatLab Toolbox kullanılacaktır [4] [5]. Projenin sonraki aşamalarında gerekebilecek diğer algoritmalar ise, MatLab programlama dili kullanılarak el ile tatbik edilecektir.

[1] OSN: Online Social Network (Çevrimiçi sosyal ağ). Kağıtta ilerleyen kısımda da aynı kısaltma terimsellik adına kullanılacaktır.

Donanım kısmında, bir önceki paragrafta atıfta bulunulan programlama dillerini derleme yeteneğine sahip derleyicileri ve gerekli diğer yazılımsal altyapıyı barındıran bir bilgisayar dışında farklı araçlara ihtiyaç duyulmamıştır. Ancak, yapılan işlemlerin çok fazla işlem gücü gerektiği bilgisini göz önünde bulundurarak, kullanılacak bilgisayarın halihazırdaki güncel teknolojilerin kullanıldığı görece işlem kapasitesi yüksek bir bilgisayar olması önerilmektedir.

III. Veri seti

Projede kullanılacak olan veri seti, kaggle.com'da bulunmuştur. [5] Kaggle, Google, Inc.'in sahip olduğu çevrimiçi bir veri bilimci ve makine öğrencisi topluluğudur ve kullanıcıların veri kümelerini bulmalarını, yayınlamalarını, web tabanlı bir veri bilimi ortamındaki modelleri keşfetmelerini, üretmelerini, diğer veri bilimcileriyle ve makine öğrenmeleriyle çalışmasını sağlar[6]. Veri seti 20.000 satırdan oluşmaktadır ve eksik değerler içermemektedir. Veri setinin boyutu, çalışmanın yürütülebilmesi için yeterlidir ve projenin ana motivasyonu sınıflandırma olduğundan, eksik değerlerin bulunmadığı bir veri seti çalışmanın amacına uygundur. (Bölüm IV'te ayrıntılı olarak açıklanan proje ilerlemesinde karşılaştığımız zorluklar nedeniyle veri setinin boyutu azaltılmıştır.)

Veri seti aşağıdaki alanları içermektedir:

unit_id	name
unit_state	profile_yn_gold
trusted_judgments	profileimage
last_judgement_at	retweet_count
gender	sidebar_color
gender:confidence	tweet_created
profile_yn	user_timezone text
profile_yn:confidence	tweet_coord
created, description	tweet_id
fav_number	tweet_count
gender_gold	tweet_location
link_color	

Tablo.1: Veri setinin sunduğu alanlar

Veri kümesi, kullanıcılar hakkındaki gerçek cinsiyet bilgilerini içerdiğinden, sonuçta algoritma ve gerçek değerler tarafından bulunan tahminler karşılaştırılarak değerlendirilebilmektedir. Her bir kullanıcıya ait tahmin sonucunda elde edilen veri, elde olan ve doğruluğu bilinen veri ile kıyaslanmakta, sonucunda ayrık bir sonuç dönülmekte ve bu sonuçlar bir başka dizi içinde

tutulmaktadır. Bu karşılaştırma yönteminin izlenmesi neticesinde, projenin sonucuyla ilgili gerçekçi bir doğruluk yüzdesi verilebilmektedir.

Karmaşık ve ham verilerle çalışırken, bunların seçimi makine öğrenme algoritmasının performansını önemli ölçüde etkileyebilmektedir [7]. Bu bağlamda, cinsiyet tahmininde kullanılacak özellikler el ile seçilmektedir, başka bir deyişle, cinsiyet kümesinde cinsiyet ile ilgili olabilecek en iyi doğruluğu verdiğine inandığımız veri setindeki özellikleri seçilmektedir.

IV. Karşılaşılan zorluklar

Projenin işleyişi esnasında, birden fazla süreçsel zorluk ile karşılaşılmıştır. Bu zorluklar, ilerleyen dönemde benzer işler üretmek isteyen kişilere zaman kaybı yaratabileceğinden, projeye dair sürecin sadeleştirilebilmesi adına burada aktarılmaktadır. İlk zorluk, projenin veri setinin işlenmesi sürecinde yaşanmıştır. Kaggle'dan indirilen veri kümesi, tablo verilerini depolamak için kullanılan ve farklı araçlarda kullanılmak üzere başka bir dosya uzantısı şeklinde içe veya dışa aktarılabilen basit bir dosya formatı olan .csv dosya uzantısına sahiptir. Veri kümesini düzenlemenin ilk adımı olarak, bu .csv uzantılı dosya, ilgili araçların yardımıyla açıldıktan sonra .xlsx uzantılı bir belgeye dönüştürülmektedir, böylece Microsoft Excel ile açılabilmekte ve daha sonra kolayca işlenebilmektedir. Veri kümesi dikkatlice incelendiğinde, bazı satırların alt satıra kaydığı görülmüştür, çünkü kullanıcının tweetini gösteren sütunlar gibi çok uzun bazı sütunlar, kendi doğal hücre boyu limitlerini aşarak alt satırda sunulan veri kümesinde kalan sütunların üzerine yazılmıştır. Veri setini MatLab'a sokmaya yönelik ilk adımlarda, kod verileri satır satır okuduğundan, bu durum basitçe beklenti dışında bir veri kümesi elde edilmesine sebebiyet vermektedir. Ek olarak, indirilen veri kümesindeki verinin kirliliği, yani farklı karakterleri kullanımı gibi problemler de görülmüştür, nitekim ASCII karakterleri MatLab'daki veri kümesinin temsilinde görünmemektedir. Bahsi geçen sorunların çözümü için veri seti Java'da yazılan bir kod kullanılarak düzenlenmiş ve temizlenmiştir ve sonuç olarak *reduced_datafile* isminde bir Excel dokümanına temiz veri kayıt edilmiştir.

Karşılaşılan bir başka zorluk, üç etiketin mümkün olduğu veri kümesi cinsiyet kategorisindedir: kadın, erkek ve marka. Marka olarak bir cinsiyet seçeneği sunulmasının sebebi, Twitter sosyal ağında markaları temsilen hesapların da var olmasıdır. Bu bağlamda, cinsiyet marka olarak belirtilmişse, temel olarak Twitter hesabının bir kişiye değil ticari amaçlarla belirli bir markaya ait olduğu anlamına gelir. Bunun bir sonucu olarak, veri setindeki, cinsiyet özelliğinde 'marka' olarak etiketlenen satırlar kaldırılmıştır.

Kullanıcının profil resminden cinsiyet çıkarımı kısmında ise bazı internet adreslerinde resim bulunmadığı fark edilmiş, bu durum da bazı hesapların kapalı olduğunu düşündürmüştür. Bahsi geçen hesap kapalılığı çalışmayı yürüten kişilerin kontrolü dışında bir durum olduğundan, kullanıcının profil resimlerinden çıkarım yaparken, bu satırları çıkarılmıştır. Resimlerle ilgili karşılaşılan bir diğer zorluk, kullanıcının geçtiğimiz ay içinde profil resmini değiştirmesi durumunda (veri kümesi bir ay önce paylaşılıyorsa), profil resmi Twitter hesabı açıldığında varsayılan resim olarak atanan boş profil resmi olarak gelmektedir, bu da küçük ve boş bir resim alanına işaret eder ve tahmin algoritmasında kullanılan MatLab kodu tarafından okunamadığı anlamına gelmektedir. Bu sorunu çözmek için, kullanıcının Twitter kullanıcı adını içeren URL'yi kullanarak doğrudan kullanıcının profil resmi alınmaktadır. Tüm bu işlemlerin sonunda yaklaşık olarak yüzde 35lik bir sadeleşme elde edilerek 13000 satırdan oluşan bir veri setine indirgenme sağlanmıştır.

V. Kullanılan algoritmalar ve yöntemler

Yapılan araştırma ve çalışma süresince, sağlıklı ve kullanılabilir çıkarımlar oluşturmak için ham verilere birçok yöntem uygulanmıştır. Her şeyden önce, bir kullanıcının erkek veya kadın olduğuna dair öngörüde bulunurken altı niteliği takip etmek gerekmektedir:

- Kullanıcı adı
- Kullanıcının profilinin kenar çubuğu rengi
- Kullanıcının profilinin rengi
- Kullanıcının yazdığı tweetlerin sayısı
- Kullanıcının yazdığı tweetlerin konusu
- Kullanıcının profil resmi

Dolayısıyla bu özellikler cinsiyet sınıflandırmasının yapılması için seçilmiştir, çünkü bunların kullanıcı cinsiyetlerinin doğru sınıflandırılmasında yüksek doğruluk sağlayan özellikler olduğuna inanılmaktadır.

a. Kullanıcı Adından Cinsiyet Çıkarımı

Kullanıcı adından çıkarım yapmak için, Gender API'si kullanılmaktadır. [2] Gender API, ismi girildiğinde, verilen ismin cinsiyetini doğrulukla veren bir uygulama programlama arayüzüdür. Veri kümesindeki kullanıcılar çoğunlukla e-postalarından otomatik olarak alınanlar gibi olduklarından, cinsiyetin e-posta adresleriyle belirlendiği API tarafından sağlanan işlevsellik kullanılmaktadır. Bu yüzden API'ye kullanıcı adı verildinde, kullanıcı adını standart e-posta adresi komut dosyasıyla birleştirecek ve kişinin cinsiyeti olasılığını sunacaktır.

b. Kenar Çubuğu Rengi ve Bağlantı Rengi kullanılarak yapılan çıkarım

Bu özelliklerin her ikisi için, CIE - Lab renk uzayı kullanılarak, kullanıcının kenar çubuğu rengini ve bağlantı rengini, cinsiyetlerin renk tercihleri hakkında yazılan kağıtta verilen en yakın renge atanacağı bir algoritma kullanılmıştır [7]. En yakın renge tayin edildikten sonra, cinsiyet tarafından seçilen rengin arka olasılığı elde edilmekte olup; bu, belirli bir cinsiyetin Twitter profili tercihinde o rengi seçmesi anlamına gelmektedir. Bu olasılık hesaplamaları, kullanılan makalede sunulan deneysel verilere dayanarak yapılmaktadır.

c. Tweet Sayısından Çıkarım

Histogramlar veri setinden tweet sayısı kullanılarak oluşturulmuştur. Tweet sayısının karşılaştırılmasında günlük yazılmış tweet sayıları kontrol edilir. Kullanıcıların Twitter'a kaydolma tarihleri farklı olduğundan, doğrudan tweet sayısına bakmak doğru bir sonuç vermeyecektir. Günlük tweet, kullanıcının toplam tweet sayısı ile Twitter'a girilen kullanıcı sayısı ile bu veri setinin yayınlandığı tarih arasındaki süreye bölünerek hesaplanmaktadır.

$$Günlük\ tweet\ sayısı = \frac{Toplam\ tweet\ sayısı}{Veri\ seti\ yayınlanma\ tarihi - kullanıcı\ kayıt\ tarihi}$$

d. Tweet Konusundan Çıkarım

Konuyu kullanıcının tweet'lerinden çıkarmak için Latent Dirichlet Allocation yöntemi kullanılmıştır. Latent Dirichlet Allocation (LDA), verilerin bazı bölümlerinin neden benzer olduğunu açıklayan gözlemlenmemiş gruplar tarafından açıklanmasına izin veren üretken bir istatistiksel modeldir.[8] Kullanıcı tweetinin konusunu aldıktan sonra, veri kümesinin yarısı kullanılarak uygulama basit bir şekilde eğitilmiş ve kullanıcının erkek veya kadın olma olasılığı elde edilmiştir.

e. Profil Resminden Çıkarım

Kullanıcının profil resminden çıkarım yapmak için, belge ile birlikte sağlanan MatLab kodu kullanılmış, böylece resimlerdeki cinsiyet sınıflaması yapılmıştır. [3] [4]. Veri kümesi çok büyük olduğundan ve her bir örnek için internete sıfırdan bağlanmanın maliyeti yüksek olduğu için, aynı anda çalışma süresini azaltmak ve zamandan tasarruf etmek için görüntü indirme işlemleri gerçekleştirilmiştir.

f. Lojistik regresyon

Yukarıda sunulan özelliklerin her biri, kullanıcının ait olduğu cinsiyet sınıfını belirleme olasılığının hesaplanmasında farklı bir ağırlığa sahip olacaktır. Örneğin, kenar çubuğu rengi, kullanıcının cinsiyeti hakkında, kullanıcının profil resminden daha zayıf bir sezgi vermektedir. Her bir özellik tarafından ayrı ayrı verilen olasılıklar hesaplandıktan sonra, erkek veya kadın olmasına katkıda bulunan her özelliğin ağırlıklarını hesaplamak için lojistik regresyon kullanılmıştır.

VI. Yürütme planı

Veri kümesi, biri eğitim için diğeri test için eşit büyüklükte iki alt gruba bölünecektir. Veri seti, kullanıcıların gerçek cinsiyet bilgilerini içerdiğinden, değerlendirme için, çıkarımın doğruluğu bir karışıklık matrisi ile gösterilecektir. Amaç mevcut cinsiyet tahmini yöntemlerinin doğruluk seviyesine ulaşmak veya en azından mevcut yöntem doğruluğunun yüzde 20'sinden daha az olmayan bir doğruluk seviyesine ulaşmak olacaktır.

Yukarıda belirtildiği şekilde, veri kümesi .csv biçimindeki Kaggle'dan alınmıştır ve bazı satırlar karışık, bazıları ise ASCII olmayan karakterlere sahiptir. Önce karışık ve ayrılmış satırlar bir Java uygulamasıyla sabitlenmekte (hakimiyet sebebi ile kullanılan Java yerine, farklı programlama dilleri de kullanılabilmektedir), ardından ASCII olmayan karakterleri satırlardan kaldıran bir döngüsü eklenerek veri dosyasını temizlenmektedir.

Bir sonraki adım, eğitim ve test aşamasında kullanılmayacak olan sütunların kaldırılmasıdır. Bu görece kolay bir işlemdir, sütunlara sadece boş hücreler atanmaktadır. Bu sürüm *ham* adında bir matrise koyulduktan hemen sonra Preferences.mat dosyasına kaydedilmektedir.

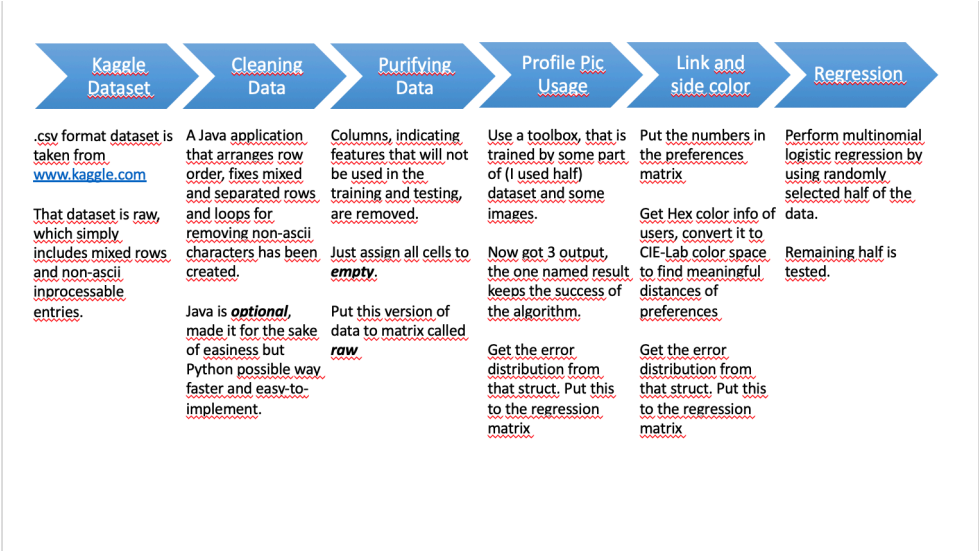
MatLab tarafından sağlanan çok terimli lojistik regresyon kullanıldığı için; regresyon, gerçek cinsiyetler ve veri kümesi özelliklerinden hesaplanan olasılıklar matrisine nominal yanıtlar ve öngörücülerin koyulması gerekmektedir. Bunun yapılması için ise, erkek olma ihtimali profil imajından, isimden, bağlantı renginden ve kenar çubuğu renginden çıkarılmalıdır. İlk olarak, cinsiyet görüntüleri çıkarmak için profil görüntüleri kullanılmaktadır. Profil görüntüleri kullanılarak cinsiyeti tahmin etmek için ise "COSFIRE_Gender_recognition_1_0" adlı klasörde bulunan bir araç kutusu kullanılmaktadır. Geliştirilen algoritma, veri kümesinin yarısını ve bazı yüz görüntülerini kullanarak eğitilmektedir ve tüm veri kümesi üzerinde test

edilmektedir. Algoritma temelde 3 çıktı yapısı sağlar ve bu çıktılar *data.mat* dosyasına kaydedilmektedir. *Result* adındaki çıktılardan biri algoritmanın başarısı hakkında bilgi tutmaktadır ve hata dağılımını o yapıdan alınmaktadır. Bundan sonra özellik sütununu bu hata dağılımından alınmaktadır. Bu sütun regresyon matrisine koyulmakta ve ileride kullanmak üzere *preferences.mat* dosyasına kaydedilmektedir.

Daha sonra, erkek olma ihtimalinin tespit edilebilmesi için link rengi ve kenar çubuğu renk bilgisi kullanılmaktadır. Burada, renklerin cinsiyetini anlamlandırabilmek adına erkeklerin ve kadınların renk tercihleri hakkında bir makale kullanılmıştır. Renklerden elde edilen numaralar, tercihler matrisine işlendikten sonra ve *preferences.mat* dosyasına kaydedilmektedir. Makalede, renk tercihi olmayan bazı insanların olduğu bilgisine erişilmiş olup, geliştirme sürecinde bu kişilerin Twitter'ın varsayılan rengini seçtiği düşünülmüştür. Her kullanıcının HEX renk bilgisini aldıktan sonra, tercihlerin anlamlı mesafelerini bulmak ve verilen renkle erkek olma hesaplanmış posterior olasılığını bulmak için onu CIE-Lab renk uzayına dönüştürüyoruz. Bunlar, linkcolorprob ve sidecolorprob matrislerine yerleştirilir ve *preferences.mat* ile regresyon matrisine kaydedilir.

Algoritmanın tahmin sürecinde son adım olarak isim bilgisi kullanılır. Hesap adları veri kümesinde bulunduğundan ve bu adları kullanarak faydalı bir bilgi alınamadığından, kullanıcı profili sayfasından HTML kodunu alan ve bu HTML kodunu işleyen Java programı kullanılmış olup kişilerin isimleri HTML kodundan tarama yöntemi ile elde edilmiş ve .txt uzantılı bir dosyaya yazılmıştır. Bu kayıt dosyası, bir Python betiği tarafından okunmakta ve isim veritabanında PostgreSQL ile bütünleşik Python kullanılarak aranmakta olup, olasılık buradan elde edilen bilgi kullanılarak hesaplanmaktadır. Olasılıklar yine bir .txt uzantılı dosyaya yazılmakta ve MatLab'daki ve regresyon matrisindeki *adeprob* ismindeki matrise konulmaktadır. Bu matris, *preferences.mat* dosyasına da kaydedilmektedir.

Veriler hazırlandığından, verinin rastgele seçilen yarısını kullanarak çok terimli lojistik regresyon gerçekleştirilmektedir. Lojistik regresyondan sonra veri setinde kalan yarı boyuttaki kısım ise teste tabii olmaktadır. Göz önünde bulundurulması gereken nokta ise, bu işlemlerin yalnızca ilk seferde geçerli olduğu gerçeğidir; nitekim verilerin tamamı dosyalara kaydedildiğinden bahsi geçen karmaşık adımlara her zaman gerek duyulmaz, dosyalarıyalnızca programın başında yüklenmektedir.

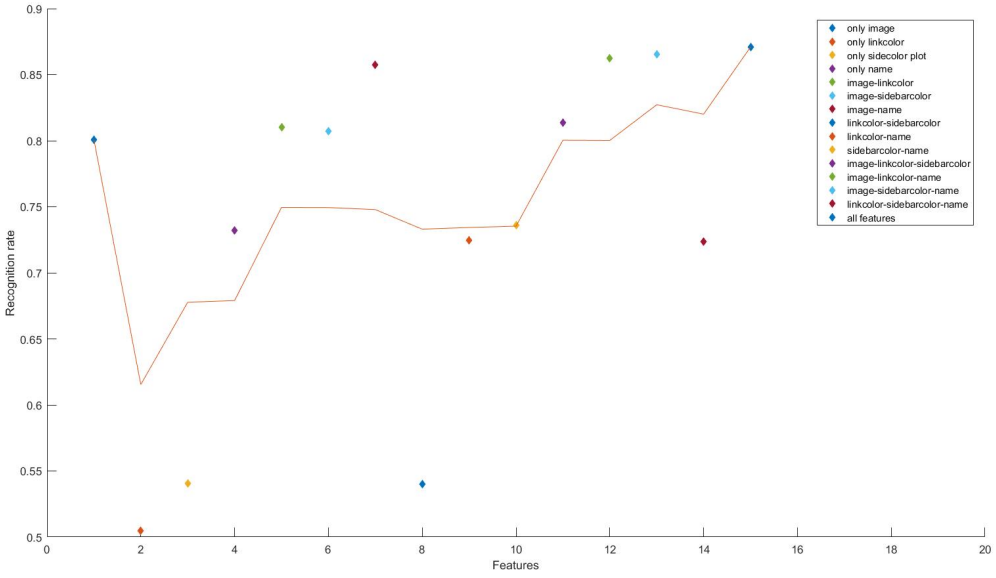


Resim.1: Çalışmanın yapılması sürecinde atılan adımların gösterimi (İngilizce)

VII. Sonuçlar

Yapılan çalışmanın neticesinde, yukarıda bahsedilen veri seti ve yöntemler kullanılarak elde edilen sonuçlara göre, yaklaşık olarak 90% civarında bir doğruluk yüzdesi ile cinsiyet tahmini elde edilmektedir. Farklı olmayı hedefleyen renk ya da konu tercihleri gibi dış faktörler göz önünde bulundurulduğunda, çalışmanın hedeflenen çerçevede görevini yerine getirdiği söylenebilir. Ayriyeten, kullanılan özellik sayısının bu veri seti için belli sayıda olması ve başka veri setlerinde artırılabilme ihtimali, çıkan sonuçlara dair umut üretmektedir.

Ayriyeten, karşılaştırma yapılabilmesi amacıyla, bazı özelliklerin çıkarılması ve eklenmesine dair çalışma da gerçekleştirilmiştir. Bu çalışmada görüldüğü şekilde, yalnızca resimden cinsiyet çıkarımı yapılması, cinsiyet tahmininde bu çalışma neticesinde elde edilen sonuçlardan yaklaşık 12% kadar eksik başarı ile tamamlama sağlarken, sadece kenar çubuğunun cinsiyet tahmininde kullanılması bu oranı 65%'lere kadar düşürmüştür. Her ne kadar diğer özelliklerin rolü de ayrı şekilde temsilen bulunsa da, özellikle kenar çubuğunun seçilmesinin tahmin yönteminde doğruluğu düşürdüğü sonucu göz önünde bulundurulması gereken bir gerekliliktir.



Resim.2: Farklı özelliklerin etiketlenmesi neticesinde elde edilen çalışma sonuçları

Elde edilen grafiklere göre verinin renk faktörü değerlendirilmeye katılmadan önceki ve sonraki dağılımı da çeşitlilik göstermektedir. Sonuçlara göre, kullanıcıların seçtikleri renklerin değişkenlik göstermesi, verinin homojenliğine katkıda bulunmaktadır.

Referanslar

[1] Jin, X., Wah B., Cheng X., Wang, Y., (2015). Significance and Challenges of Big Data Research. Big Data Research (2), 2015, pp. 59-64.

[2] Miller, Z., Dickinson, B. and Hu, W. (2012). Gender Prediction on Twitter Using Stream Algorithms with N-Gram Character Features. International Journal of Intelligence Science, 02(04), pp.143-148.

[3] Gender API. (2016).Gender API. [online] Available at: https://www.gender-api.com/ [Accessed 28 Dec. 2016].

[4]George Azzopardi, Antonio Greco and Mario Vento, "Gender recognition from face images with trainable COSFIRE filters", IEEE Advanced Video and Signal-based Surveillance (AVSS) 2016, in print

[5] George Azzopardi and Nicolai Petkov, "Trainable COSFIRE filters for keypoint detection and pattern recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35(2), pp. 490-503, 2013.

[6] Lardinois, Frederic; Mannes, John; Lynley, Matthew (March 8, 2017). "Google is acquiring data science community Kaggle". Techcrunch. Archived from the original on March 9, 2017. Retrieved March 9, 2017. Sources tell us that Google is acquiring Kaggle [...] the official announcement could come as early as tomorrow.

[7] Peersman, C., Daelemans, W., Van Vaerenbergh, V., 2011, Predicting Age and Gender in Online Social Networks, Proceedings of the 3rd international workshop on Search and mining user-generated contents, 2011, pp. 37-44

[8] Blei, David M., et al. “Latent Dirichlet Allocation.” *Journal of Machine Learning Research 3*, 2003, pp. 993–1022., www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.